Introduction to clustering methods for gene expression data analysis

Giorgio Valentini

e-mail: valentini@dsi.unimi.it

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

Outline

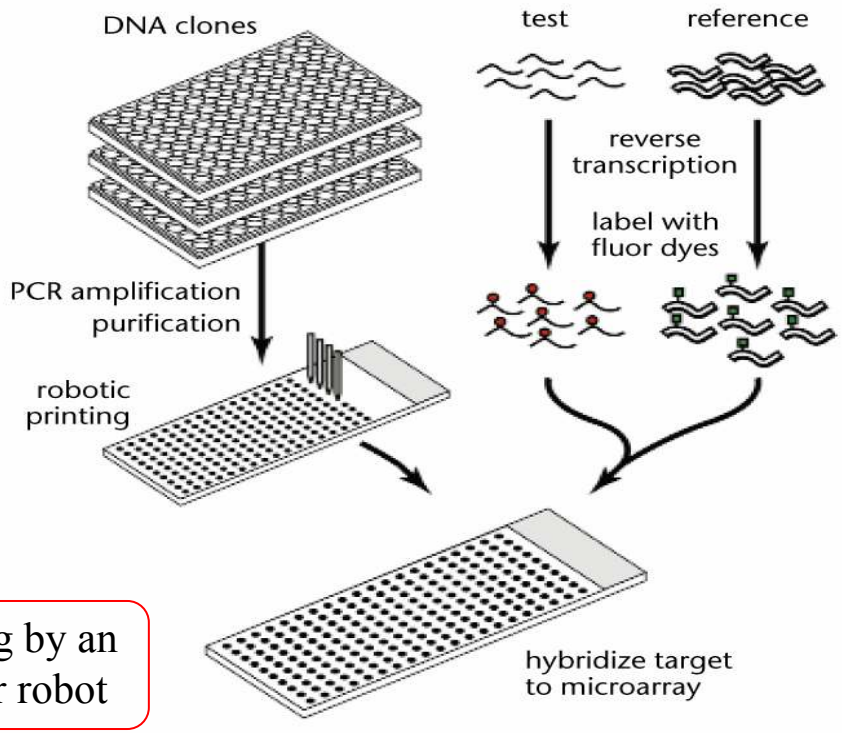
- Levels of analysis of DNA microarray data
- Clustering methods for functional class discovery: discovering gene expression signatures

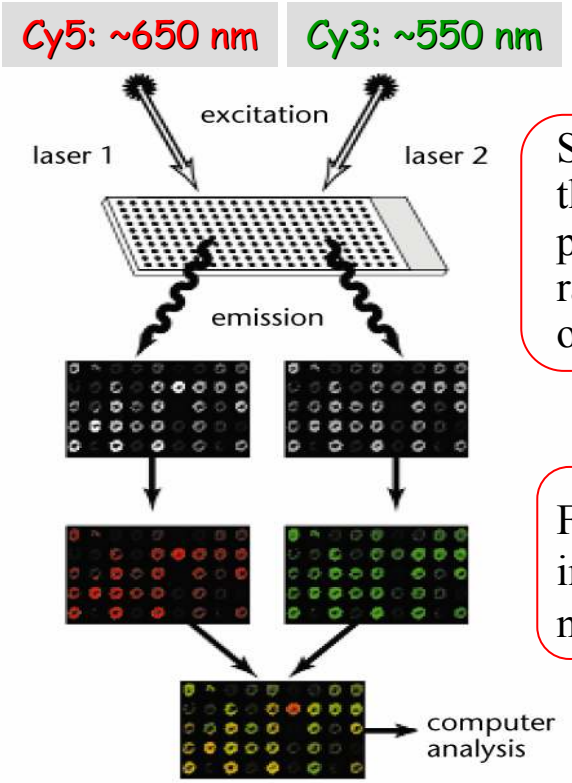
- Hierarchical clustering
- K-means family of clustering methods
- A fuzzy c-mean application to a toy problem

DNA microarray hybridization experiments

Selection of DNA probes (cDNA clones from cDNA libraries)

Preparation of mRNA and cDNA synthesis by reverse transcription





Scanning the slide to produce a raster image of the array

Fluorescent intensities ➔ mRNA levels

Printing by an arrayer robot

Hybridization of the cDNA sequences with the DNA samples of the microarray

Data preprocessing and data analysis

Measuring gene expression through DNA microarray experiments

Large scale gene expression profiling:

- Different tissues
- Different conditions
- Different developmental stage
- Time series experiments
- Responses to external stimuli (drugs, environments, hormones)

To discover:

- Gene function
- Gene regulation
- Metabolic, gene and signaling networks
- Genetic mechanisms of diseases
- ...

