

---

## Laboratory 2:0 Topics in analysis of microarray data: clustering and discrimination

Day 2: Day August 17, 2004: 15:45 – 17:15

Ben Bolstad, Biostatistics, University of California, Berkeley

---

### Key Concepts

- Clustering
- Discrimination

### What you will be able to do at end of this section

- Perform clustering on a dataset using R/BioC software.

### Introduction

This lab introduces you to some basic preprocessing tools for cDNA microarray data. We will make use of the BioConductor software package *marray*.

### First steps

Start R. At the command prompt type

```
library(cluster)
```

```
library(multtest)
```

This will load the cluster and multtest libraries. In this case we are using the multtest library because it contains the dataset that we will use in this lab.

Next load the data and find out more about it.

```
data(golub)
```

```
?golub
```

### Partitioning methods

First consider clustering genes using the kmeans algorithm.

```
Kmeans1 <- kmeans(golub,2)
attributes(Kmeans1)
Kmeans1$cluster
```

We could also consider clustering samples rather than genes. Do this by typing

```
Kmeans2 <- kmeans(t(golub),2)
Kmeans2$cluster
golub.cl
```

Another method is partitioning around medoids

```
Pam1 <- pam(golub,k=2,diss=FALSE)
attributes(Pam1)
Pam1$clustering
```

You can use different distance measures to cluster upon.

```
Pam2 <- pam(dist(golub,method="euclidean"),k=2,diss=TRUE)
```

Read the documentation for this functions and experiment.

```
?pam
?dist
```

## Hierarchical Clustering

The main function for hierarchical clustering in R is `hclust`. The first thing you should do is read the documentation by typing `?hclust`. We will only consider clustering samples for this lab to avoid any problems with computational speed. The following commands will carry out the clustering, draw the dendrogram and then cut the tree into two groups.

```
Hclust1 <- hclust(as.dist(1-abs(cor(golub))))
```

```
plot(Hclust1)
cutree(Hclust1, k=2)
```

what do you observe?

We could use a different distance metric. Now what do you observe?

```
Hclust2 <- hclust(dist(t(golub), method="manhattan"))
plot(Hclust2)
cutree(Hclust2, k=2)
```

## Using silhouettes to choose the number of clusters

First read the documentation on silhouette

```
?silhouette
```

Now lets try pam with different numbers of clusters.

```
par(mfrow=c(2,3))
for (k in 2:7){
  PamSil <- pam(t(golub), k=k, diss=FALSE)
  print(summary(silhouette(PamSil))$avg.width)
}
```

How many clusters do you think is sensible for this data? Does the number you get using the silhouettes agree?

## Heatmaps

Read the documentation about the heatmap function by typing `?heatmap`. Lets try hierarchical clustering of both genes and arrays. For speed we will use a subset of the genes.

```
library(marray)
```

```
heatmap(golub[1:200,], col=maPalette(50, low="green", high="red", mid="white"))
```

Consider clustering arrays on some correlation metric. This can also be visualized using a heatmap. Type

```
heatmap(1-  
abs(cor(golub)), col=maPalette(50, low="green", high="red", mid="white"))
```

You should experiment with different distance measures.

## Filtering

Up to this point we have not considered the effect that filtering might have on our clustering. We have just clustered all the data. Suppose that we think that the most variable genes are the ones that are important. Lets find the top 200 most variable genes. And create a heatmap based upon them.

```
vars.xsamples <- apply(golub, 1, var)  
heatmap(golub[rank(-  
vars.xsamples)<200,], col=maPalette(50, low="green", high="red", mid="white"))
```

Do you see more or less structure than before?

If you have time, consider filtering on a t-statistic and observing what happens to your heat maps.

## Appendix

### 1. Resources

#### i) Original Papers

- Alizadeh et al, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 2000

- Golub et al, "Molecular Classification of Cancer: Class Discovery and Class prediction by Gene Expression Monitoring ", Science, 1999
- Dudoit, et al, :Comparison of discrimination methods for the classification of tumors using gene expression data, JASA, 2002
- Dudoit and Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset ", Genome Biology, 2002
- Dudoit and Fridlyand, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics, 2003

**ii) Software**

- [www.biconconductor.org](http://www.biconconductor.org)
- [www.r-project.org](http://www.r-project.org)
- <http://rana.lbl.gov/EisenSoftware.htm>

**iii) Text books:**

- Hastie, Tibshirani, Friedman "The Elements of Statistical Learning", Springer, 2001
- Speed (editor) "Statistical Analysis of Gene Expression Microarray Data", Chapman & Hall/CRC, 2003

**2. Demonstration Overheads**

© 2004 Canadian Genetic Diseases Network