**Algorithm in Bioinformatics**

Programming Assignment Three

By

Mahmood uz Zaman

E01407909

Instructor: Dr Li Zhang

COSC616

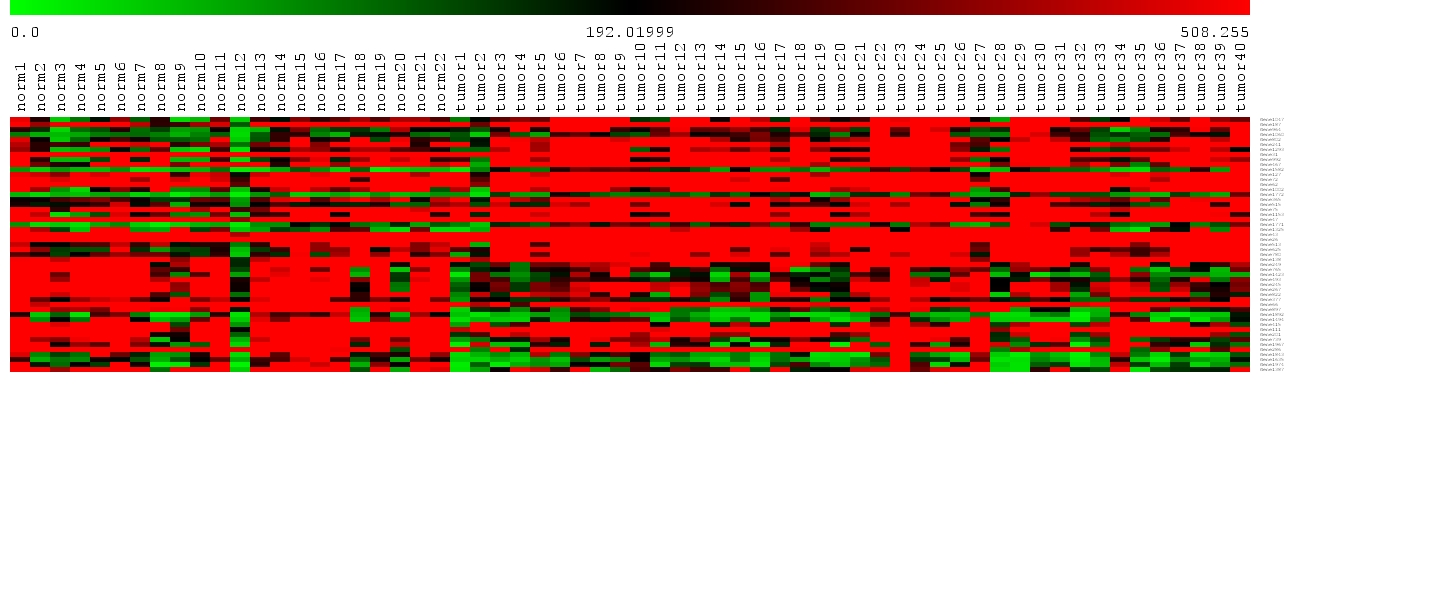
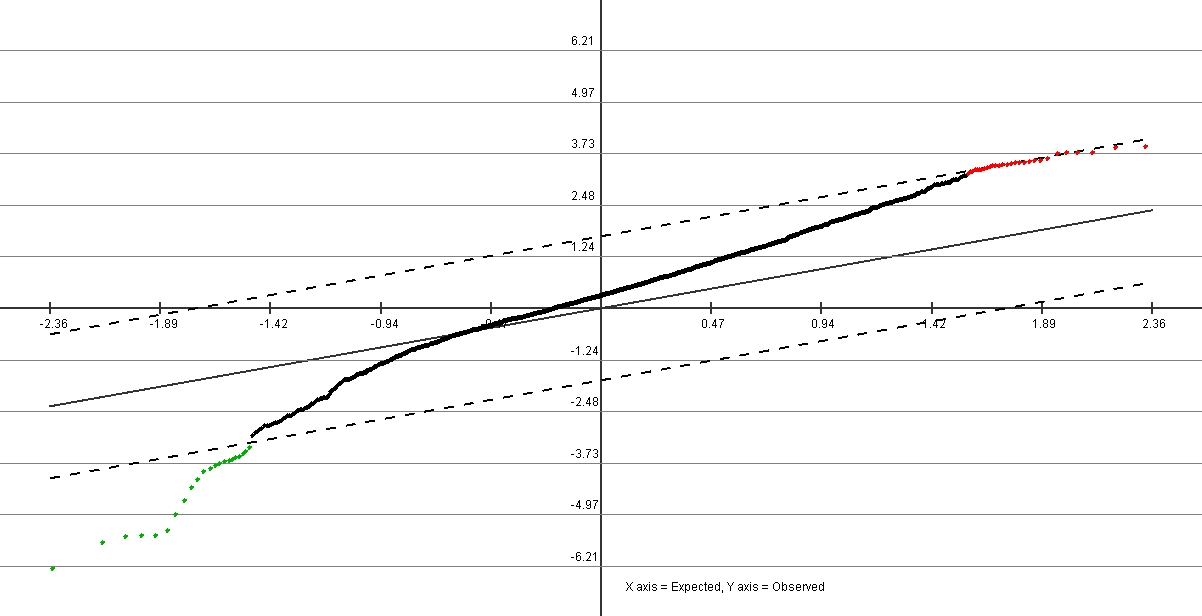
Fall 2016

Date Submitted: Dec 20, 2016

**Part-1 Supervised Learning:**

**SAM:** One significant/informative genes sample of 50 is taken for testing and training to apply hold out, reason of choosing these numbers in middle side is that on the SAM graph when we scroll to lower no it looks like you are ignoring some significant genes and if we took in 70’s it look like we are including the common pattern of genes for me no between 40 to 60 should include all the scatter points on the graph and other reason of not selecting 40 we don’t want our classifier to be relying on very less significant genes.