Very large amounts of data

Levels of analysis of DNA microarray data

A. - Image analysis
   - Preprocessing and normalization

B. Single gene analysis: Detecting differential expression of single genes

C. Multiple genes pattern discovery: analysis of interactions, common functionalities, co-regulations:
   - Analysis of expression signatures
   - Discovering groups of genes correlated with functional status of a cell
   - Discovering (sub)classes of cell/tissues on functional basis

D. Classification and prediction of functional classes of genes or tissues
   - Diagnosis of polygenic diseases based on gene expression data
   - Predicting subsets of genes related to a particular disease.

E. Pathway analysis: analysis of the relationships between networks of interacting molecules

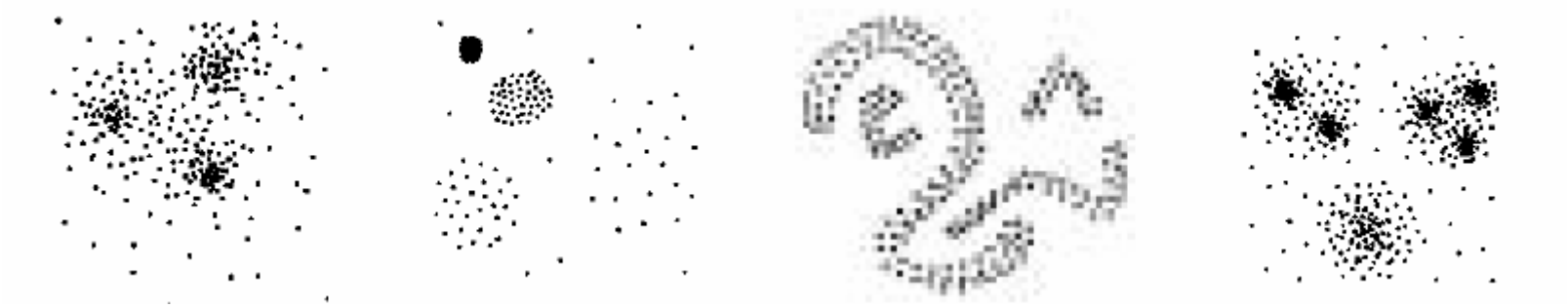
Clustering

- High level overview of the data
- First analytical step in study that involves other anaytical methods
- Unsupervised methods

Goals:

- Discovering the underlying structure of the data
- Discovering groups of co-expressed genes/tissues

Clustering - 1



- Grouping a set of data objects into clusters
- Cluster: a collection of data objects:
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering is an ***unsupervised method*** (no labeled examples)

Typical usage:

As a *stand-alone tool* to get insight into data distribution

As a *preprocessing step* for other algorithms

Clustering - 2

Goals:

- Inferring unknown gene functions from clusters
- Discovering functionally related sets of genes
- Discovering new subclasses of diseases
- Discovering regulatory networks

