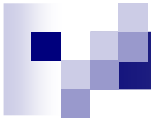


Finding patterns in genes/samples: Clustering Methods for Class Discovery

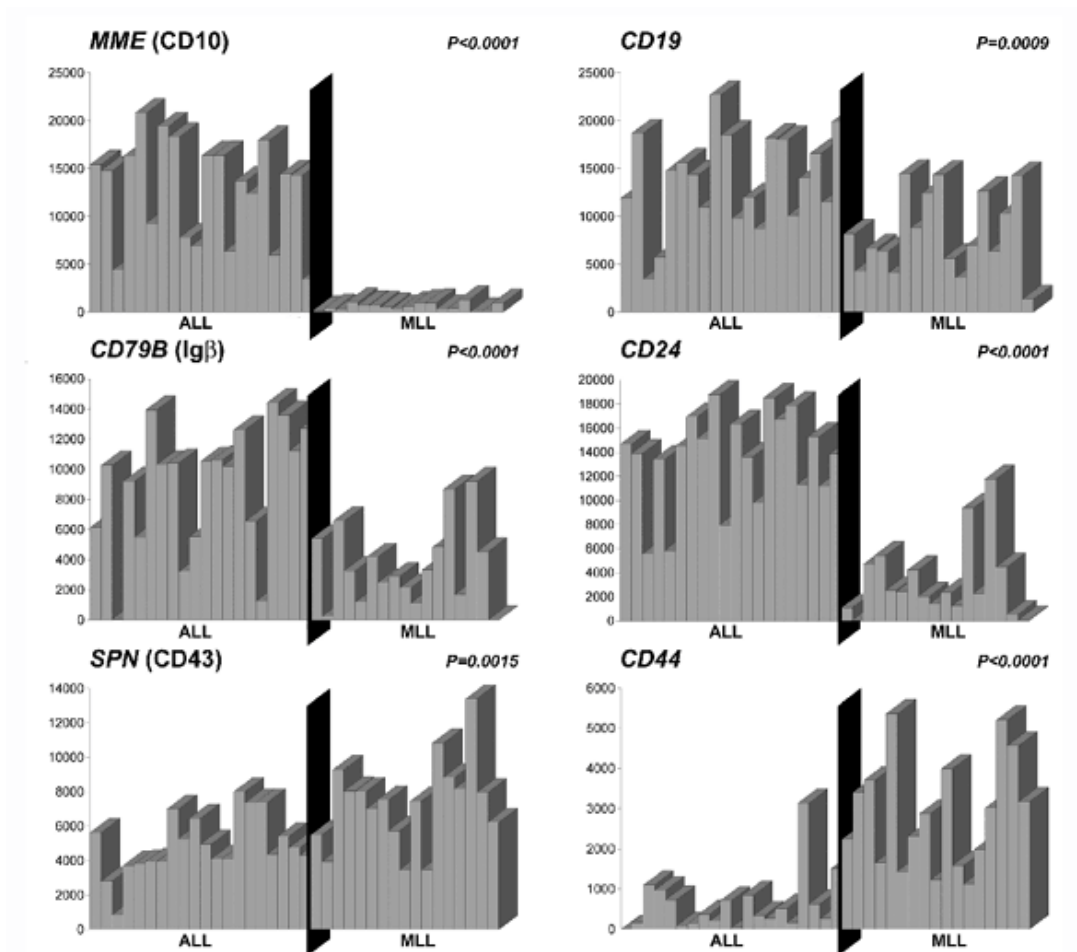
Alex Sánchez

Statistics Department. University of
Barcelona



Introduction

Gene expression profiles



- Many genes show definite changes of expression between conditions
- These patterns are called *gene profiles*



The problem of finding patterns

- In a microarray experiment it is common to have
 - Groups of genes showing coordinate changes across conditions
 - This may suggest co-expression, co-regulation, etc
 - Groups of samples showing similar expression patterns
 - This may suggest the existence of groups in samples
 - This may be useful to detect outliers or batch effects when “unexpected” groupings appear
- The goal of this chapter is:
 - Learn to find patterns in genes and/or samples.
 - Learn to distinguish *interesting* or *real* patterns from *meaningless variation*, at the level of the gene.



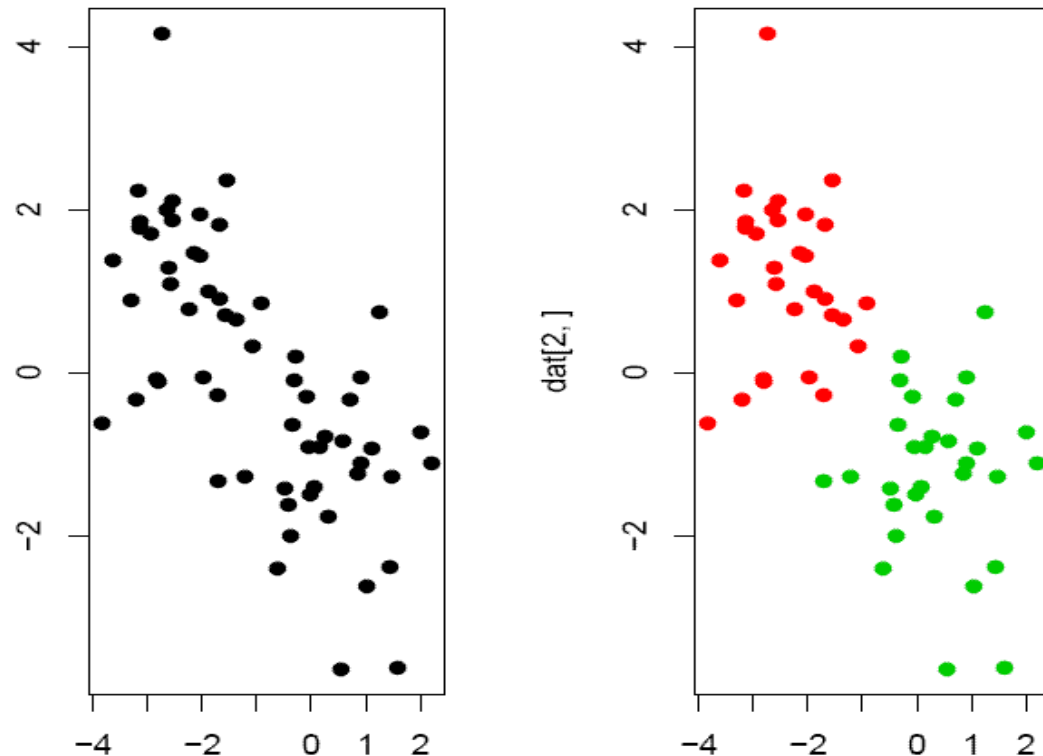
Finding patterns: Two approaches

- If patterns already exist we may be interested in Profile comparison (Distance analysis):
 - Find the genes whose expression fits specific, predefined patterns.
 - Find the genes whose expression follows the pattern of predefined gene or set of genes.
- If we wish to discover new patterns we may be interested in carrying out some kind of exploratory analysis to see what expression patterns emerge.
- This is the goal of: *Cluster analysis (class discovery)*

Clustering: grouping objects by their similarity

Cluster: a set of objects that are similar to each other and separated from the other objects.

By forming clusters one can **discover** groups in data.