Significant Genes(50)



**Hold Out:**

Hold out of 50-50 is with different taken with different sample sets

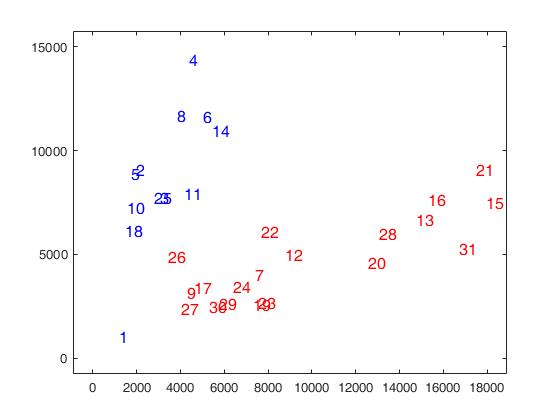
1. Training: [1:11 23:42] Testing: [12:22 43:62];
2. Training: [1:6 17:22 23:32 43:52 ], Testing [7:16 33:42 53:62];
3. Training: [6:11 12:17 32:41 52:62 ],Testing [1:5 18:22 23:31 42:51]

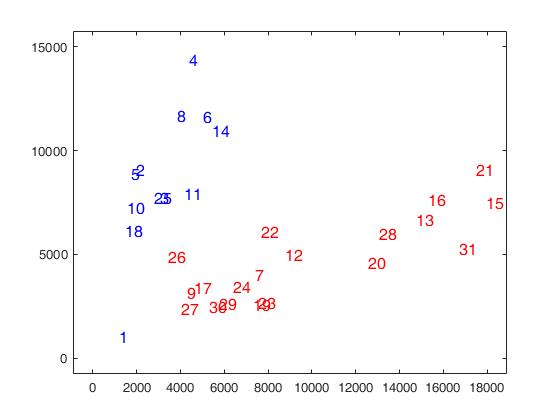
**Result Summary:**

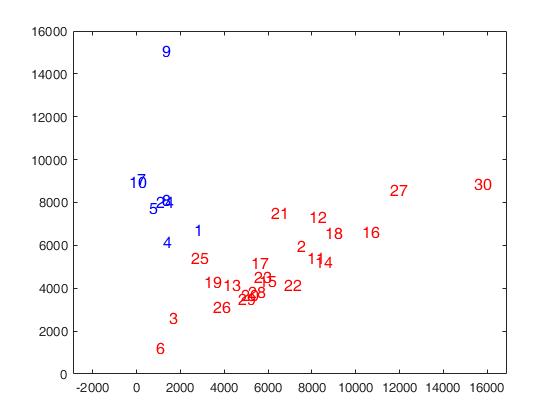
|  |  |  |  |
| --- | --- | --- | --- |
| **Distance** | **Hold out-1** | **Hold out-2** | **Hold out-3** |
| **Euclidean** | 19.3548 | 13.33 | 13.79 |
| **Correlation** | 16.1290 | 13.33 | 13.69 |

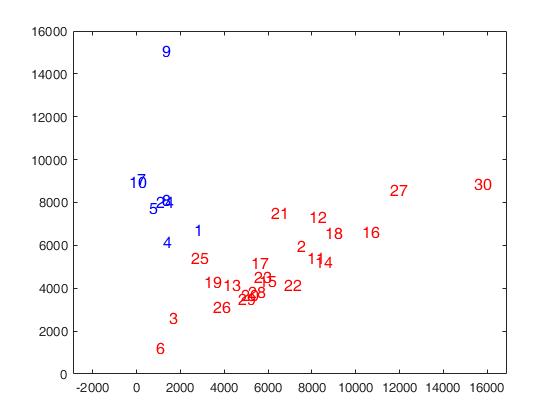
We can see from the result taking continuous samples didn’t provide good result rather taking scattered samples gives us good results with less error and these are more consistent.

**Scatter plot of result:**

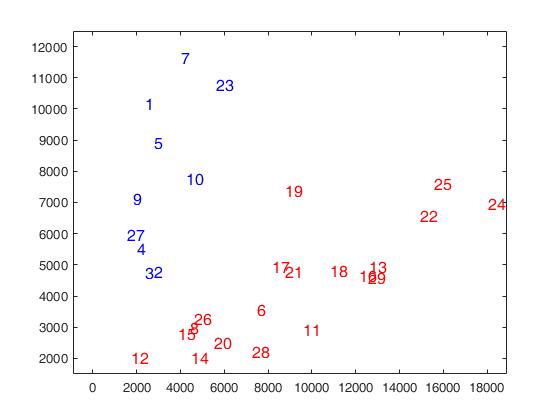
Hold out-1 Correlation Hold out-1 Eucledean

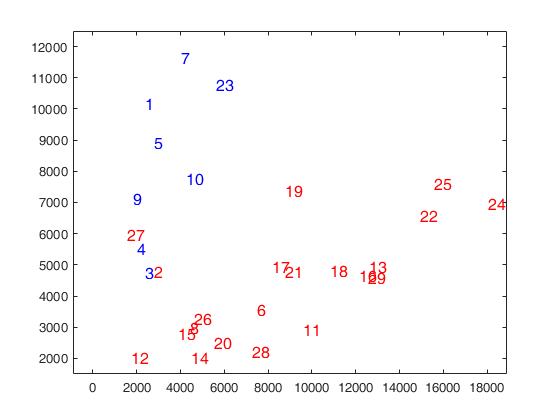


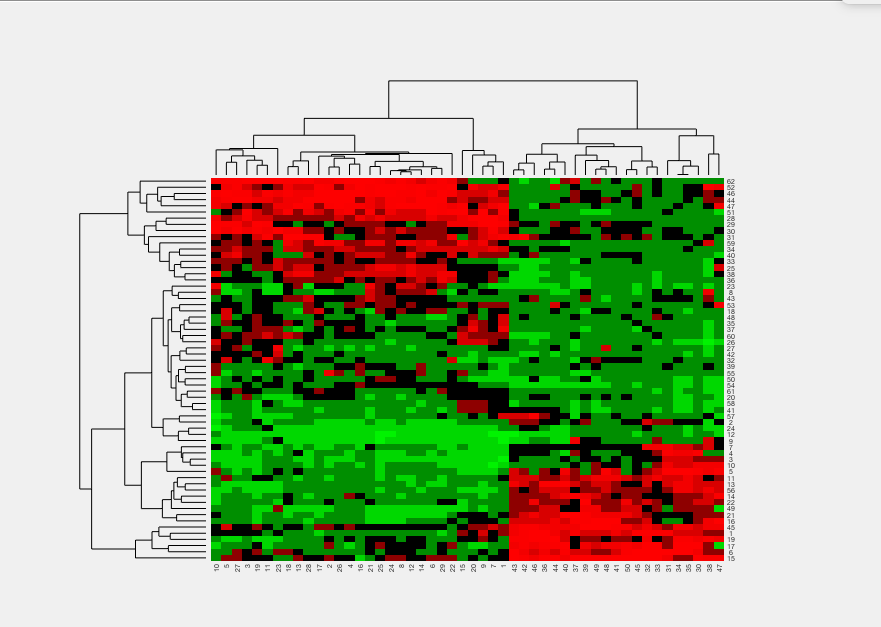
Hold out-2 Correlation Hold out-2 Eucledean