For instance, clustering permits us to

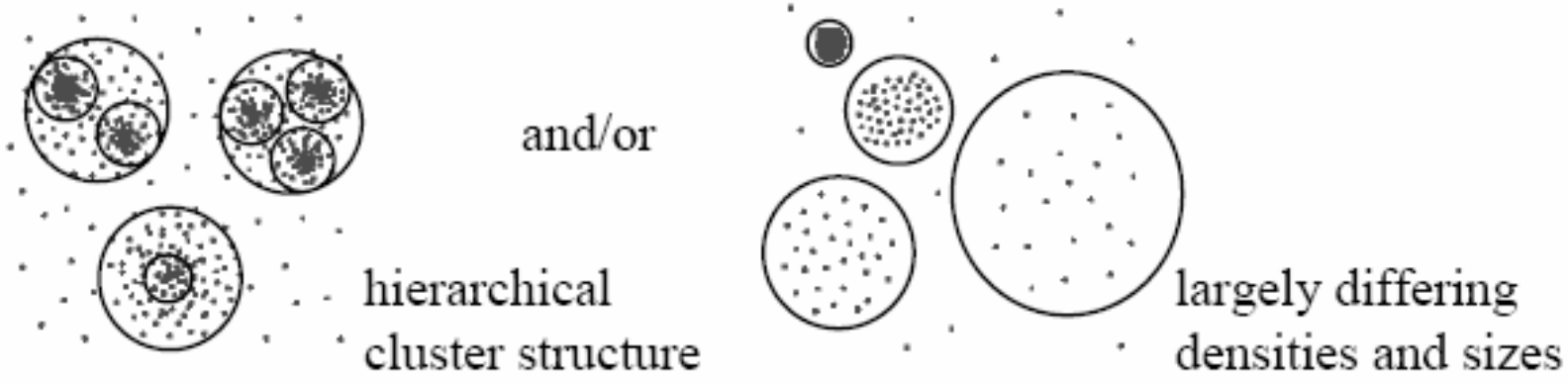
– group together genes that respond similarly across several experimental conditions

– group together experimental conditions tissue types giving similar expression patterns across the whole genome

Clustering methods applied to gene expression data analysis

- Hierarchical clustering (Eisen et al., 1998)
- K-mean family clustering (Tavazoie et al., 1999; Gasch and Eisen, 2002)
- Self-organizing maps (SOM) (Tamayo et al., 1999)
- Methods based on graph theory (Sharan and Shamir, 2000)
- Methods based on within cluster maximization and between cluster similarity minimization (De Veet et al., 2002)
- Biclustering methods (Tanay et al., 2002)
- Ensemble methods (Dudoit et al. 2003)