Example: green/red data points were generated from two different normal distributions

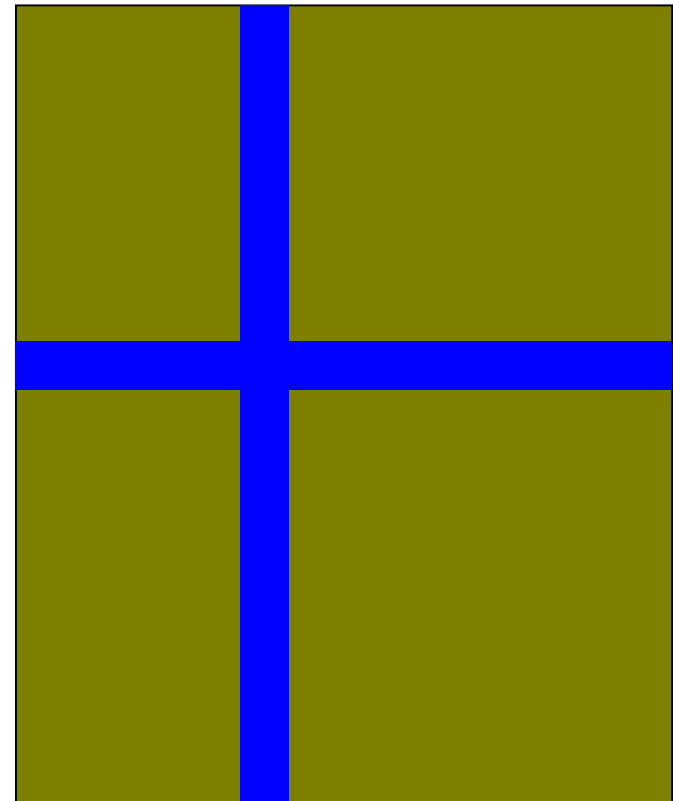
Clustering microarray data

- In a gene expression matrix
 - **rows represent genes and**
 - **columns represent measurements**from different experimental conditions measured on individual arrays.
- The values at each position in the matrix characterise the expression level (absolute or relative) of a particular gene under a particular experimental condition

gene expression data matrix

n experiments

p genes





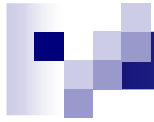
Clustering microarray data

- Cluster can be applied to genes (rows), mRNA samples (cols), or both at once.
- Cluster samples to
 - identify new classes of biological (e.g. cell or tumour) subtypes.
 - Identify problems such as batch effects or outliers
- Cluster rows (genes) to
 - identify groups of possibly co-regulated genes.
 - Identify spatial or temporal patterns (e.g. in cell cycle or analysis of different brain areas)
 - Reduce redundancy e.g. for variable selection in predictive models.



Advantages of clustering

- Clustering leads to readily interpretable figures.
- Clustering strengthens the signal when averages are taken within clusters of genes (Eisen).
- Clustering can be helpful for identifying patterns in time or space..
- Clustering is useful, perhaps essential, when seeking new subclasses of cell samples (tumors, etc).



Drawbacks of clustering

- It is an exploratory technique
 - Usually no significance available.
- Different approaches often yield different groupings
 - Difficult to decide which is the best (or the “real”)
 - Difficult to avoid the temptation of selecting the grouping that best fits our hypotheses.
- Any dataset can be clustered
 - Difficult to decide if clustering is real or random.

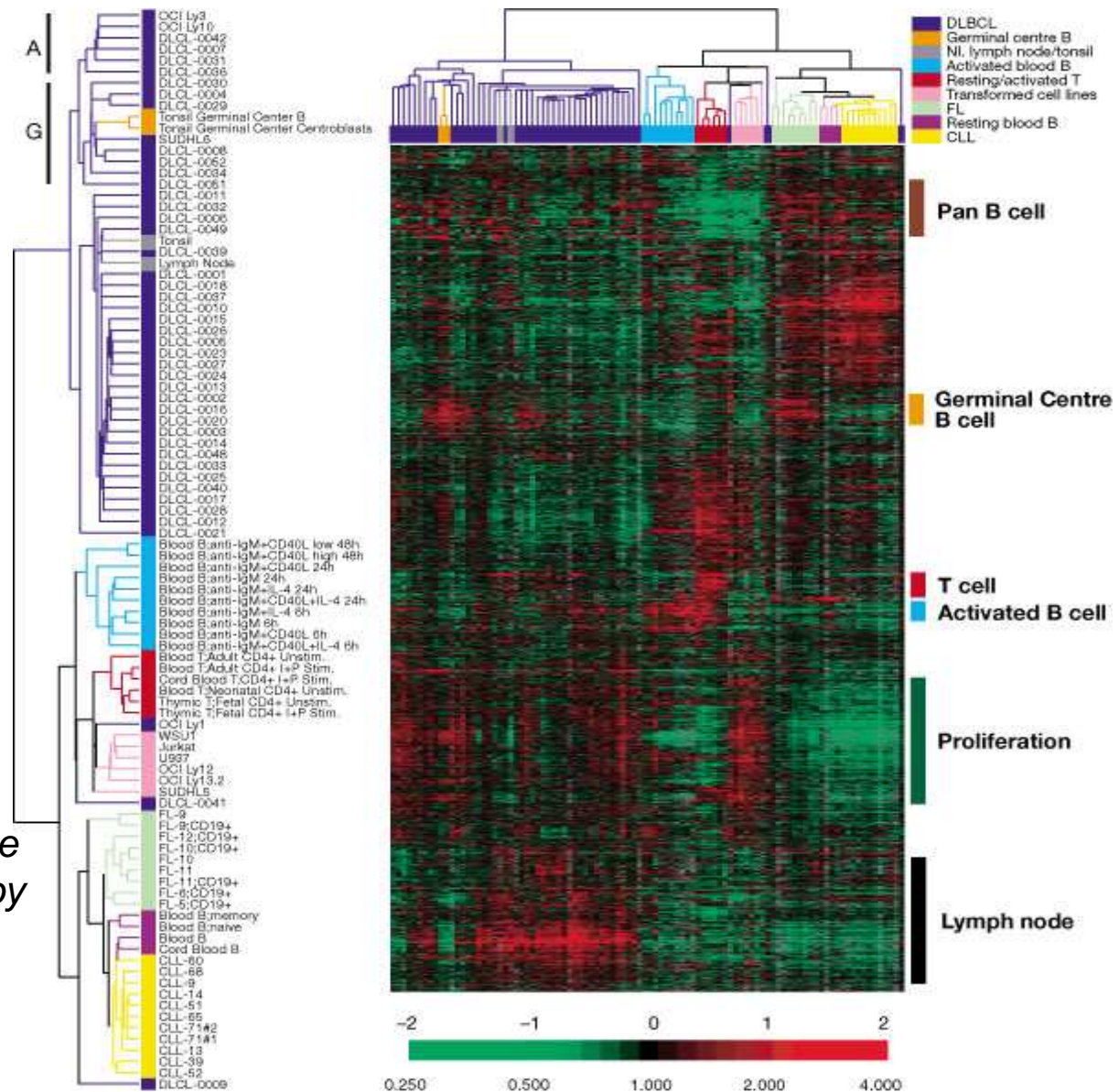


Applications of clustering (1)

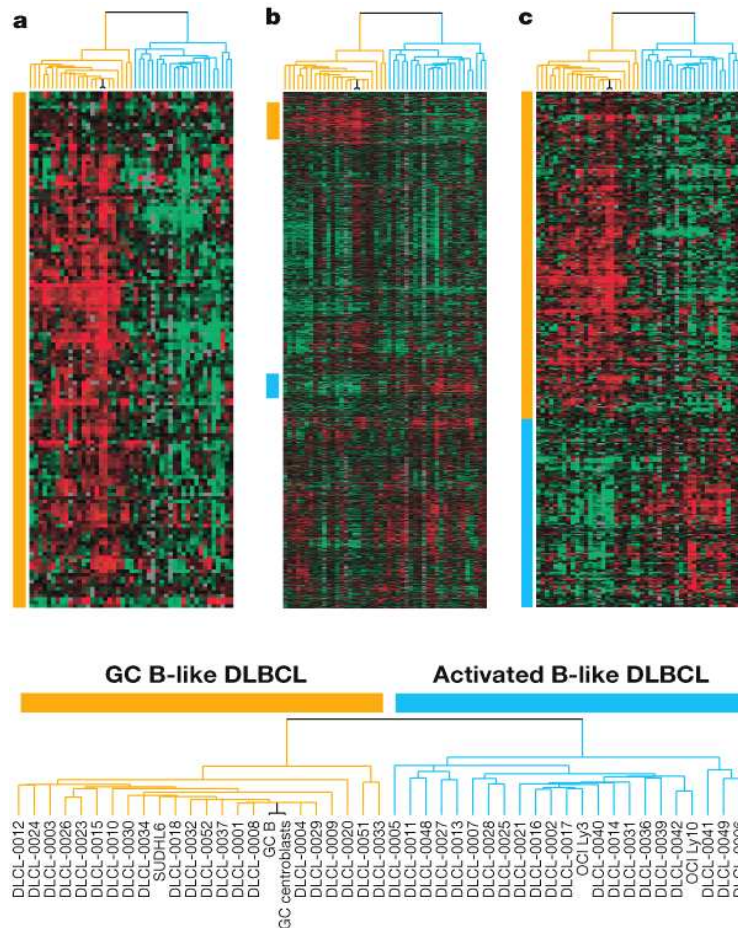
- Alizadeh et al (2000) *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.*
 - Three subtypes of lymphoma (FL, CLL and DLBCL) have different genetic signatures.
 - (81 cases total)
 - DLBCL group can be partitioned into two subgroups with significantly different survival. (39 DLBCL cases)

