C.Eng.465 Hw#4 Report

Clustering Microarray Samples

Utku Şirin

1560838

05.06.2012

In this report we have analyzed the microarray dataset of 30 samples. There are 22215 human transcripts. We have tried to cluster the samples into two groups, one for healthy tissues and one for disease tissues. After clustering the samples, we have tried to find 10 genes that are highly expressed in healthy tissues and not expressed in diseased tissues; 10 genes that for vice versa. In order to cluster the samples we have done seven different experiments by using the tool MeV(Multi Experiment Viewer) and in order to find the 10 + 10 genes, we have written our own code by java and try to find genes that are expressed more than a threshold in healthy and diseased tissues. The threshold is determined empirically.

PART I: CLUSTERING THE SAMPLES

In order to cluster the samples we have done seven different experiments. The first and second ones are k-means tests with different distance measures. In one test the distance measure is Euclidean distance and in the other test the distance measure is Pearson correlation. The third and fourth ones are hierarchical clustering (HCL) tests with Euclidean distance and Pearson correlation distance measures. The fifth and sixth ones Self-Organizing Map (SOM) tests by using Euclidean distance and Pearson correlation measure at each. The last test is also a hierarchical clustering test with one different distance measure, cosine similarity. This test is required to distinguish one last remaining sample as healthy or diseased. The details explained in this section, later. After applying each different algorithm, we have extracted three lists: Healthy, Suspicious, Diseased and try to identify the suspicious ones.

K-Means Experiments: Since k-means results differently with respect to the initial conditions, we have done five different k-means tests for the two different distance measure, Euclidean distance and Pearson correlation; build the healthy, suspicious and diseased groups by checking all those ten experiments. We have done 500 iterations.

The groups for Euclidean distance as below (only healthy group is written):

Exp1: 2, 3, 10, 14, 17, 21, 13, 15, 18

Exp2: 2, 3, 9, 10, 13, 15, 17, 18, 25, 27, 14, 19, 21

Exp3: 2, 10, 15, 17, 18, 21, 25, 3, 9, 13, 14, 19

Exp4: 2, 9, 10, 15, 19, 25, 27, 21

Exp5: 2, 9, 14, 17, 3, 13, 15, 18, 21, 10, 19

The results are grouped as 5-times occurred, 4-times occurred etc… :

5-times: 2, 10, 21, 15

4-times: 3, 14, 17, 13, 18, 9, 19

3-times: 25

2-times: 27

1-times: -

The groups for Pearson Correlation as below (only healthy group is written):

Exp1: 2, 3, 9, 10, 17, 19, 21, 27, 8, 14, 15, 18, 25

Exp2: 13, 14, 15, 17, 19, 21, 28, 29, 2, 3, 8, 10, 18, 22

Exp3: 2, 8, 10, 13, 17, 19, 21, 28, 3, 14, 15, 18

Exp4: 2, 3, 8, 9, 15, 18, 19, 21, 25, 10, 14, 17, 27

Exp5: 2, 8, 9, 10, 13, 17, 19, 3, 14, 15, 18, 21

The results are grouped as 5-times occurred, 4-times occurred etc… :

5-times: 2, 3, 10, 19, 17, 21, 8, 14, 15, 18

4-times: -

3-times: 9, 13

2-times: 25, 27, 28

1-times: 22, 29

After having the tests available, we have assumed 5-times and 4-times occurred are confident. Hence intersection of two sets of experiments for 5-times and 4-times comprises our healthy set for k-means tests. We also have assumed that 3-times and 2-times occurred are in suspicious list. Beside, the ones that are out of the intersection of 5-times and 4-times sets are also in suspicious list (sample 9 and 13, in this case). So the first healthy and suspicious lists are as below. We do not write the diseased ones, as the rest is obviously diseased ones.

Healthy: 2, 3, 10, 17, 19, 21, 14, 15, and 18

Suspicious: 8, 9, 13, 25, 27, and 28

Hierarchical Clustering Experiments (HCL): We have built two HCL trees based on Euclidean distance and Pearson correlation, and using average linkage. The trees are shown in Figure 1 and 2.