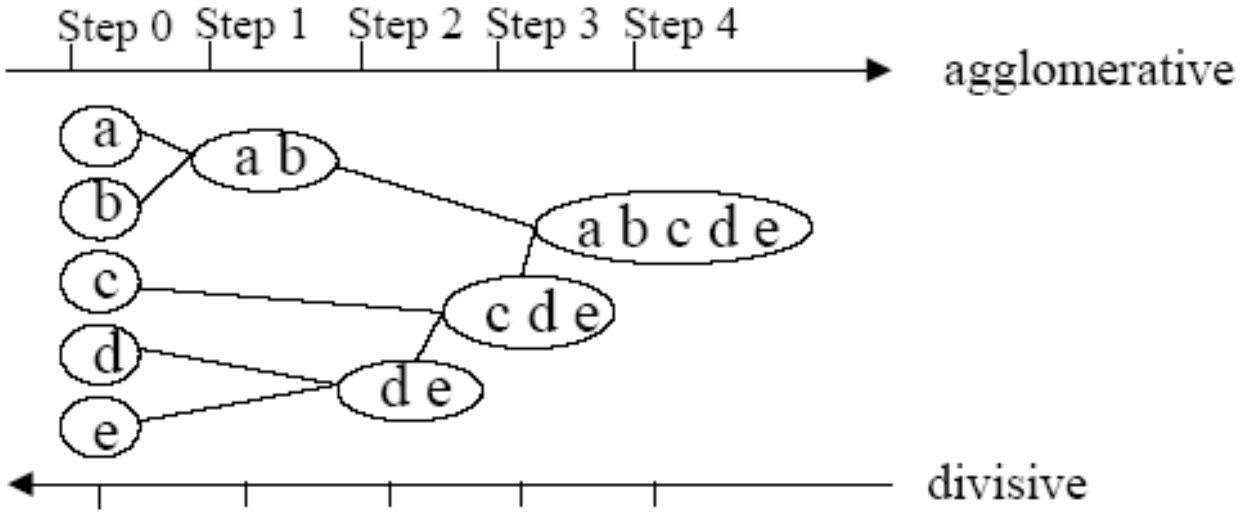
Data may have a hierarchical structure



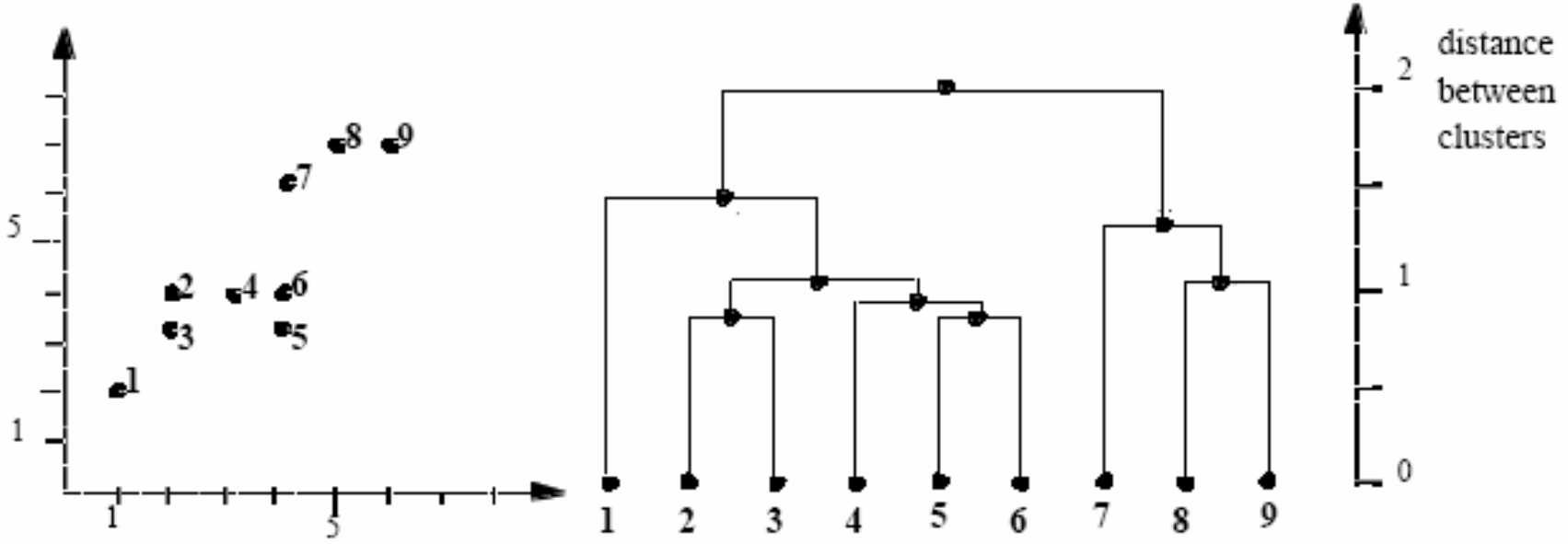
It could be useful a hierarchical clustering algorithm in these situations