Hold out-3 Correlation Hold out-3 Eucledean

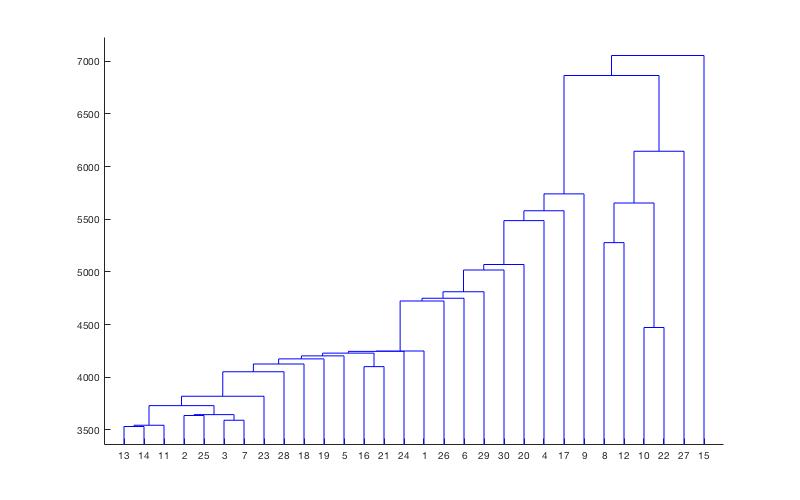
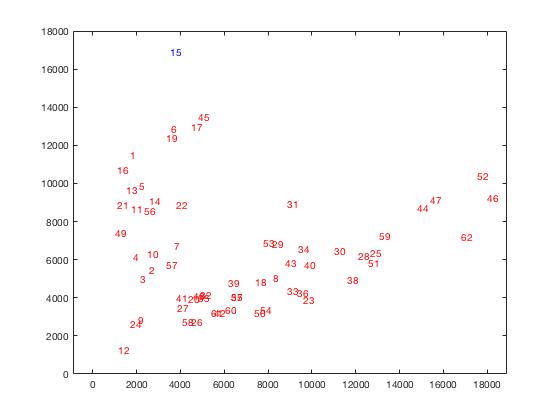