Clusters on both genes and arrays

Taken from
Nature February, 2000
Paper by Allizadeh. A *et al*
Distinct types of diffuse large B-cell lymphoma identified by Gene expression profiling



Discovering tumor subclasses

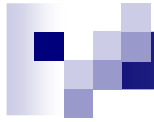


- DLBCL is clinically heterogeneous
- Specimens were clustered based on their expression profiles of GC B-cell associated genes.
- Two subgroups were discovered:
 - GC B-like DLBCL
 - Activated B-like DLBCL



Applications of clustering (2)

- A naïve but nevertheless important application is assessment of experimental design
- If one has an experiment with different experimental conditions, and in each of them there are biological and technical replicates...
- *We would expect* that the more homogeneous groups tend to cluster together
 - Tech. replicates < Biol. Replicates < Different groups
- Failure to cluster so suggests bias due to experimental conditions more than to existing differences.



Basic principles of clustering



Basic principles of clustering

- **Aim:** to group observations that are “similar” based on predefined criteria.
- **Issues:**
 - Which genes / arrays to use?
 - Which similarity or dissimilarity measure?
 - Which clustering algorithm?
- It is advisable to **reduce** the number of genes from the full set to some more manageable number, before clustering.
- The basis for this reduction is usually quite context specific, see later example.



Cluster analysis

- Generally, cluster analysis is based on two ingredients:
 - **Distance measure:**
Quantification of (dis)similarity of objects.
 - **Cluster algorithm:**
A procedure to group objects. Aim: small within-cluster distances, large between-cluster distances.

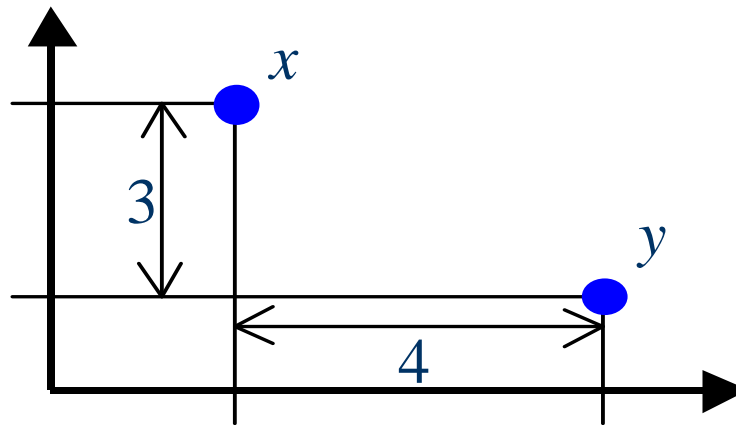


Some distance measures

Given vectors $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n)$

- Euclidean distance:
$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$
- Manhattan distance:
$$d_M(x, y) = \sum_{i=1}^n |x_i - y_i|.$$
- Correlation distance:
$$d_C(x, y) = 1 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Some example of distances



1, Euclidean distance : $\sqrt{4^2 + 3^2} = 5$.

2, Manhattan distance : $4 + 3 = 7$.

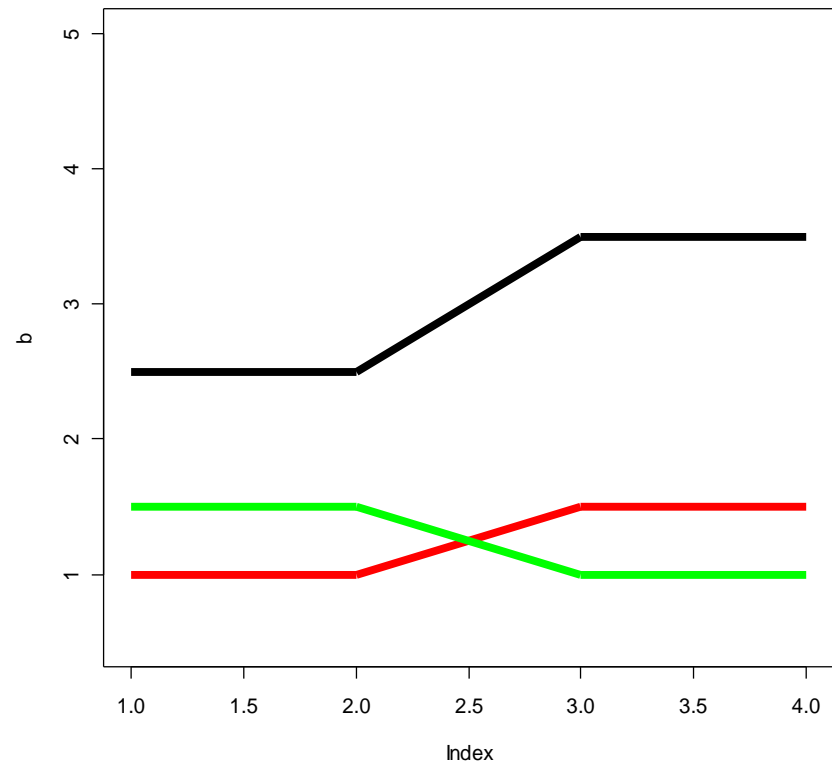
3, "sup" distance : $\max\{4, 3\} = 4$.

Which distance measure to use?

Choice of distance measure should be based on the application area.

- What sort of similarities would you like to detect?
- Correlation distance d_c measures trends/relative differences:

$$d_c(x, y) = d_c(ax + b, y) \text{ if } a > 0.$$



$$\mathbf{x} = (1, 1, 1.5, 1.5)$$

$$\mathbf{y} = (2.5, 2.5, 3.5, 3.5) = 2\mathbf{x} + 0.5$$

$$\mathbf{z} = (1.5, 1.5, 1, 1)$$

$$d_c(\mathbf{x}, \mathbf{y}) = 0, d_c(\mathbf{x}, \mathbf{z}) = 2.$$

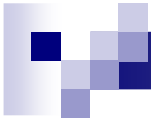
$$d_E(\mathbf{x}, \mathbf{z}) = 1, d_E(\mathbf{x}, \mathbf{y}) \sim 3.54.$$



Which distance measure to use?

