Data Mining and Machine Learning in Bioinformatics

Exercise Series 7

Group members (Name, Student ID, E-Mail):

- Baldomero Valdez, Valenzuela, 2905175, baldmer.w@gmail.com
- Omar Trinidad Gutierrez Mendez, 2850441, omar.vpa@gmail.com
- Shinho Kang, 2890169, wis.shinho.kang@gmail.com

Task 1:

A) perform an `unpaired t-test` on colonCA (____/2)

Code

```
data(colonCA)
colon.ds = log(exprs(colonCA))

pvalues <- apply(colon.ds, 1, function(x) {
  return (t.test(x[colonCA$class=='t'], x[colonCA$class=='n'])$p.value)
})
alpha = 0.0001
pvalues[pvalues <= alpha]
```

Result

```
Hsa.8125  4.03E-05      Hsa.3331  2.61E-05
Hsa.957   3.42E-05      Hsa.549   5.12E-07
Hsa.8147  1.73E-05      Hsa.462   1.62E-05
Hsa.821   3.36E-05      Hsa.3016  9.91E-05
Hsa.36689 5.10E-06      Hsa.1832  1.03E-05
Hsa.5971  1.48E-05      Hsa.2928  4.53E-06
Hsa.37937 2.51E-07      Hsa.2097  1.17E-05
Hsa.831   1.51E-06      Hsa.627   4.00E-06
Hsa.6472  5.49E-05      Hsa.2645  8.72E-05
Hsa.3306  9.81E-06      Hsa.601   3.19E-05
Hsa.773   1.95E-05      Hsa.6814  2.33E-06
Hsa.36952 6.27E-06      Hsa.2291  7.39E-05
```

B) Install EMA package, do clustering and produce heatmaps (____ /4)

Install EMA package is a tedious process, first install biomaRt dependencies:

- Ubuntu: sudo apt-get install libcurl4-openssl-dev libxml2-dev
- install.packages("XML")
- install.packages("RCurl")

After that install EMA dependencies and EMA library:

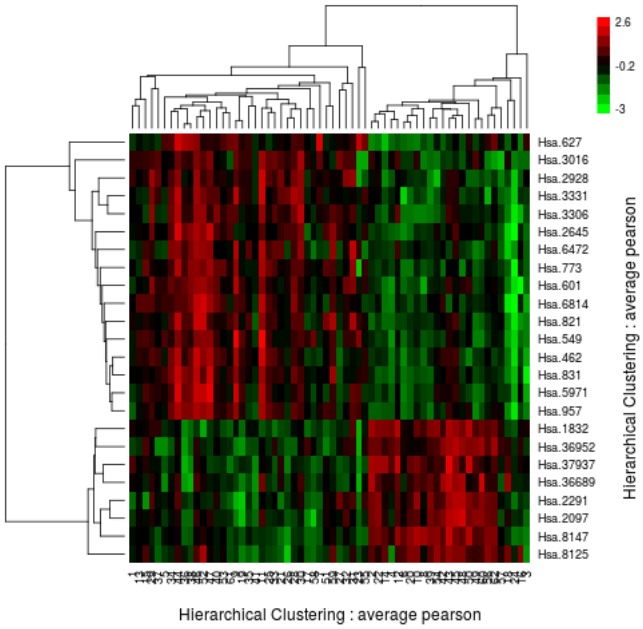
- source("https://bioconductor.org/biocLite.R")
- biocLite("biomaRt")
- biocLite("affy")
- biocLite("siggenes")
- biocLite("gcrma")
- biocLite("AnnotationDbi")
- install.packages("EMA")

code

```
diff.exp = colon.ds[pvalues <= alpha,]

cluster.complete.genes = clustering(diff.exp, metric="pearson",  method = "complet
e")
cluster.complete.samples = clustering(t(diff.exp), metric="pearson",  method = "co
mplete")

cluster.average.genes = clustering(diff.exp, metric="pearson", method = "average")
cluster.average.samples = clustering(t(diff.exp), metric="pearson", method = "aver
age")

cluster.ward.genes = clustering(diff.exp, metric="pearson", method = "ward")
cluster.ward.samples = clustering(t(diff.exp), metric="pearson", method = "ward")
```