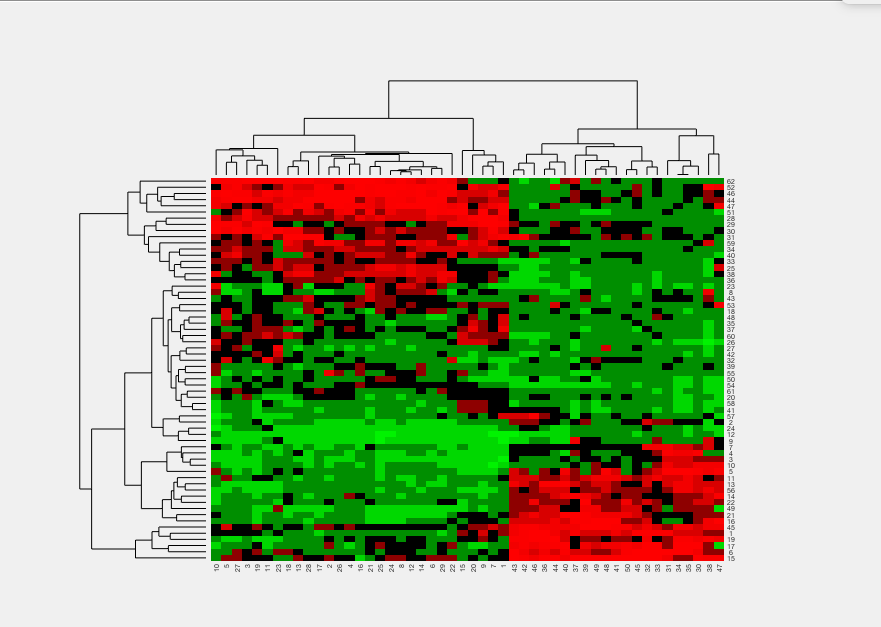


**Tree View**

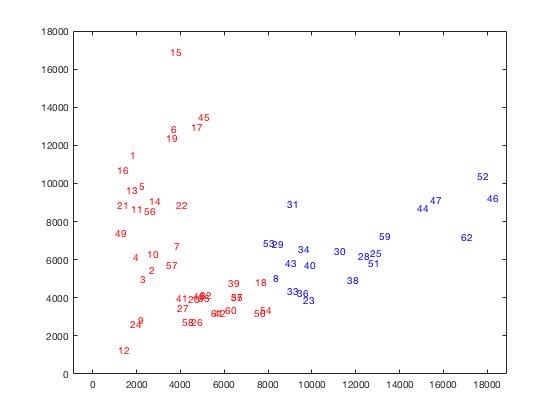
**Part 2 un supervised clusters:**

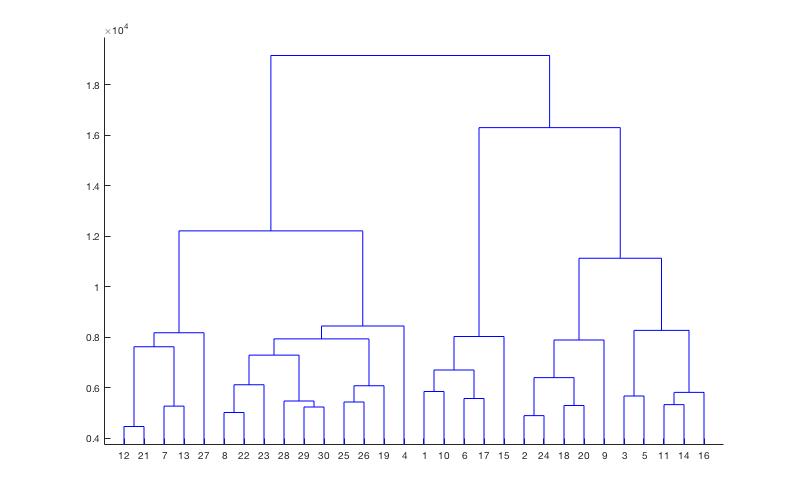
**HC Single Euclidean:**

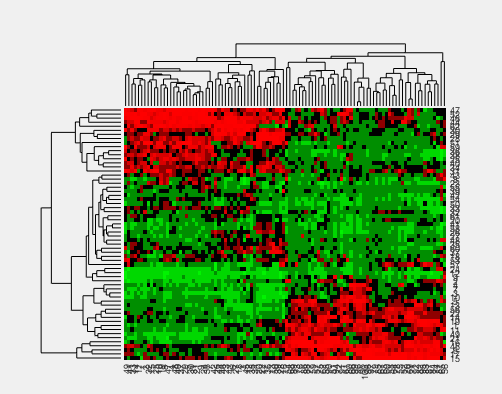




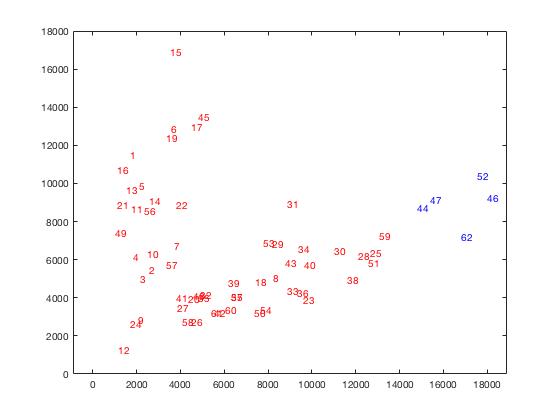
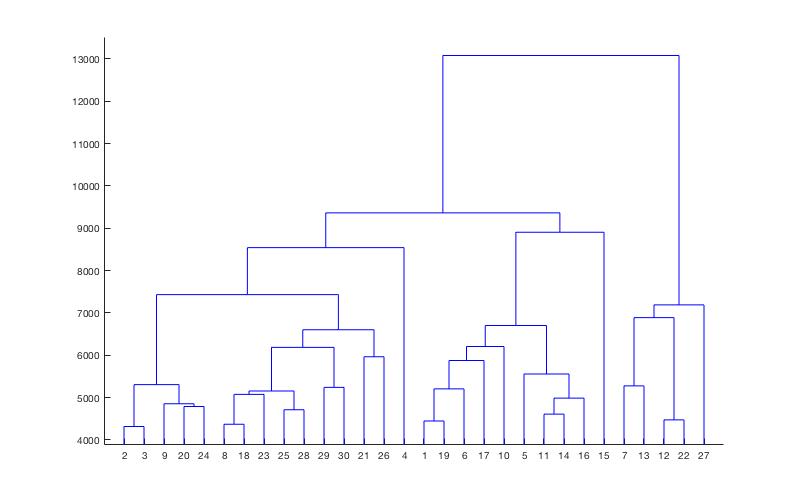
**Error: 66.1290**

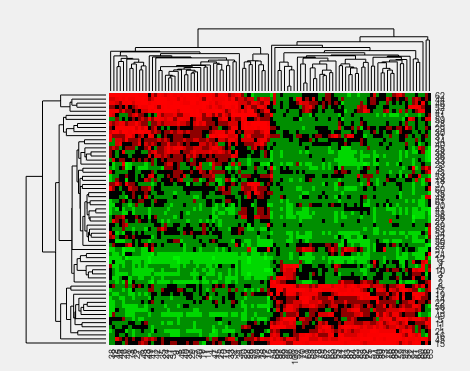
**HC average Euclidean:**



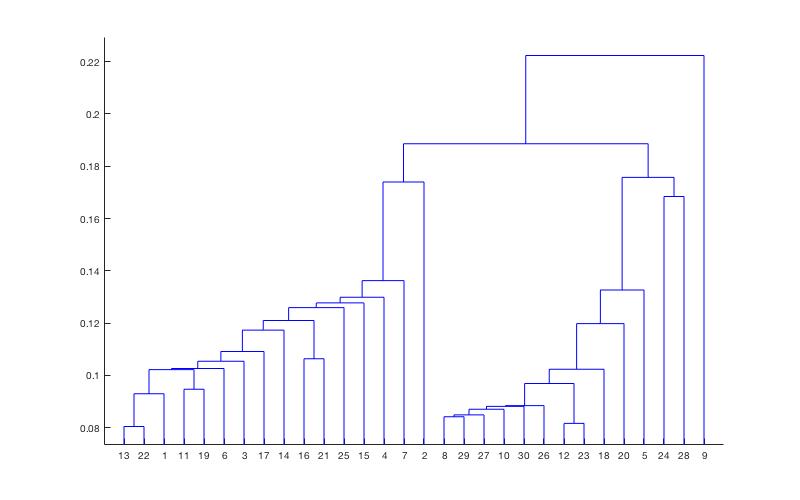


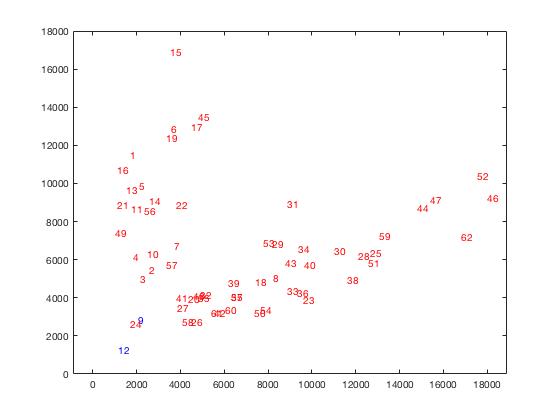
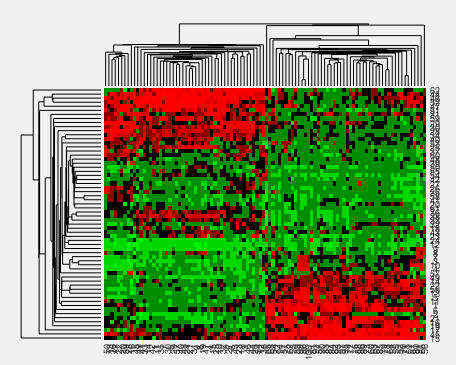
**Error: 33.87**