Hierarchical clustering

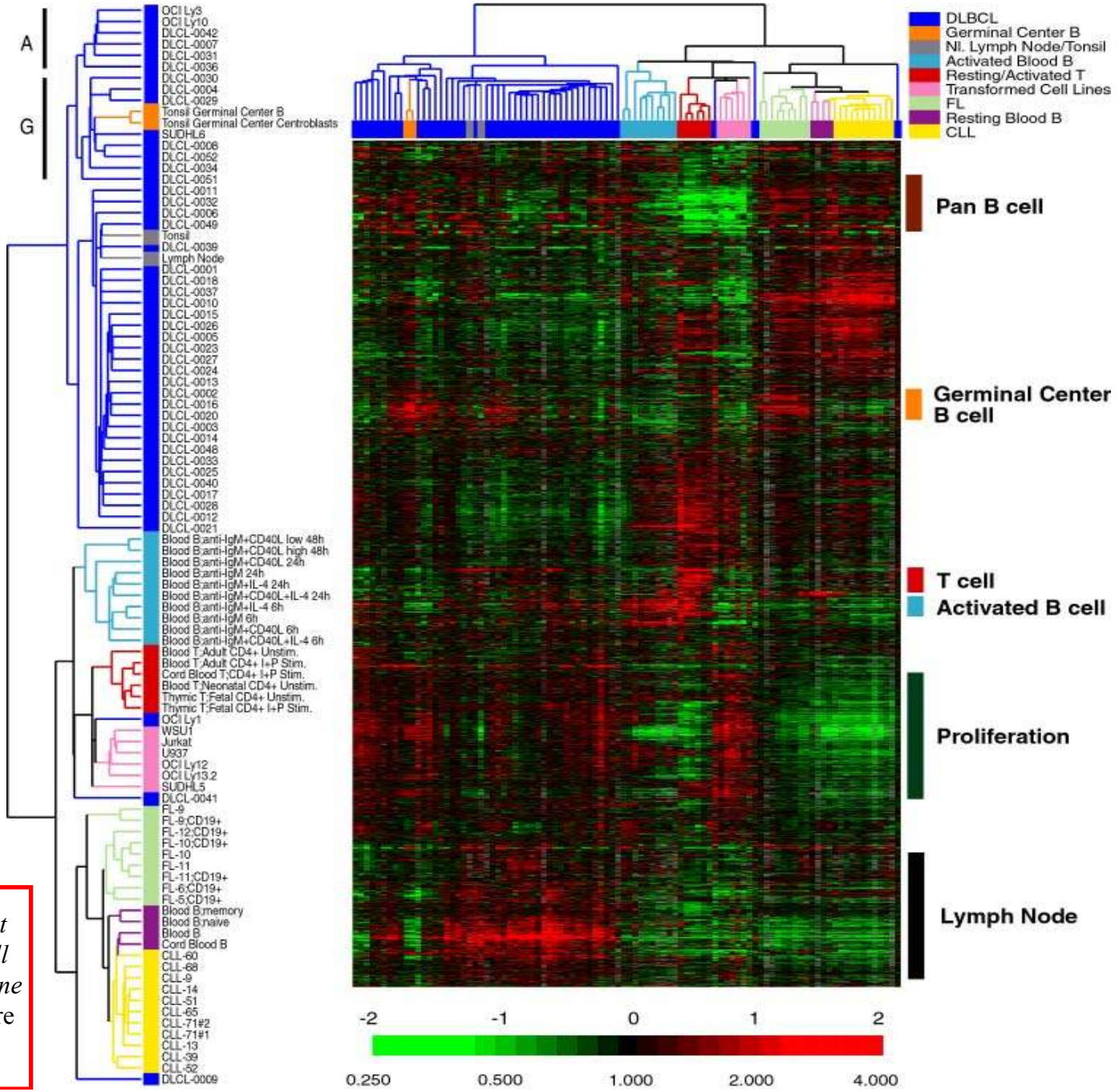
- Hierarchical decomposition of the data set (with respect to a given similarity measure) into a set of nested clusters
- Result represented by a *dendrogram*
- Nodes in the dendrogram represent possible clusters
- They can be constructed bottom-up (agglomerative approach) or top down (divisive approach)



Dendrograms



- The *root* represents the whole data set
- A *leaf* represents a single object in the data set
- An *internal node* represent the union of all objects in its sub-tree
- The *height* of an internal node represents the distance between its two child nodes



Alizadeh, A. et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature 403, p. 503-511, 2000

Advantages and disadvantages of Hierarchical clustering

Advantages

- Does not require the number of clusters to be known in advance
- No input parameters (besides the choice of the (dis)similarity)
- Computes a complete hierarchy of clusters
- Good result visualizations integrated into the methods

Disadvantages

- May not scale well: runtime for the standard methods: $O(n^2 \log n^2)$
- No explicit clusters: a "flat" partition can be derived afterwards (e.g. via a cut through the dendrogram or termination condition in the construction)
- No automatic discovering of "optimal clusters"

The K-means algorithm

Input : A set S of examples (vectors of gene expression levels), a number K of clusters