- Euclidean and Manhattan distance both measure absolute differences between vectors. Manhattan distance is more robust against outliers.
- May apply **standardization** to the observations: Subtract mean and divide by standard deviation:
- After standardization, Euclidean and correlation distance are equivalent:

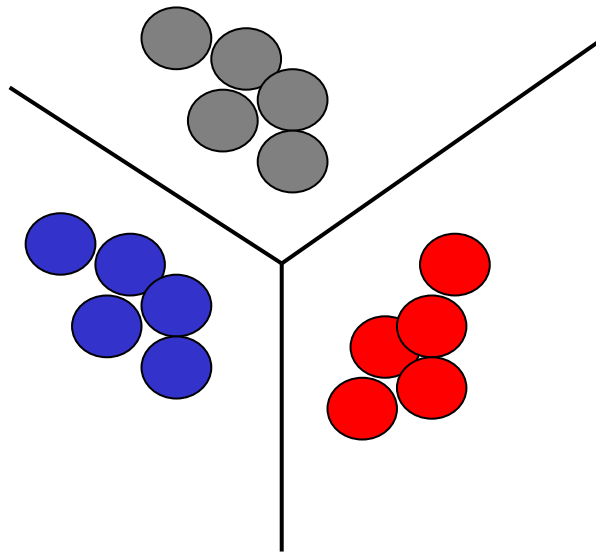
$$d_E(x_1, x_2)^2 = 2nd_C(x_1, x_2).$$



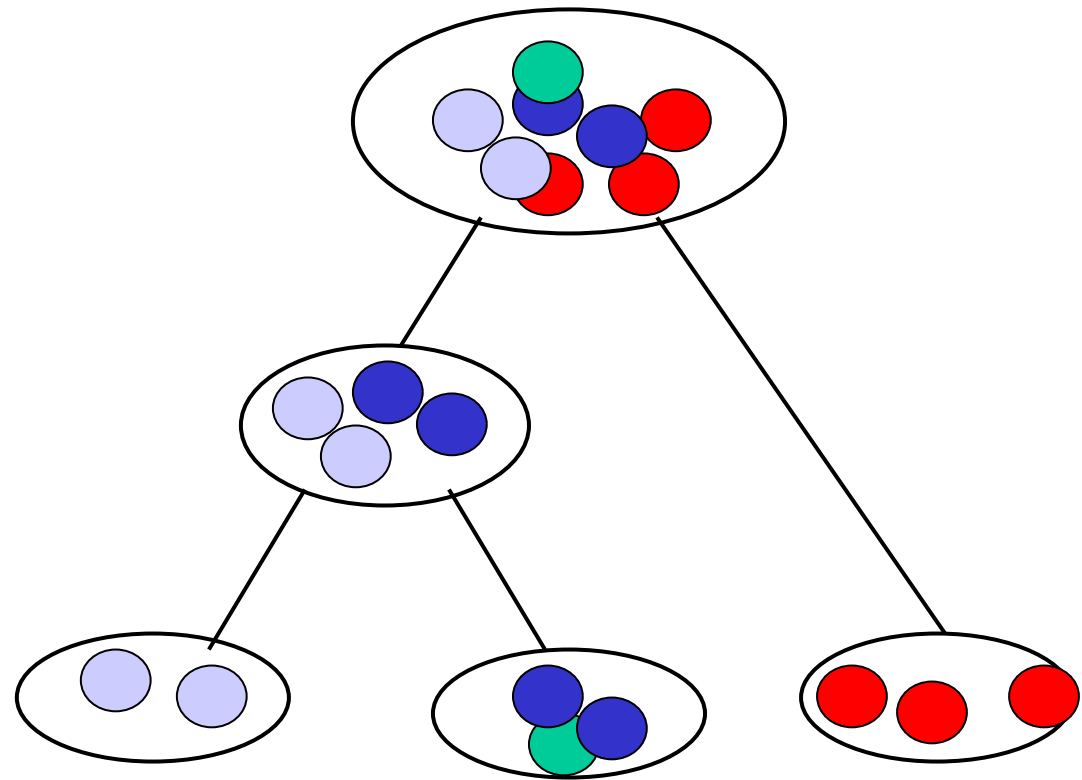
Clustering methods and algorithms

Two basic types of methods

Partitioning



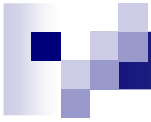
Hierarchical






Partitioning methods

- Partition the data into a pre-specified number k of mutually exclusive and exhaustive groups.
 - Iteratively reallocate the observations to clusters
 - until some criterion is met, e.g. minimize within
 - cluster sums of squares.
- *Examples:*
 - k -means, self-organizing maps (SOM), *PAM*, etc.;
 - Fuzzy: needs stochastic model, e.g. Gaussian mixtures.



Hierarchical methods

- Hierarchical clustering methods produce a **tree** or **dendrogram**.
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.
- The tree can be built in two distinct ways
 - bottom-up: **agglomerative** clustering;
 - top-down: **divisive** clustering.



Agglomerative hierarchical clustering

- **Bottom-up** algorithm (top-down (divisive) methods are less common).
- Start with n clusters.
- At each step, merge the two closest clusters using a *measure of between-cluster dissimilarity*, which reflects the shape of the clusters.

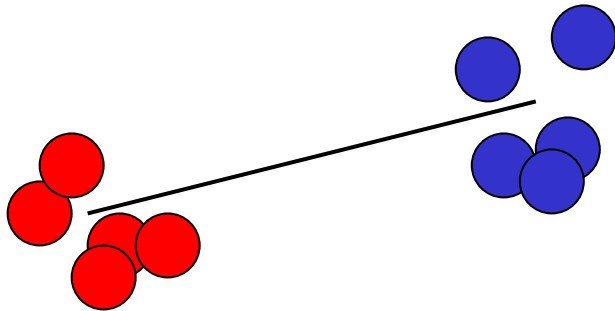


Distances between clusters used for hierarchical clustering

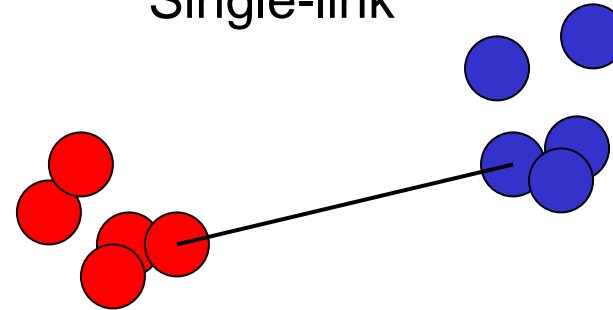
- Calculation of the distance between two clusters is based on the pairwise distances between members of the clusters
 - Mean-link: average of pairwise dissimilarities
 - Single-link: minimum of pairwise dissimilarities.
 - Complete-link: maximum of pairwise dissimilarities.
 - Distance between centroids
- Complete linkage gives preference to compact/spherical clusters.
- Single linkage can produce long stretched clusters.

Between-cluster dissimilarity measures

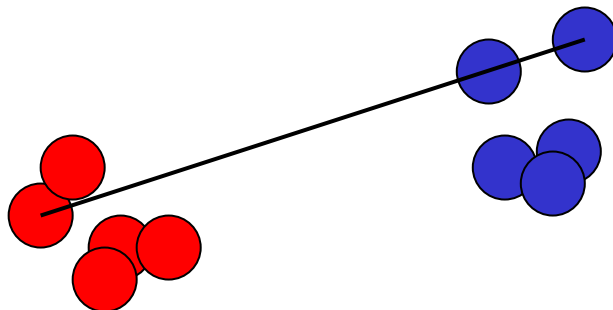
Distance between centroids



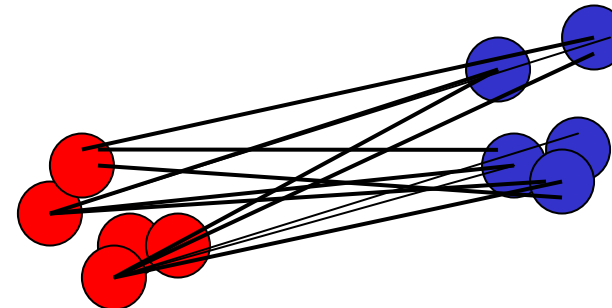
Single-link



Complete-link



Mean-link





Hierarchical divisive methods

- Start with only one cluster.
- At each step, split clusters into two parts.
- Split to give greatest distance between two new clusters
- *Advantages.*
 - Obtain the main structure of the data, i.e. focus on upper levels of dendrogram.
- *Disadvantages.*
 - Computational difficulties when considering all possible divisions into two groups.



Which genes to cluster?