**HC complete Euclidean**

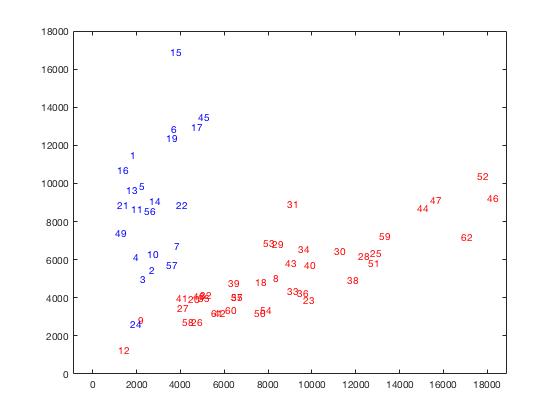
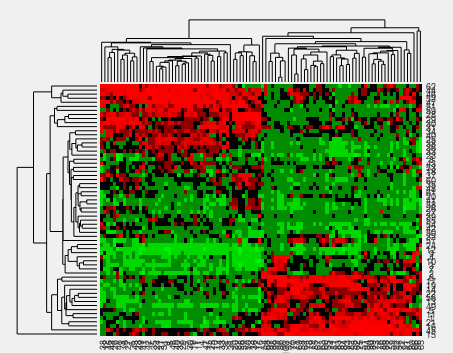


**Error: 56.45**

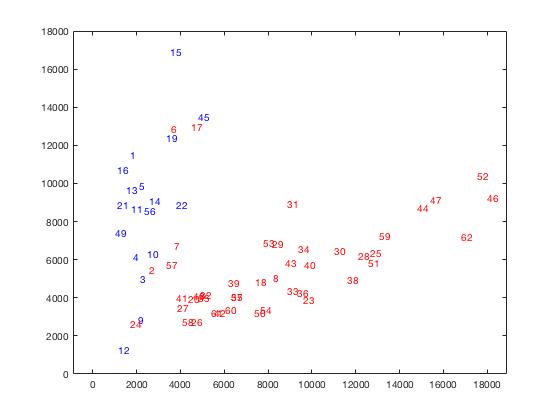
**HC Single Correlation**

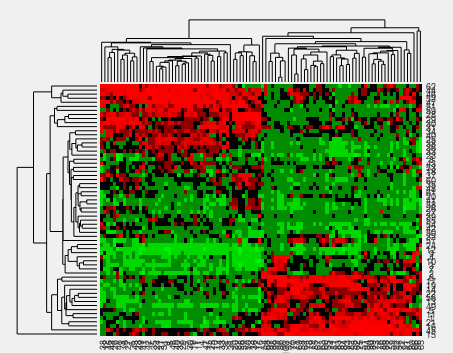


**Error:67.74**

**HC Average Correlation:**

**Error: 16.12**

**HC Complete Correlation:**



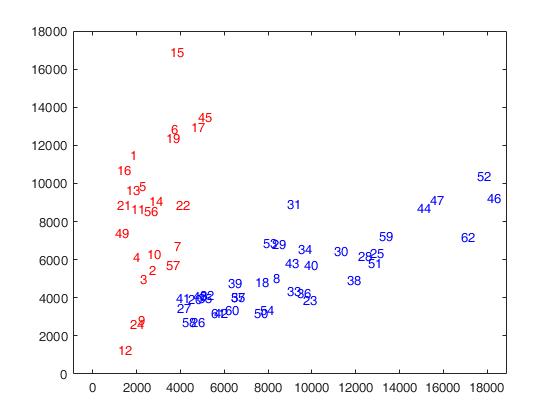
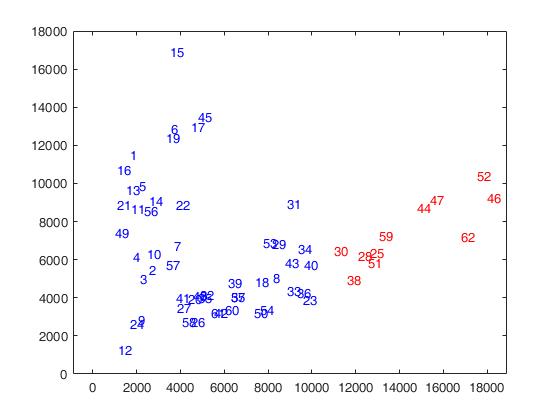
**Error: 16.12**

**Comparision and conclusion**

|  |  |  |
| --- | --- | --- |
| **Method of analysis** | **Distance** | **Error %** |
| HC single linkage | correlation | 67.74 |
| HC avg linkage | Correlation | 16.12 |
| HC complete Linkage | Correlation | 16.12 |
| HC single linkage | euclidean | 66.12 |
| HC avg linkage | euclidean | 33.81 |
| HC complete Linkage | euclidean | 56.45 |

From the comparison table we can clearly see two combinations, average and complete linkage with correlation perform much better than the others and we can also see from the scatter plot in previous pages clusters are segregated very clearly. But Euclidean distance result cant be relied as they perform very poorly that gives one big clusters of all genes.