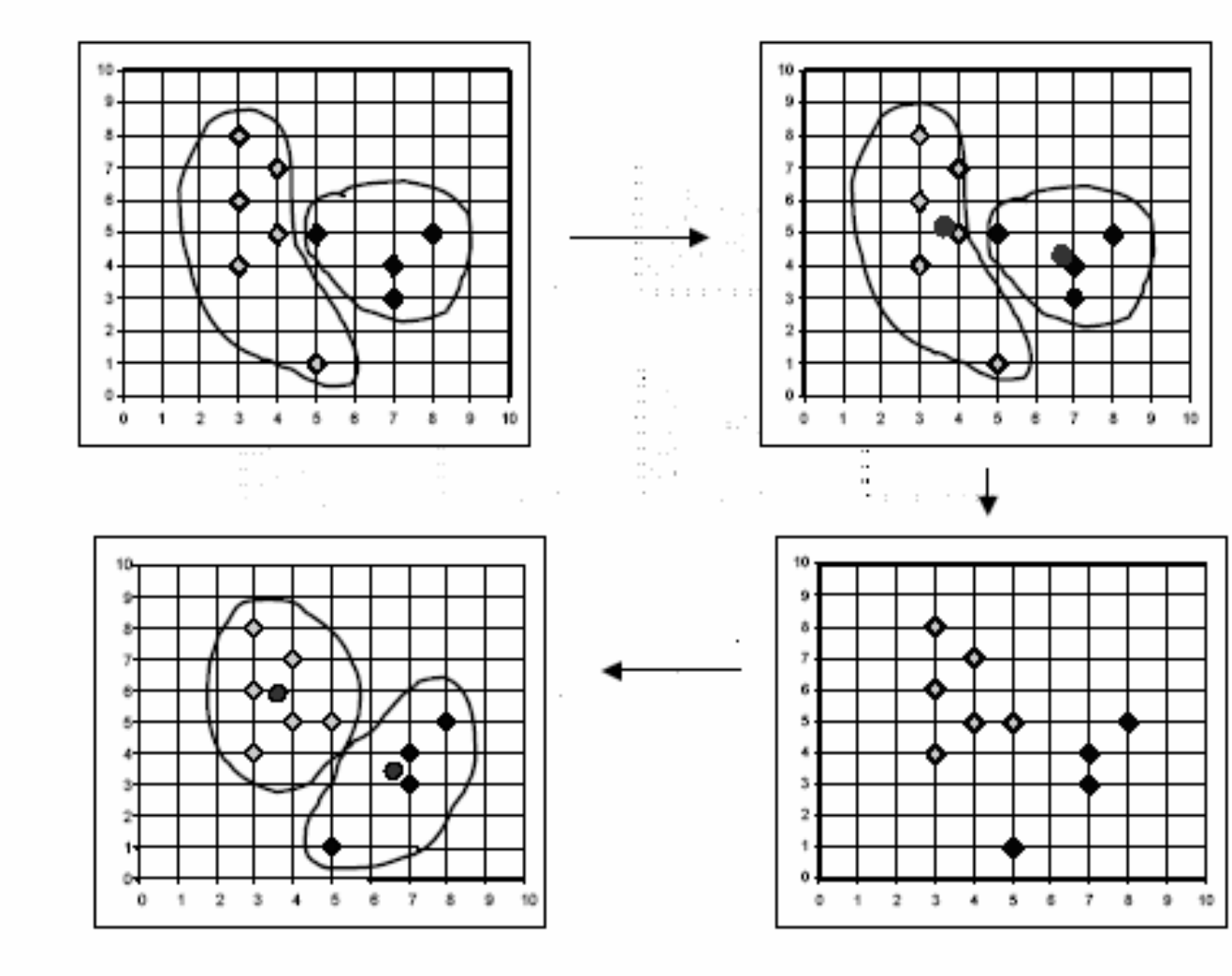
1. Initialization: assign the examples randomly to the K clusters

loop:

2. Compute the mean fo each cluster
3. Assign each example to the “nearest cluster”
4. If stop condition reached then exit loop, else repeat loop