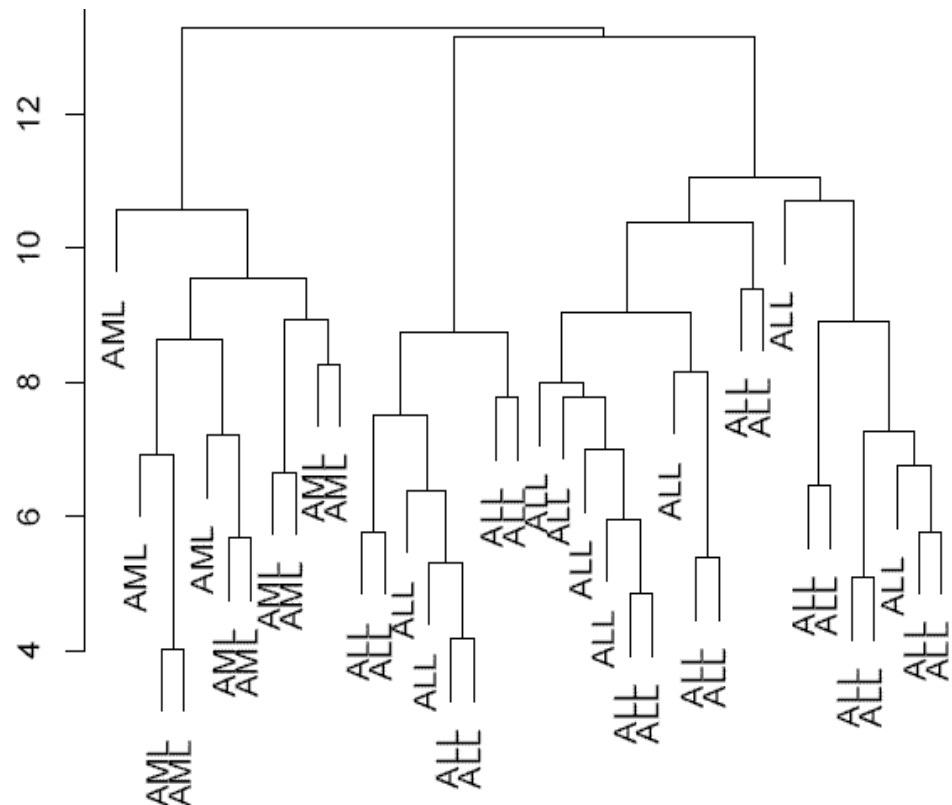
The role of feature selection

- Sometimes, people first select genes that appear to be differentially expressed between groups of samples. Then they cluster the samples based on the expression levels of these genes. Is it remarkable if the samples then cluster into the two groups?
- No, this doesn't prove anything, because the genes were selected with respect to the two groups! Such effects can even be obtained with a matrix of i.i.d. random numbers.

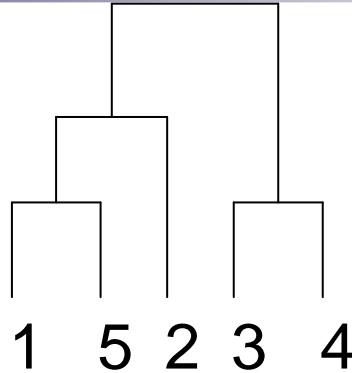
Example of hierarchical clustering

- The height of a node in the dendrogram represents the distance of the two children clusters.
- Loss of information: n objects have $n(n-1)/2$ pairwise distances, tree has $n-1$ inner nodes.
- The ordering of the leaves is not uniquely defined by the dendrogram: 2^{n-2} possible choices.

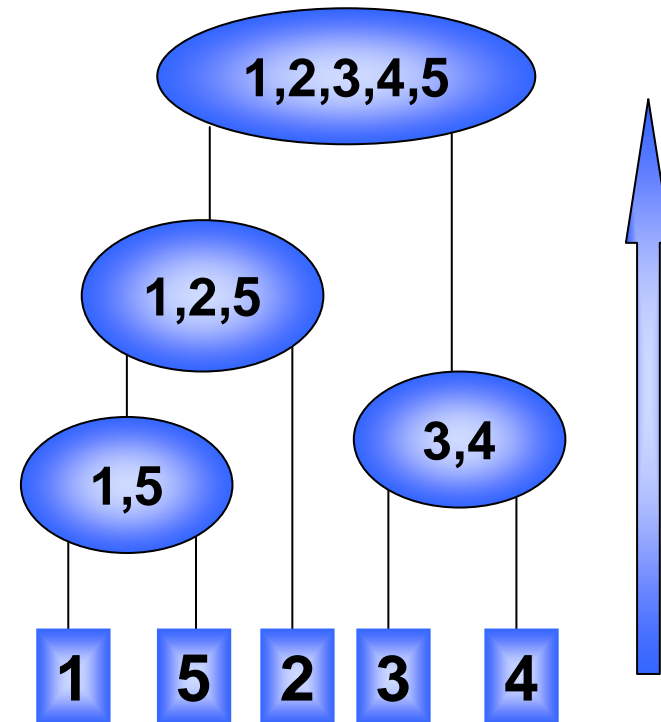
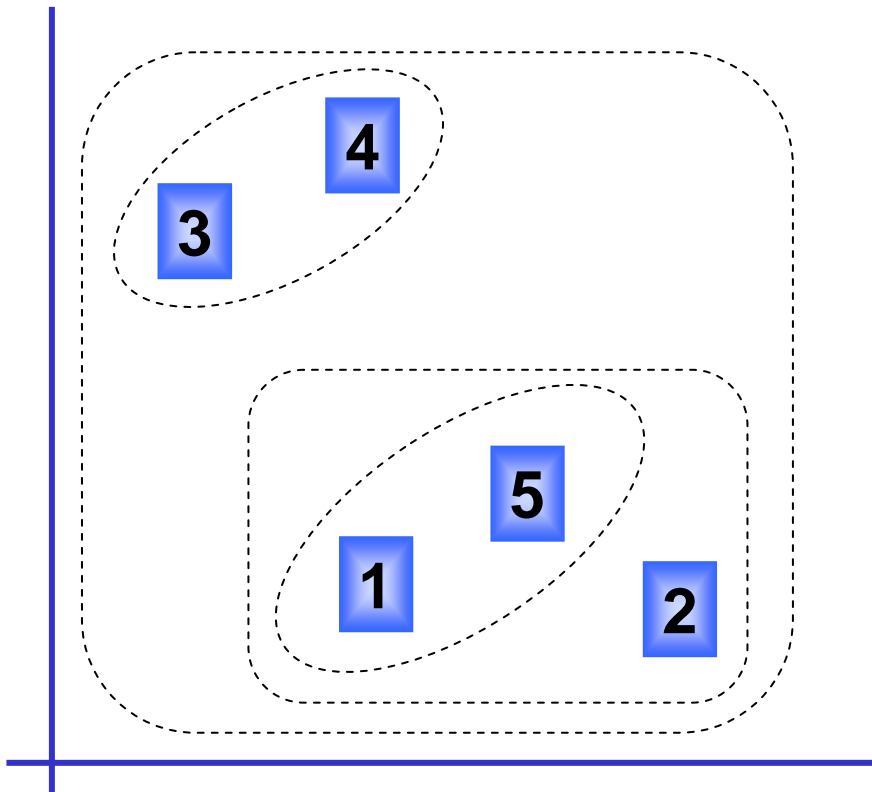
Golub data: different types of leukemia. Clustering based on the 150 genes with highest variance across all samples.