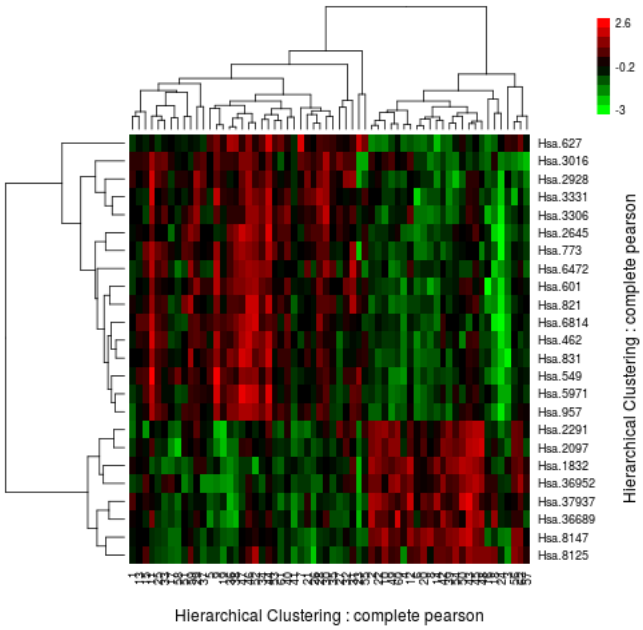
Heatmap for cluster with average linkage

```
clustering.plot( tree = cluster.average.genes, tree.sup = cluster.average.samples,
 data = diff.exp)
```



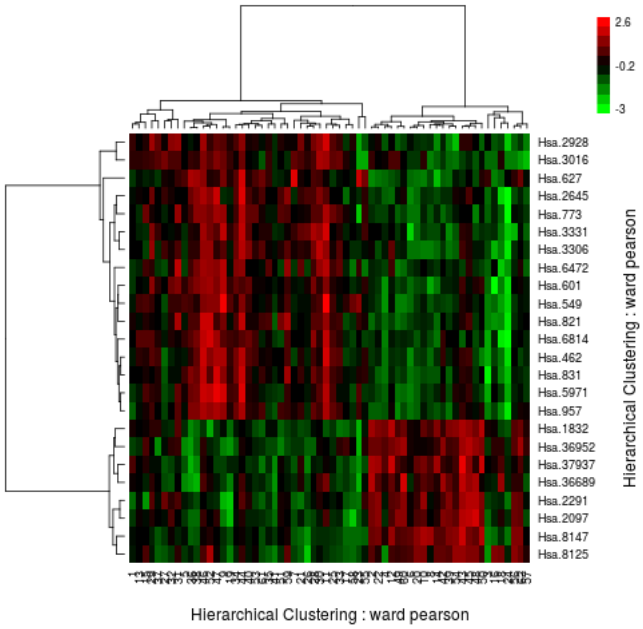
Heatmap for cluster with complete linkage

```
clustering.plot( tree = cluster.complete.genes, tree.sup = cluster.complete.sample
s, data = diff.exp)
```



Heatmap for cluster with Ward's method

```
clustering.plot( tree = cluster.ward.genes, tree.sup = cluster.ward.samples, data
= diff.exp)
```



The dendogram visualization shows that clustering with average and complete linkage produce a similar result. The dendogram for Ward's method shows that this clustering grouping produce fewer levels.

C) Comparison between Pearson correlation and Euclidean distance (____ /2)

When clustering colonCA data it is better to use Pearson correlation as distance because we are comparing the similarity between several dimensions. Another advantage from Pearson correlation over Euclidean distance: is unit independent.

Task 3:

A) Which behavior would you expect, if you plot the distortion in dependency of the number of clusters?

The distortion is sum of the square of distances(difference of points) between each point and its cluster center. So, as the number of cluster increases the distortion decreases.

B) Consider k relatively well separated clusters, which can be embedded into spheres. Would GMM clustering yield a very different result than k-means?

Gaussian mixture models can reach the maximization of likelihood using the EM algorithm. It computes parameters of each gaussian by several iterations. Assume that GMM is looking for k gaussians and the given environment, it should find similar result as k-means clustering.

Task 4:

Which of the clustering algorithms discussed in the lecture so far can you apply to cluster biological sequences? Give reason for your answer!

Hierarchical clustering is widely used for detecting clusters in genomic data. It generates a set of partitions forming a cluster hierarchy. According to linkage criteria, there are three hierarchical clustering methods including single-linkage clustering (SL), complete-linkage clustering (CL) and average-linkage clustering (AL). With SL, clusters may be merged together due to single sequences being close to each other, even though many of the sequences in each cluster may be very distant to each other. CL tends to find compact clusters of approximately equal diameters. With CL, all objects in a cluster are similar to each other. AL can be seen as an intermediate between single and complete linkage clustering, resulting in more homogeneous clusters than those obtained by the single-linkage method. Hierarchical approaches may yield fairly good results, but they require the similarity of all pairs of sequences and quickly arrive at a bottleneck in terms of computational time and memory usage for large-scale data sets.

On the other side, K-means (KM) is a commonly used method of partitional clustering methods. KM has a lower order of computational complexity and demands less physical memory than the hierarchical method. It is suitable for clustering large gene data. The major drawback of KM compared to hierarchical clustering algorithms is the lack of hierarchical relationships in its results.