Output: A set of K clusters

K-means algorithm



K-means for gene expression

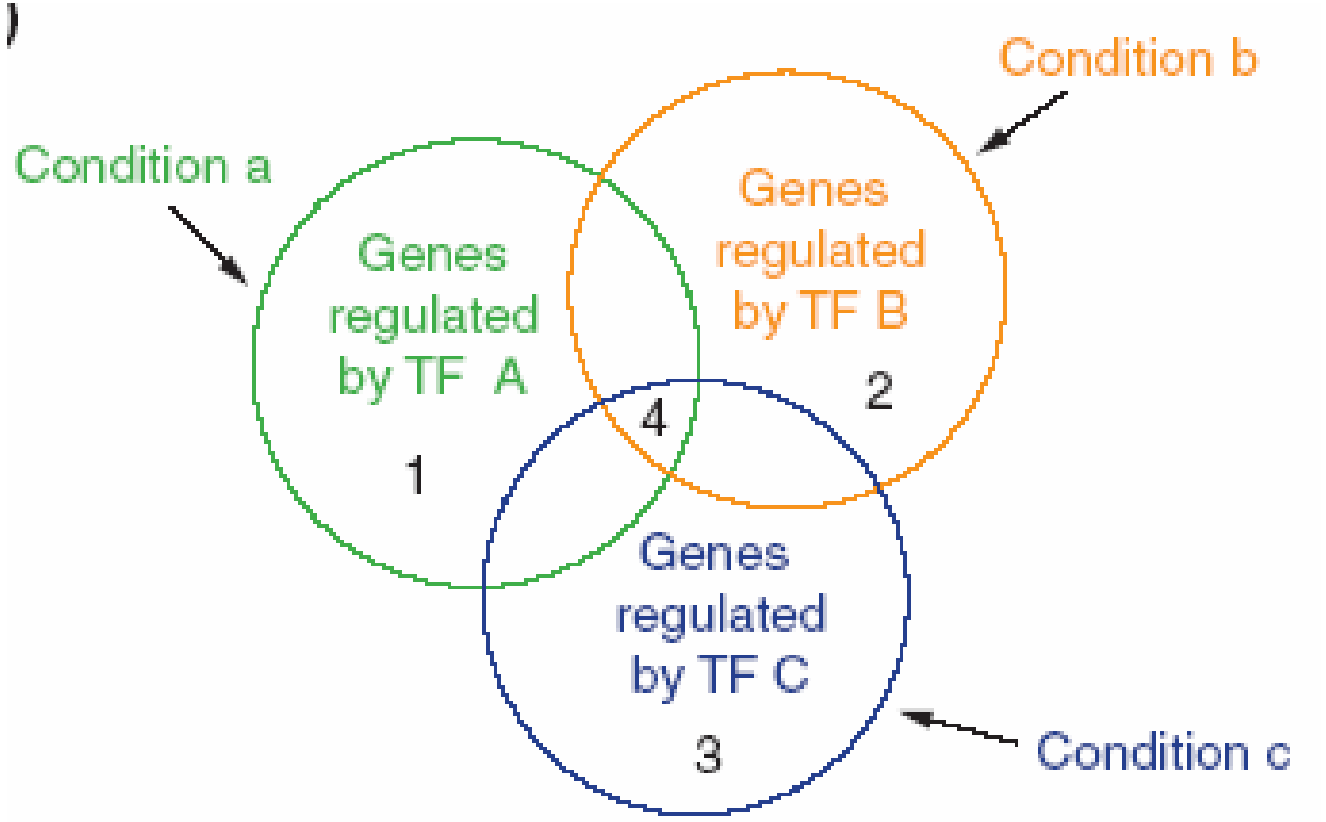
Strength:

- Relatively efficient: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t << n$.
- Easy implementation.

Weakness:

- Need to specify k, the number of clusters, in advance.
- Sensitive to noisy data and outliers
- Clusters are forced to have convex shapes
- Result and runtime are dependent on the initial partition;
- often terminates at a *local optimum*
- Genes are forced to belong only to 1 cluster

Genes can be regulated by multiple transcription factors



Fuzzy C-mean

- Each gene can belong to multiple clusters: its *membership* can vary from 0 to 1.
- As in the “hard” version, it is an *iterative algorithm*, but the means are weighted through the memberhips, and the memberships are updated at each iteration

•The *output* is a fuzzy partition of the data, where each gene/sample can belong to more than 1 cluster

•We can obtain also *different partitions* with genes belonging or not to a cluster according to a selected membership cut-off

Using the fuzzy c-mean algorithm

- Main parameters to be tuned:
  - The number of clusters
  - The index of fuzziness
- Initialization of the algorithm
- The reliability of the clusters
- Selecting a membership level to assign a gene to a specific cluster.