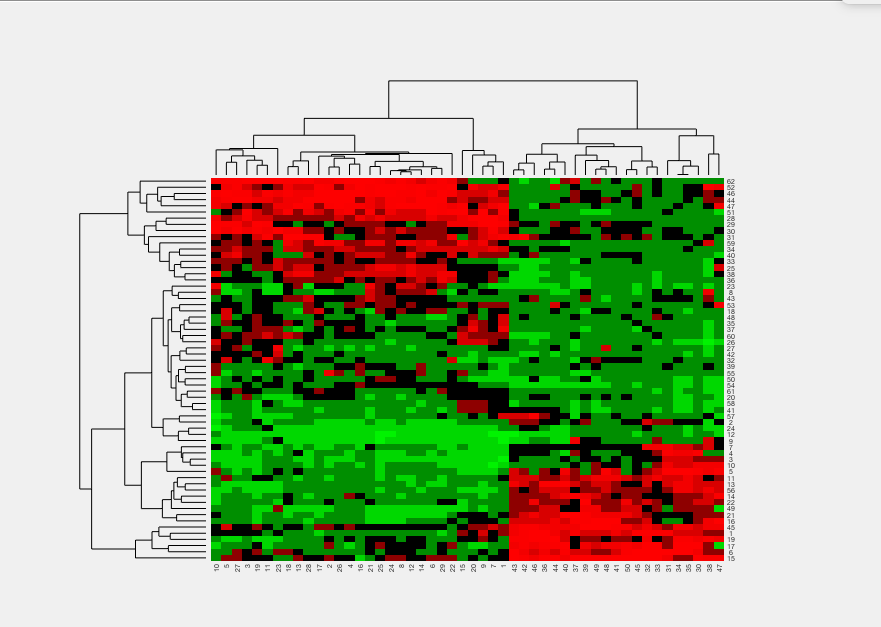
**KMC clustering Results:**

**Distance: Correlation KMean Distance: Seqlidean**

**Error: 12.9 Error: 46.72**

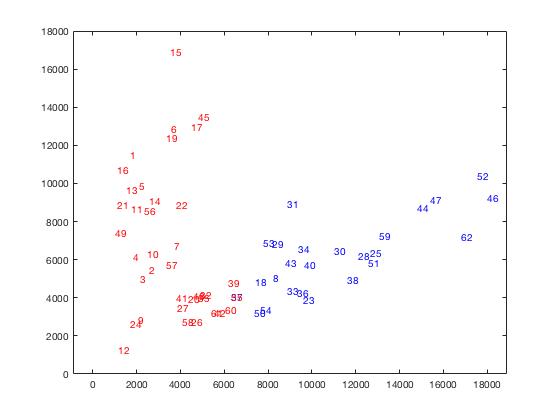
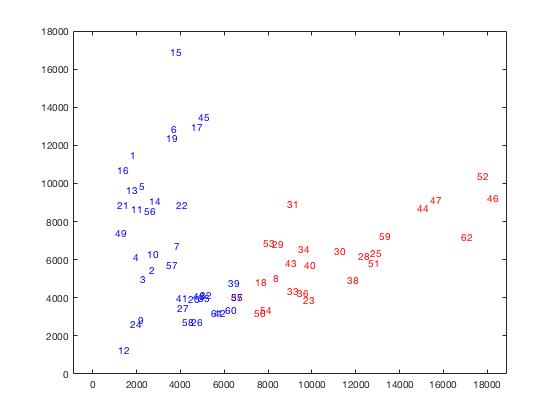
**Observations:**

We can see from the error and scatter plots that K mean with correlation distance gives good result but Euclidean distance mess up the condense area of clusters that ruined our result.

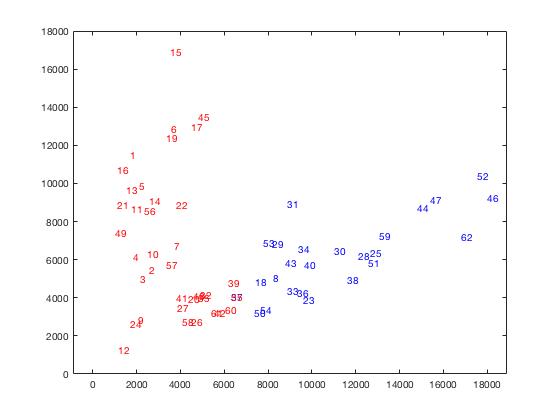
**Tree view:**

**SOM Result:**

100 Iteration Scatter plot**:** 1000 Iterations:



Error:30.64 Error:30.64 (same as last one)

10000 Iterations(Same result) Error:30.64 (same as last one)

**Conclusion:**

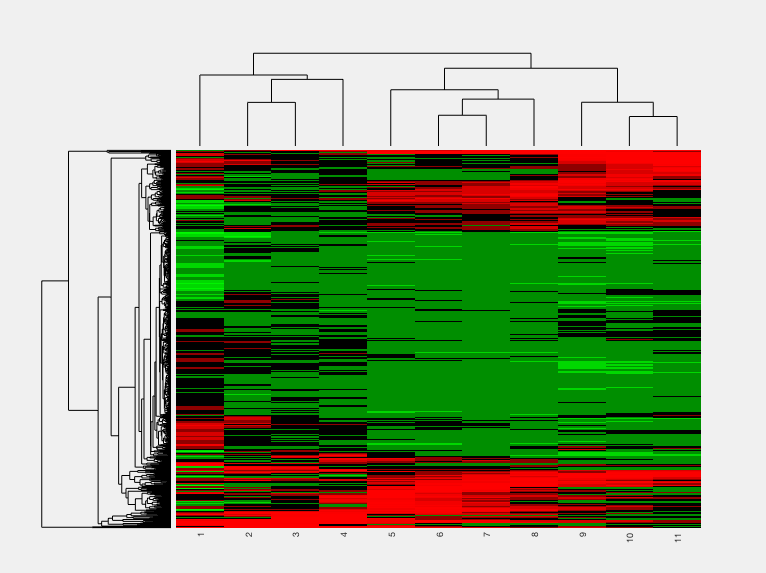
|  |  |  |
| --- | --- | --- |
| **Method of analysis** | **Distance** | **Error %** |
| KMC | Correlation | 12.9 |
| KMC | sqeuclidean | 46.7 |
| SOM | Iteration:100 | 30.64 |
| SOM | Iteration:1000 | 30.64 |
| SOM | Iteration:10000 | 30.64 |
|  |  |  |
| HC single linkage | correlation | 67.74 |
| HC avg linkage | Correlation | 16.12 |
| HC complete Linkage | Correlation | 16.12 |
| HC single linkage | euclidean | 66.12 |
| HC avg linkage | euclidean | 33.81 |
| HC complete Linkage | euclidean | 56.45 |

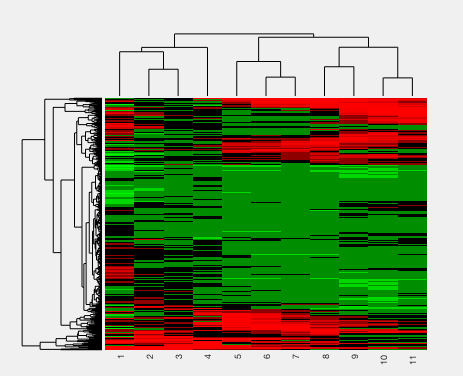
**Conclusion Unsupervised clustering:**

We can see from the table KMC with distance correlation gives better results than SOM, the reason being it’s unsupervised learning and results are unknow so it didn’t affect the end result. On the other hand KMC with correlation gives good result that are comparable with HC with correlation with average and complete linkage and this close error can lead me to these two clustering techniques can be preferred in this case.