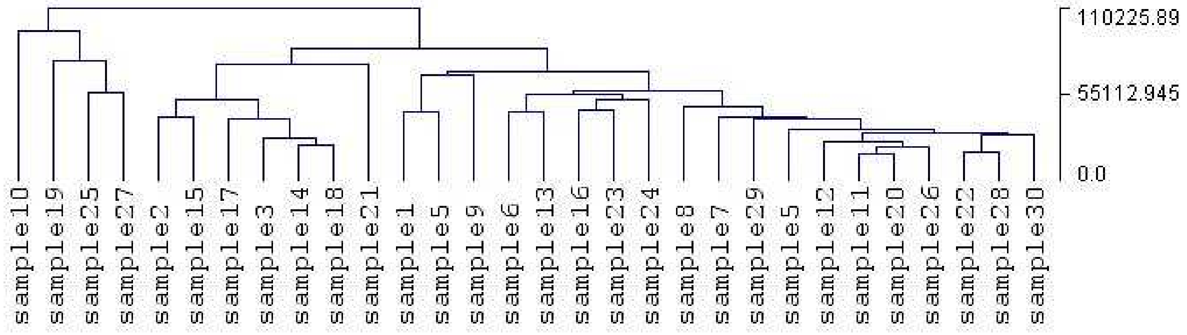

Figure 1: Hierarchical Tree based on Euclidean distance

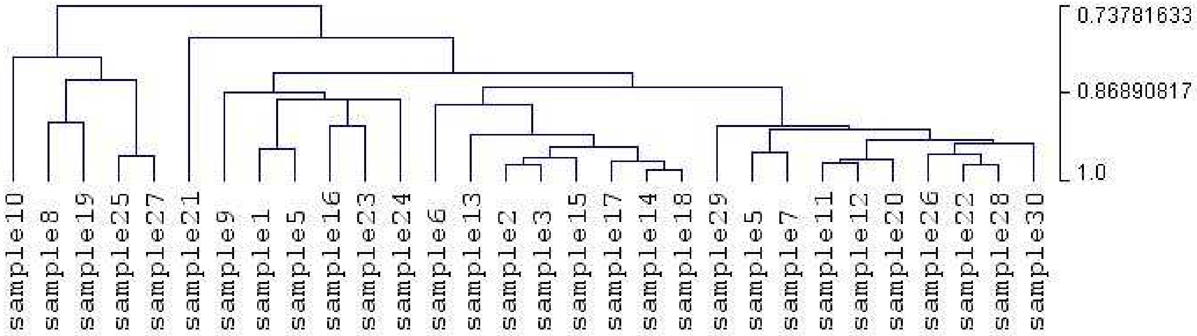

Figure 2: Hierarchical Tree based on Pearson correlation

Here, we have tried to identify the ones that are found in k-means together, in which we have seen together in both of the trees, in different places.

The groups based on Euclidean distance are as below:

Very-Close: 2, 15, 17, 3, 14, 18, 21

Close: 10, 19, 25, 27

Here, the Very-Close group is the ones at the second accumulated part of the HCL in Figure 1. The Close group on the other has same distance to the Very-Close group and to the other samples in the three. However, from the results of the k-means we see that they are either in healthy or in suspicious group. Hence, we have included them as Close.

The groups based on Pearson correlation are as below:

Very-Close: 2, 3, 15, 17, 14, 18

Semi-Close: 6, 13, 21

Close: 10, 8, 19, 25, 27

Here, the Very-Close and Close groups are same, but Semi-Close group stands for the middle steps between the Very-Close and Close group. They are closer to the Very-Close group and also seen in the k-means results also. After having the HCL experiment results, we have grouped the samples as healthy and suspicious as below:

Healthy: 2, 3, 15, 17, 14, and 8

Suspicious: 8, 10, 19, 25, 27, 21, 6 and 13

Here we have take the intersection of the two Very-Close groups as Healthy and put the rest to the suspicious group.