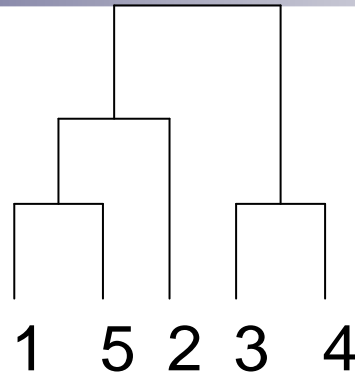
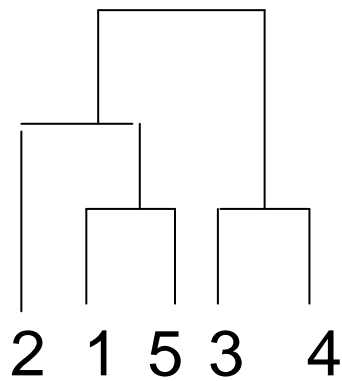


**Illustration of points
In two dimensional
space**

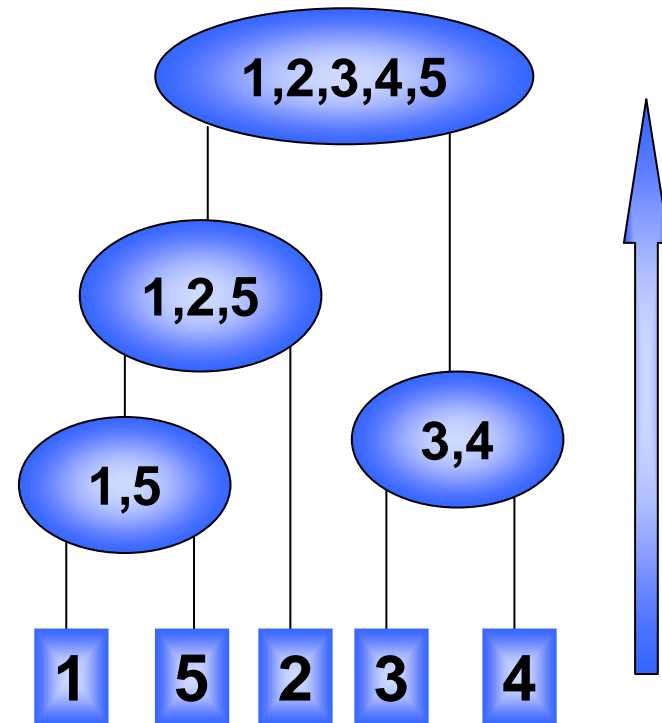
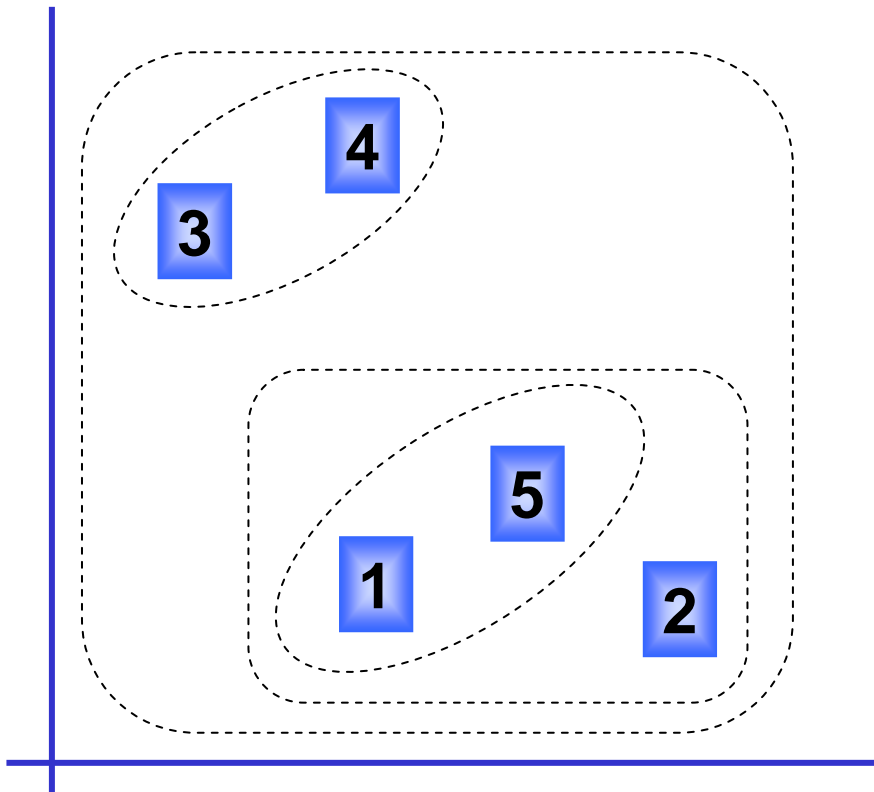


Agglomerative





Agglomerative





K-means clustering

- Input: N objects given as data points in \mathbf{R}^p
- Specify the number k of clusters.
- Initialize k cluster centers. Iterate until convergence:
 - Assign each object to the cluster with the closest center (wrt Euclidean distance).
 - The centroids/mean vectors of the obtained clusters are taken as new cluster centers.
- K -means can be seen as an optimization problem:
Minimize the sum of squared within-cluster distances,

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

- Results depend on the initialization. Use several starting points and choose the “best” solution (with minimal $W(C)$).

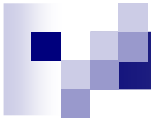


Partitioning around medoids (PAM)

- K -means clustering is based on Euclidean distance.
- Partitioning around medoids (PAM) generalizes the idea and can be used with any distance measure d (objects x_i need not be vectors).
- The cluster centers/prototypes are required to be observations: $m_j = x_{i_j}, j = 1, \dots, K$.
- Try to minimize the sum of distances of the objects to their cluster centers,

$$\sum_{i=1}^n d(x_i, m_{j(i)}),$$

using an iterative procedure analogous to the one in K -means clustering.

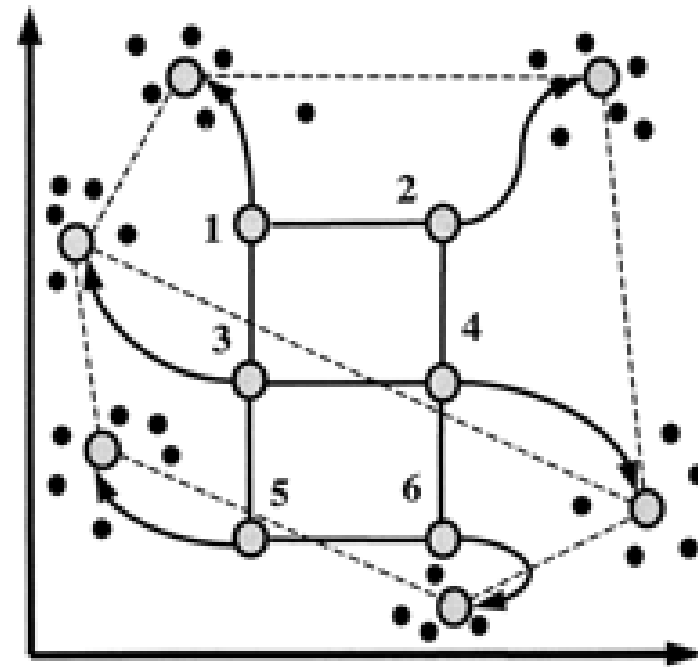


K -means/PAM: How to choose K (the number of clusters)?

- There is no easy answer.
- Many heuristic approaches try to compare the quality of clustering results for different values of K (for an overview see Dudoit/Fridlyand 2002).
- The problem can be better addressed in model-based clustering, where each cluster represents a probability distribution, and a likelihood-based framework can be used.