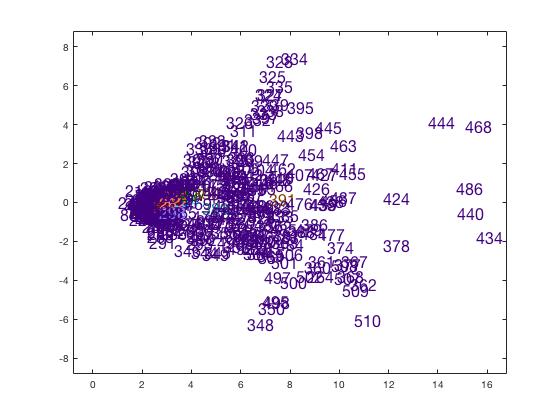
Although we get not very good result in this case that might be due to data samples because for healthy genes we have only 22 samples that is not big data set. But for the given set of conditions and based on the observations we can say KMC with HC will produce good results.

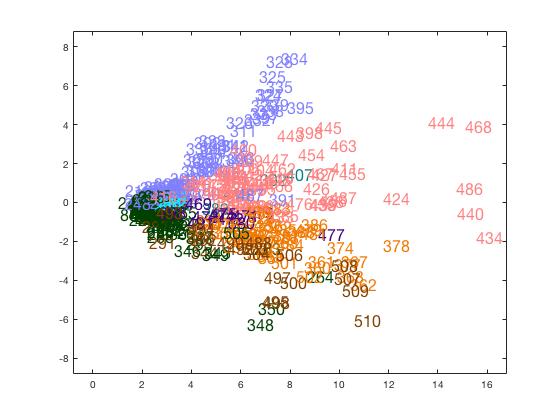
**Part-3 Gene Based Analysis:**

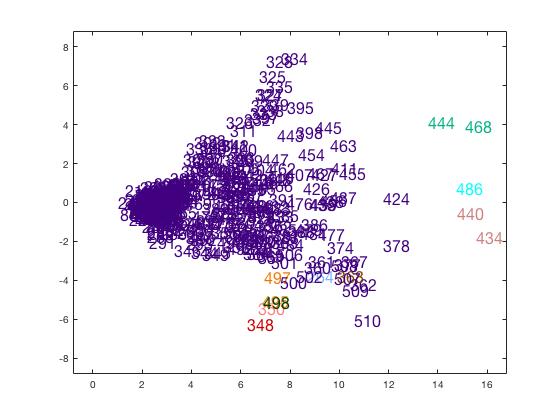
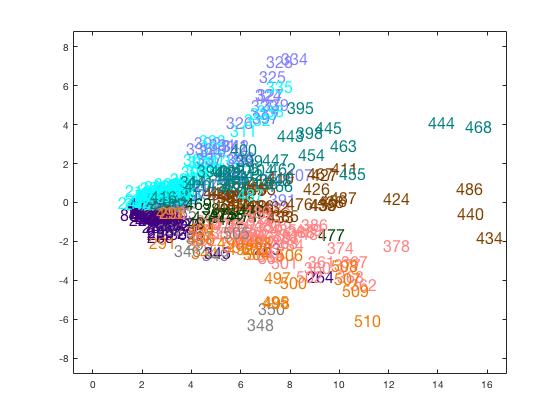
**Visulization of data:** **Average Linkage**

**Single Linkage:**

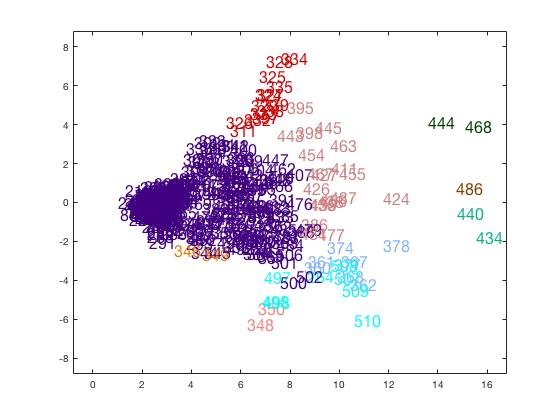
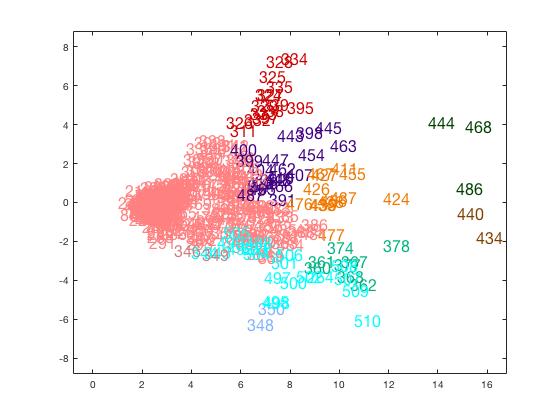
**Result of HC,KMC and SOM**

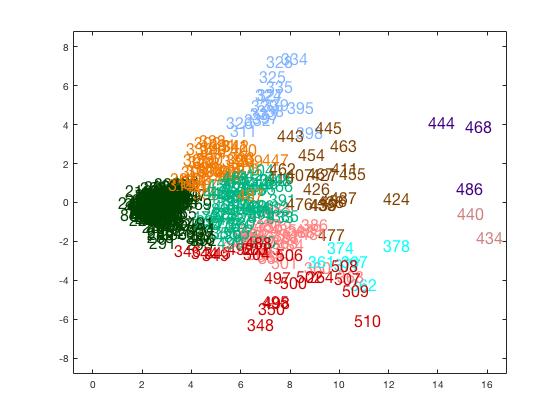
HC Correlation Single Linkage HC Correlation average Linkage