Self-Organizing Map Experiments (SOM): We have again built two self-organizing maps based on Euclidean distance and Pearson correlation. We have set the resolution as 2 x 2. Although it may seem SOMs are not so suitable for our task, since we want to cluster the samples into two groups but at least we have 2 x 2 resolution, hence 4 groups, the results are very confident in terms of clustering. After having the experiments, we got only 2 cells full and the other 2 cells are empty. That improved our confidence to the microarray samples since they are already grouped into 2, although the system allows them to be divided four. Only for the Pearson correlation case, we got sample 13 alone in one cell. The results are as below.

Euclidean distance results:

Healthy: 18, 13, 14, 3, 17, 9, 15, 21, 2, 27, 19, 10, and 25

No suspicious!

Pearson correlation results:

Healthy: 18, 15, 8, 14, 17, 3, 9, 2, 21, 19, 10, 27, and 25

Suspicious: 13 (alone in one cell in the SOM)

So, by intersection the healthy ones and putting the rest to suspicious the overall result for SOM experiments is as below:

Healthy: 18, 14, 3, 17, 9, 15, 21, 2, 27, 19, 10, and 25

Suspicious: 8, 13

Just before concluding the healthy and diseased results, let us consider one more subtle point for the sample 8.

The special experiment for sample 8: One interesting result from the experiments is that the sample 8 is always in healthy list for the experiments based on Pearson correlation but absolutely in diseased list for the experiments based on Euclidean distance. Hence we have done one more experiment to be able to understand the sample 8 class. It is a hierarchical experiment based on cosine similarity. Our aim is to understand the position of sample 8. Since Pearson correlation subtracts the mean, we cannot understand exactly the position of sample 8 with respect to the other. In the Figure 3 we see the HCL tree based on cosine similarity.

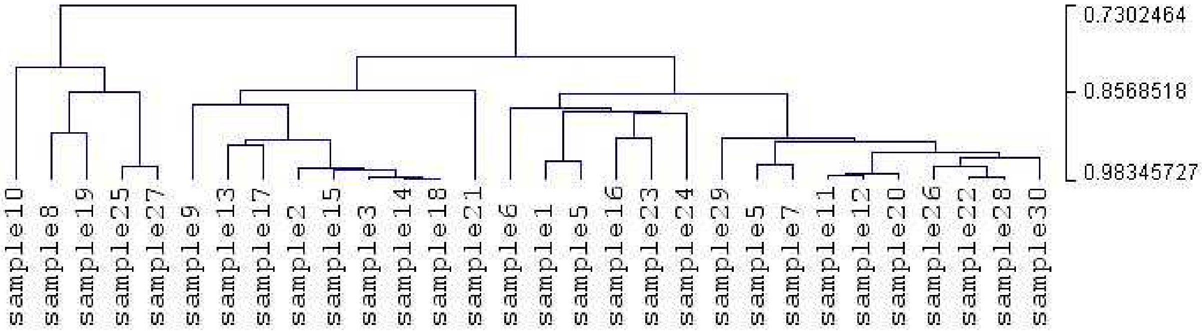

Figure 3: Hierarchical Tree based on Cosine similarity

As it can be seen from the tree, sample 8 is at the same place in the HCL as the place in the HCL tree based on Pearson correlation. This shows that, sample 8 is close to healthy samples in terms of the angles between them, however, it is far away them spatially since Euclidean distance measure have given very high results. Hence, we conclude that sample 8 is not in healthy group.

Overall result: By combining three healthy and suspicious groups we are able to obtain the ultimate results. The summary of the results are as below:

Summary of the results:

K-Means:

Healthy: 2, 3, 10, 17, 19, 21, 14, 15, and 18

Suspicious: 8, 9, 13, 25, 27, and 28

HCL Tree:

Healthy: 2, 3, 15, 17, 14, and 8

Suspicious: 8, 10, 19, 25, 27, 21, 6 and 13

SOMs:

Healthy: 18, 14, 3, 17, 9, 15, 21, 2, 27, 19, 10, and 25

Suspicious: 8, 13

In order for doing this, we have labeled each candidate samples as below.