Self-organizing maps

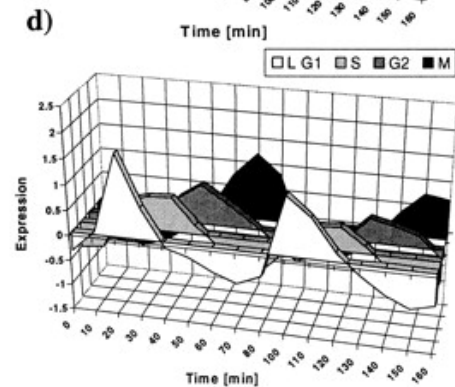
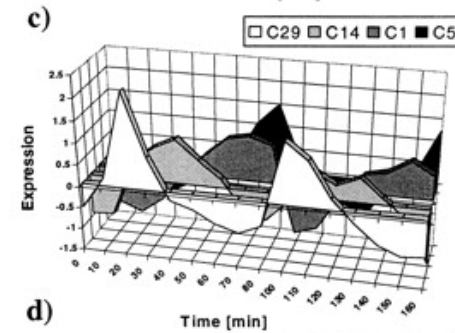
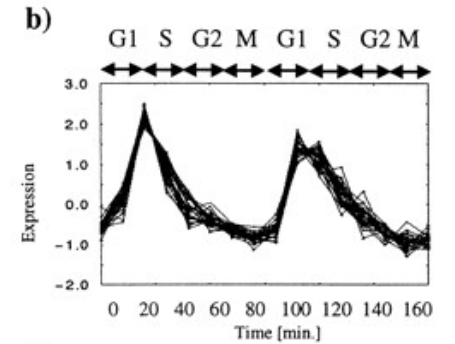
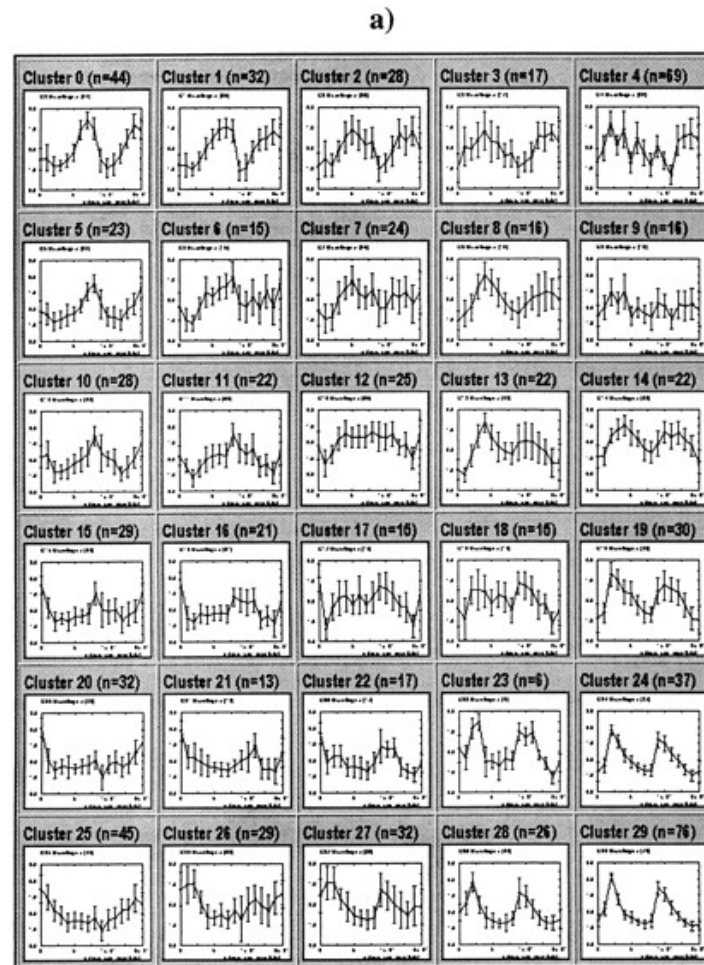
- $K = r*s$ clusters are arranged as nodes of a two-dimensional grid. Nodes represent cluster centers/prototype vectors.
- This allows to represent similarity **between** clusters.
- Algorithm:
Initialize nodes at random positions.
Iterate:
 - Randomly pick one data point (gene) \mathbf{x} .
 - Move nodes towards \mathbf{x} , the closest node most, remote nodes (in terms of the grid) less. Decrease amount of movements with no. of iterations.



from Tamayo et al. 1999

Self-organizing maps


from Tamayo
et al. 1999
(yeast cell
cycle data)





Partitioning or Hierarchical?

- Partitioning:
 - Advantages
 - Optimal for certain criteria.
 - Genes automatically assigned to clusters
 - Disadvantages
 - **Need initial k;**
 - **Often require long computation times.**
 - **All genes are forced into a cluster.**
- Hierarchical
 - Advantages
 - Faster computation.
 - Visual.
 - Disadvantages
 - Unrelated genes are eventually joined
 - Rigid, cannot correct later for erroneous decisions made earlier.
 - Hard to define clusters.



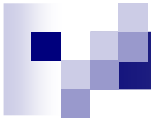
Estimating number of clusters using silhouette

- Define silhouette width of the observation as :
$$S = (b-a)/\max(a,b)$$
- Where a is the average dissimilarity to all the points in the cluster and b is the minimum distance to any of the objects in the other clusters.
- Intuitively, objects with large S are well-clustered while the ones with small S tend to lie between clusters.
- **How many clusters:** Perform clustering for a sequence of the number of clusters k and choose the number of components corresponding to the largest average silhouette.
- *Issue of the number of clusters in the data is most relevant for novel class discovery, i.e. for clustering samples*



Estimating number of clusters using the bootstrap

- There are other resampling (e.g. Dudoit and Fridlyand, 2002) and non-resampling based rules for estimating the number of clusters (for review see Milligan and Cooper (1978) and Dudoit and Fridlyand (2002)).
- The bottom line is that none work very well in complicated situation and, to a large extent, clustering lies outside a usual statistical framework.
- It is always reassuring when you are able to characterize a newly discovered clusters using information that was not used for clustering.



Some examples using R



The data

We will use the dataset presented in van't Veer *et al.* (2002) which is available at <http://www.rii.com/publications/2002/vantveer.htm>.

These data come from a study of gene expression in 91 breast cancer node-negative tumors.

The training samples consisted of 78 tumors, 44 of which did not recur within 5 years of diagnosis and 34 did.

Among the test samples, 7 have not recurred within 5 years and 12 did.

The data were collected on two color Agilent oligo arrays containing about 24K genes.

.



Preprocessing

The data has been filtered using procedures described in the original manuscript.

Only genes showing 2-fold differential expression and p-value for a gene being expressed < 0.01 in more than 5 samples are retained.

There are 4,348 such genes.

Missing values were imputed using k-nearest neighbors imputation procedure (Troyanskaya, et al, 2001).

There, for each gene containing at least one missing value we find 5 genes most highly correlated with it and take average of their value for the sample in which a value for a given gene is missing.

The missing value is replaced with the average.



R data

The filtered gene expression levels are stored in a 4348×97 matrix named **bcdata** with rows corresponding to genes and columns to mRNA samples.

Additionally, the labels are contained in the 97-element vector **surv.resp** with 0 indicating good outcome (no recurrence within 5 years after diagnosis) and 1 indicating bad outcome (recurrence within 5 years after diagnosis).



Hierarchical clustering (1)

- Start performing a hierarchical clustering on the mRNA samples
- using correlation as similarity function and
- complete linkage agglomeration

```
library(stats)
```

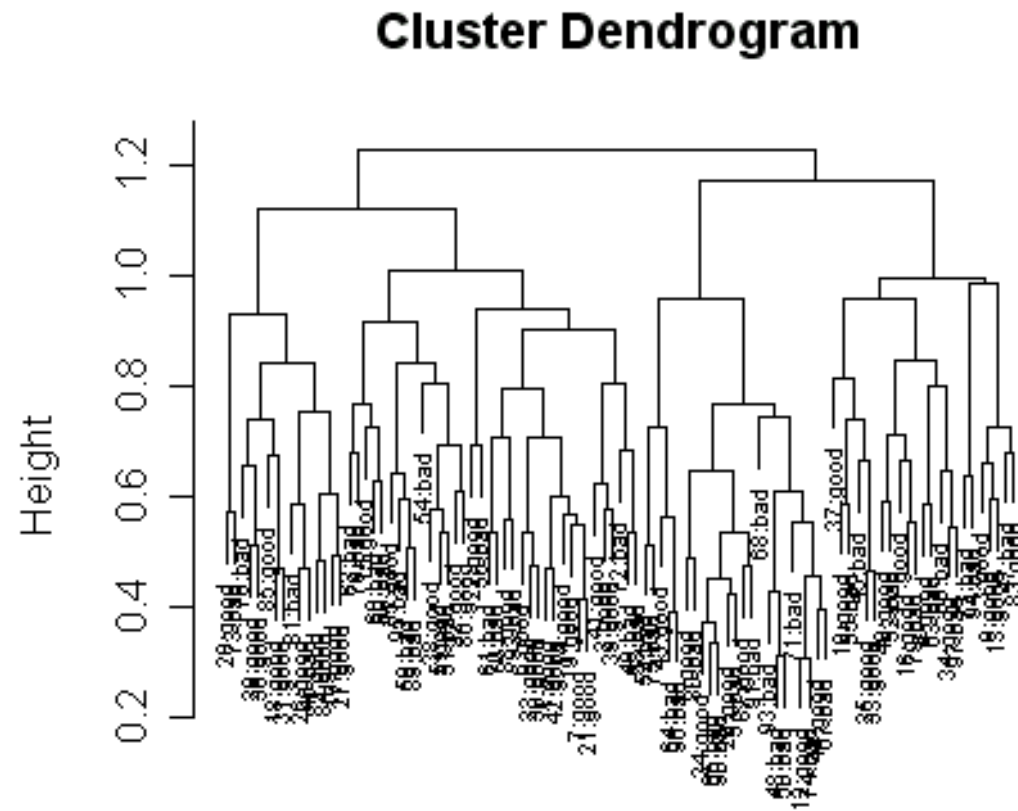
```
x1<-as.dist(1 - cor(bcdata)
```

```
clust.cor <- hclust(x1),  
method="complete")
```

```
plot(clust.cor, cex = 0.6)
```



Hierarchical clustering (1)



```
as.dist(1 - cor(bcddata))
hclust(*, "complete")
```



Hierarchical clustering (2)

- Perform a hierarchical clustering on the mRNA samples using Euclidean distance and
- average linkage agglomeration.
- Results can differ significantly.