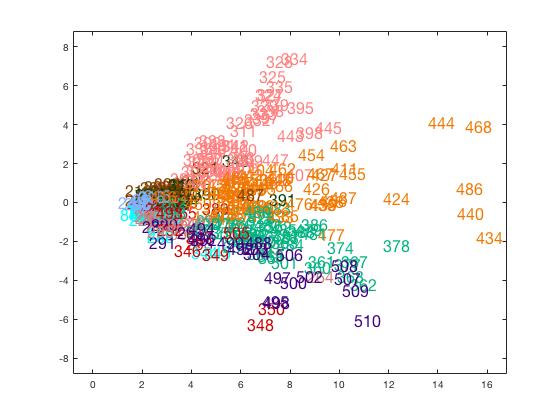


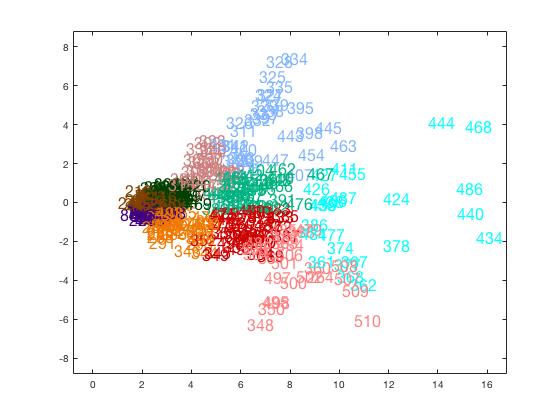
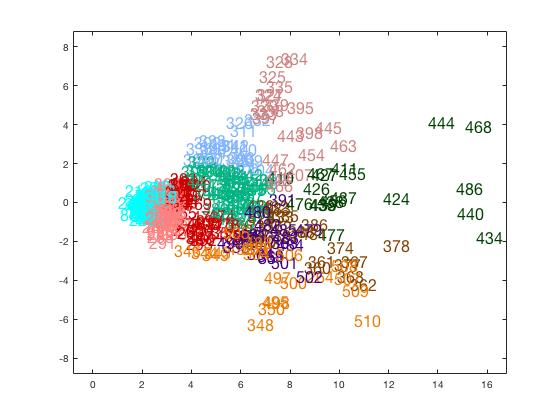
HC Correlation Complete Linkage HC Euclidean Sinle Linkage

HC Euclidean complete Linkage HC Euclidean average Linkage



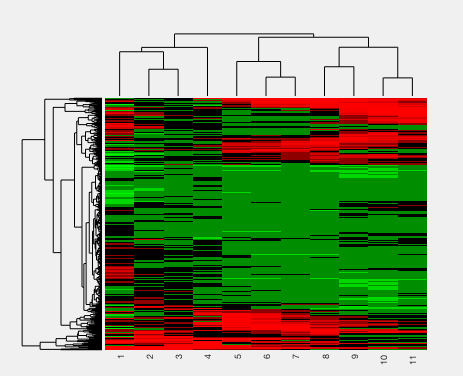
**KMC with Euclidean Distance KMC with correlation Distance:**

**SOM:**

**Iterations:100 Iterations:1000**

From all these scatter plots we can see the data is really dense so it becomes really hard to to decide the clusters but these scatter plots showing there are probably 4-5 branches and then there is some dense data(4-5 clusters) in the middle. If we compare k mean with HC K mean cluster showing much better view of clusters than Hier archieal (HC) cluster but complete linkage of both distances showed better result.

If we look at SOM results that seems pretty logical clustering to me specially for scattered branches and also for dense data points and it somehow agrre with K mean Euclidean distance too that support the concurrency in both so can be relied.

**Tree View:**

**Appendix:**

**Matlab Script:**

data50 = load('alonSigncnt50.txt');

data50 = data';

[rows, cols] = size(data50);

%supervised classifier

I\_label = zeros(**rows**,1);  
I\_label(1:22,:) = 1;  
I\_label(23:62,:) = 2;

Train\_idx = [1:11 23:42];  
Test\_idx = [12:22 43:62];

Train\_data =data50(Train\_idx, :);  
Train\_label = I\_label(Train\_idx, :);  
Test\_data = data50(Test\_idx, :);  
Test\_label =I\_label(Test\_idx, :);

mdl = ClassificationKNN.fit(Train\_data, Train\_label, 'NumNeighbors', 5, 'Distance', 'correlation')

prediction\_labels = predict(mdl, Test\_data);

error = nnz(Test\_labels - prediction\_labels) / length(Test\_label) \* 100;

%scatter plot

PC = princomp(Test\_data );

ts502D = Test\_data \* PC(:, 1:2);

test\_labels\_and\_data = [prediction\_labels ts502D];

plotClass2D(test\_labels\_and\_data, 1, 12);

%Unsupervised Clusters

data50 = load('alon\_g100.txt');

labels = zeros(62, 1);

dist = pdist(data50, 'correlation');

tree = linkage(dist, 'single');

clusters = cluster(tree, 'maxclust', 2);

labels(1:22, 1) = 2;

labels(23:62, 1) = 1;

error = nnz(labels - clusters)/length(labels)\*100

dendrogram(tree);

clustergram(data50, 'linkage', single, 'standardize', 'col');

%scatter Plot

PC = princomp(data50 );

ts502D = data50 \* PC(:, 1:2);

ts502Dclass = [zeros(62,1) ts502D];

ts502Dclass(1:62,1)=clusters;

plotClass2D(ts502Dclass, 1, 10);

%K Mean

clusters = kmeans(data50, 2, 'dist', 'correlation');

labels(1:22, 1) = 2;

labels(23:62, 1) = 1;

error = nnz(labels - clusters)/length(labels)\*100

%SOM

Q=data50;

Q=Q';

net = newsom([min(Q,[],2) max(Q,[],2)], [2 1], 'hextop');

net.trainParam.epochs = 100;

net = train(net, Q);

clusters = vec2ind(net(Q));

clusters =  clusters'

labels(1:22, 1) = 1;

labels(23:62, 1) = 2;

error = nnz(labels - clusters)/length(labels)\*100