2, 3, 15, 17, 14, and 18 → Always in Healthy

10, 19, 21 →2-times in Healthy, 1-times in Suspicious

9 → 1-times Healthy, 1-times suspicious

25, 27 → 1-times Healthy, 2-times suspicious

13 → 3-times suspicious

6, 28 →1-times suspicious

6 → 1-times suspicious

From those,

- "Always in Healthy" is labeled as healthy
- "2-times times in Healthy, 1-times in Suspicious" is also labeled healthy
- "1-times Healthy, 1-times suspicious" is labeled as healthy. This is because k-means resulted in 4-times in Euclidean distance and 3-times in Pearson correlation as healthy. SOM also supports its healthy labeled. We can say that it is high-ranked suspicious, hence taken as healthy.
- "1-times Healthy, 2-times suspicious" labeled as healthy. Although k-means have given 3-times for Euclidean and 2-times for Pearson correlation for sample 25 and sample 27, SOM supports their healthy label. Hence, labeled as healthy.
- "3-times suspicious" also labeled as healthy. K-means resulted in 3-times in healthy, again high-ranked suspicious. Hence, labeled as healthy.
- "1-times suspicious" is labeled as diseased since it is low ranked suspicious. Only k-means based on Pearson correlation presented the sample 28 3-times over 5-times. Hence not labeled as healthy, labeled as diseased. For sample 6, only HCL tree based

on Pearson correlation presented sample 6 is somehow related, but relevance is not strong (coming after sample 13). So sample 6 also labeled as diseased.

Therefore the overall result for clustering the samples as healthy and diseased as below:

Healthy: 2 3 15 17 14 18 10 19 21 9 27 25 13

Diseased: 1 4 5 6 7 8 11 12 16 20 22 23 24 26 28 29 30

PART II: FINDING THE HIGHLY EXPRESSED GENES

To be able find the highly expressed in healthy and not expressed in diseased genes, for each gene, we have checked the minimum and maximum difference value between average of healthy samples and average of diseased samples. For each gene, we calculated the average value of healthy and diseased samples; and, checked whether it is greater than the maximum or less than the minimum. If so, set it to maximum possible difference value, or minimum possible difference value.

From all genes, the minimum and maximum difference between average of healthy and disease samples are as below:

MinVal = 4.520416259765625E-4 ~= 0

MaxVal = 27001.2138671875 ~= 27000

These values show the range of the values that the difference between healthy and disease samples may have for any gene. If a gene has a difference value between the average of its healthy samples and the average of its diseased samples such that it is almost same as to the *(MaxVal – MinVal)* , then, we can say that this gene is highly expressed in one of the tissues and not expressed in the other tissue. However, there are only 3 genes having this difference value, which is not enough in our task (10 + 10 genes should be found). Then, empirically, we have tried different threshold values than *(MaxVal – MinVal)* by dividing it progressively.

Empirically, we have found that the threshold that is equal to *(MaxVal – MinVal) / 12* produce 120 genes in which 12 of them is highly expressed and not expressed in diseased tissue, and 108 of them is highly expressed and not expressed in healthy tissues, which meets the required number of genes. The 10 + 10 genes that are chosen from those 108 + 12 genes as below.

Highly expressed in healthy samples and not expressed in diseased samples:

1. 200019_s_at
2. 200032_s_at
3. 200062_s_at
4. 200064_at
5. 200081_s_at
6. 200088_x_at
7. 200095_x_at
8. 200674_s_at
9. 200689_x_at
10. 200717_x_at

Highly expressed in diseased samples and not expressed in healthy samples:

1. 200703_at
2. 201891_s_at
3. 203455_s_at
4. 207430_s_at
5. 209118_s_at
6. 209699_x_at
7. 210297_s_at
8. 210592_s_at
9. 211628_x_at
10. 213350_at