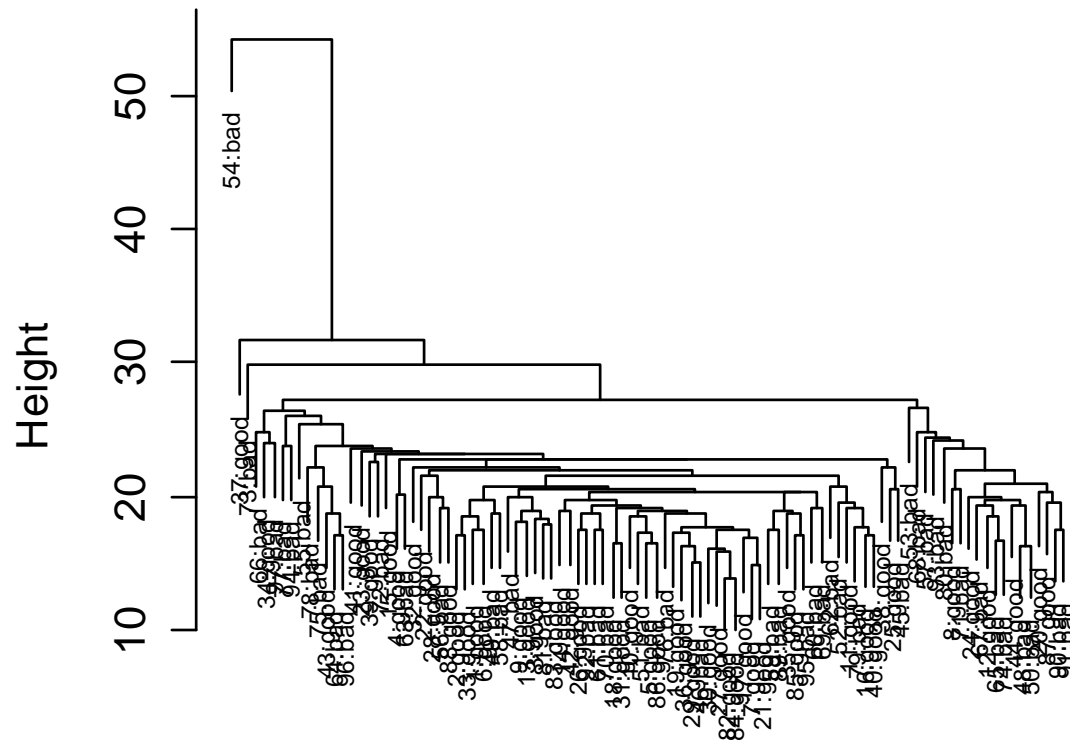
```
x2<-dist(t(bcdata))
```

```
clust.euclid <- hclust(x2),  
  method = "average")
```

```
plot(clust.euclid, cex = 0.6)
```


Hierarchical clustering (2)

Cluster Dendrogram



dist(t(bcdata))
hclust (*, "average")



Comparison between orderings

Sample	CORR.GROUP	EUC.GROUP
1:good	1	1
2:good	1	1
6:good	1	1
7:good	1	1
14:good	1	1
19:good	1	1
21:good	1	1
23:good	1	1
30:good	1	1
39:good	1	1
41:good	1	1
42:good	1	1
49:bad	1	1
58:bad	1	1
61:bad	1	1
70:bad	1	1
72:bad	1	1
88:good	1	1
92:bad	1	1
3:good	2	1
9:good	2	1
13:good	2	1
47:bad	2	1
83:good	2	1
4:good	3	1
11:good	3	1
16:good	3	1
40:good	3	1
52:bad	3	1
63:bad	3	1

- IN THIS CASE WE OBSERVE THAT:
 - Clustering based on correlation and complete linkage distributes samples more uniformly between groups
 - Euclidean-average linkage combination yields one huge group and many small one



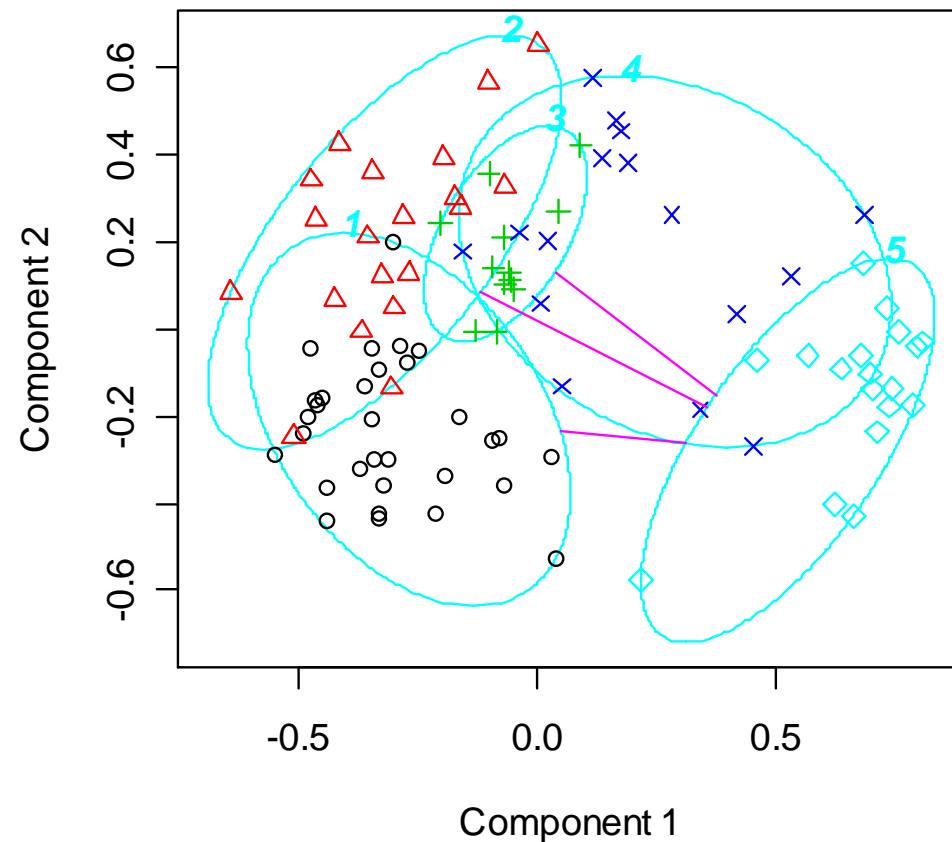
Partitioning around medoids

- If we assume a fixed number of clusters we can use a partitioning method such as PAM
- It is a robust version of k-means which clusters data around the medoids

```
library(cluster)
x3<-as.dist(1-cor(bcddata))
clust.pam <- pam(x3, 5, diss =TRUE)
clusplot(clust.pam, labels = 5,
         col.p = clust.pam$clustering)
```

PAM clustering (2)

`clusplot(pam(x = x3, k = 5, diss = TRUE))`



These two components explain 24.97 % of the point variance



More (many more) examples

http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual#R_clustering