A MATLAB implemetation of the fuzzy-c-mean algorithm

function [A, Y, H, cycles, HBindex] = fuzzy_c_mean (X, c, f, errstop, lim)
Esegue l' algoritmo di fuzzy c means secondo Bezdek
Input:
X : matrice dei dati di input.
Formato: [n,m]=size(X) con n numero dei pattern ed m dimensione dei pattern.
c : numero delle classi (def. c = 2)
f : indice di fuzziness dell' algoritmo, 1<f<inf (def. f = 2)
errstop: condizione di arresto (def. = 0.001)
lim : limite valore di membership per assegnazione di un pattern ad un cluster
Output:
A : Matrice delle membership function dei cluster.
Formato: [c,n]=size(A), c = numero dei cluster, n = numero dei pattern di input
Y : Matrice dei centri dei cluster
Formato: [c,m]=size(Y), c = numero dei cluster, m = dimensione dei pattern
H : Vettore di assegnazione dei pattern ai cluster
Formato: n=length(H), n = numero dei pattern di input
cycles : numero dei cicli effettuati dall' algoritmo
HBindex : Xie-Beni index (misura la qualità della partizione)
minore è il valore di HBindex, migliore è la partizione

High-level view of the fuzzy c-mean algorithm

```
% Step1: Select an initial pseudopartition
A = DoInitPseudoPartition(c,n);
while err > errstop
        %Step2: Cluster-centers computations
        Y = ClusterCenters(A, X, f);
        % Save the membership matrix
        oldA = A;
        % Step3: Update the memebership matrix A
        A = UpdateMembership(X, Y, c, f);
        % Step4: Compute the differences (error) between A and its previous values
        err = Error(A, oldA);
        cycles = cycles + 1;
end
H = AssignPatterns(A, lim);
HBindex = ComputeXieBeniIndex(X,Y,A,f);
% Plot the clusters
% drawfuzzycluster(X, A, lim);
```

Limitations of fuzzy C-means

- User defind membership cut-off.Some criteria:
  - Functional relationships of the genes selected
  - Biological coherence of the selected patterns
  - Statistical enrichment of sequences in gene promoters
- No “natural” visualization of the data
- “Outlier” genes forced to belong to some cluster (due to Ruspini constraint)

In general there is not the “best” clustering method

Different techniques highlight different patterns and characteristics of the data.

For example:

Fuzzy clustering: identifying genes pertaining to different regulatory networks

Hierarchical clustering: it offers an intuitive visual clue of the distribution of the data

Biclustering: identifying subsets of genes with similar behavior in subsets of experimental conditions

A final caveat about clustering methods

- We should use clustering methods if we pursue pattern discovery or dimensionality reduction
- *It is not a good idea using them to distinguish between classes of examples (if the class labels are known)*

A short bibliography on clustering methods for gene expression data analysis

Eisen, M.B. et al., *Cluster analysis and display of genome-wide expression patterns*, PNAS (25)95, p. 14863--14868 1998.

Tamayo, P. and others, *Interpreting patterns of gene expression with self-organizing maps*, PNAS 96, p.2907--2912, 1999

Golub, T.R. and others, *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science 286, p. 531--537, 1999.

Alizadeh, A. et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature 403, p. 503-511, 2000

Tanay, A. and Sharan, R. and Shamir, R., *Discovering Statistically Significant Biclusters in Gene Expression Data*, Bioinformatics 18, 2002.

Gasch P. and Eisen, M. *Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering*, Genome Biology 3(11), 2002.

Reviews of clustering methods for gene expression data analysis

Jain A. Murty M. And Flynn P. *Data clustering: a review*, ACM Computing Surveys. (5)31, p. 264-323, 1999.

Ben-Dor, A. and Shamir, R. and Yakhini, Z., *Clustering gene expression patterns*, Journal of Computational Biology, (3)6, p.281--297, 1999

D. Jiang, C. Tang, A. Zhang Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, pp. 1370-1386, 2004.

Madeira SC, Oliveira AL Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics* **1** (1): 24–45. 2004