



Multi-SOM: A novel clustering algorithm for gene expression data analysis

A. Ghouila, H. Jmel, S. BenYahia, D. Malouche and
S. Abdelhak

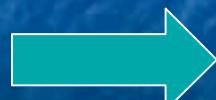
Bioinformatics for Africa, Nairobi
2007

Functional genomics

- Microarray and SAGE data analysis
- Thousands of gene expression levels are studied simultaneously



Need for tools to analyse the expression data of several thousands of genes

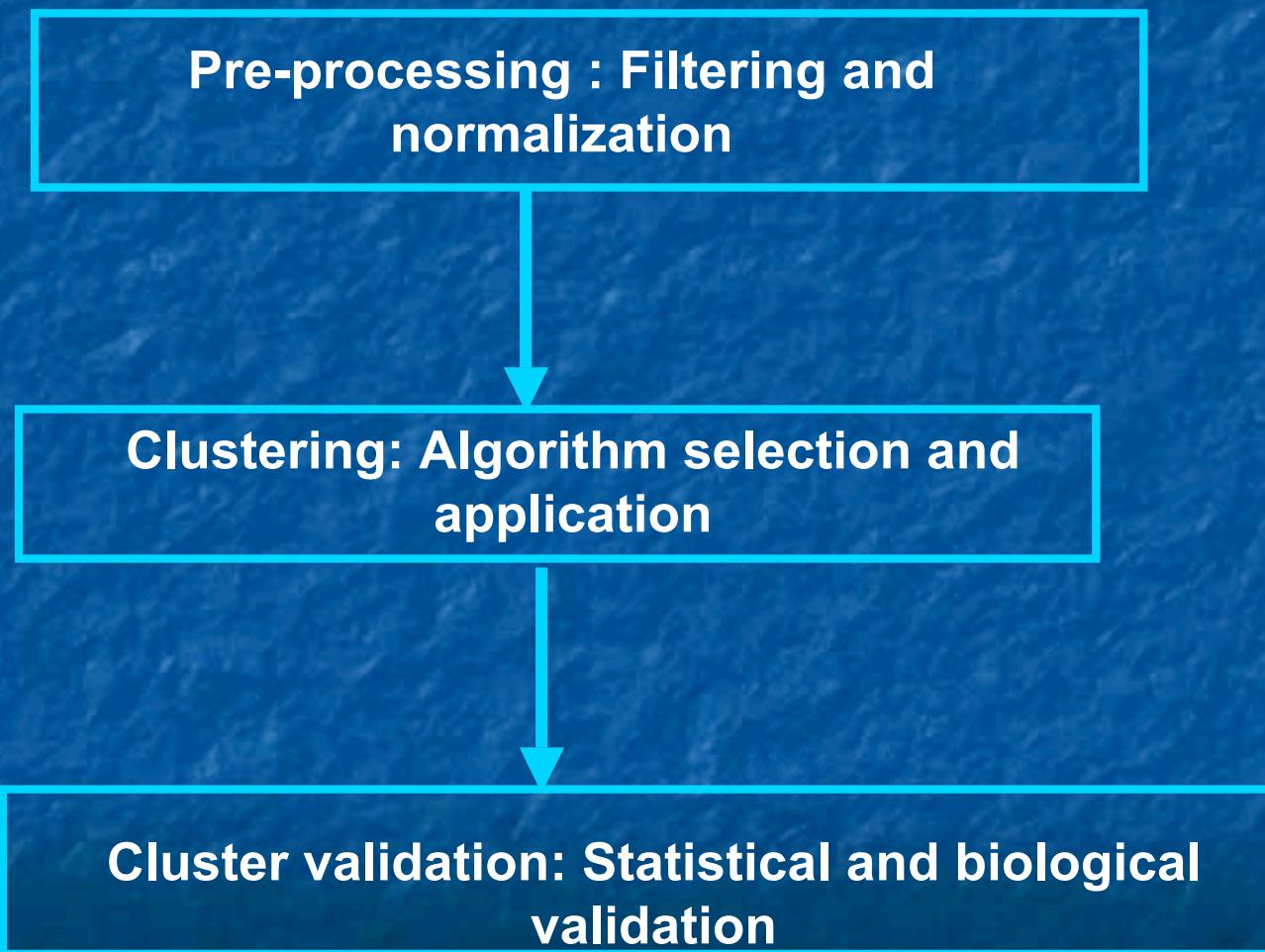


Translate the results into meaningful biological knowledge

Clustering of microarray data

- Organize the genes into meaningful groups exhibiting similar patterns of expression level
- Genes with similar expression profiles:
 - have similar biological function
 - are frequently co-regulated
 - contribute to a common pathway

Microarray data analysis steps



Clustering

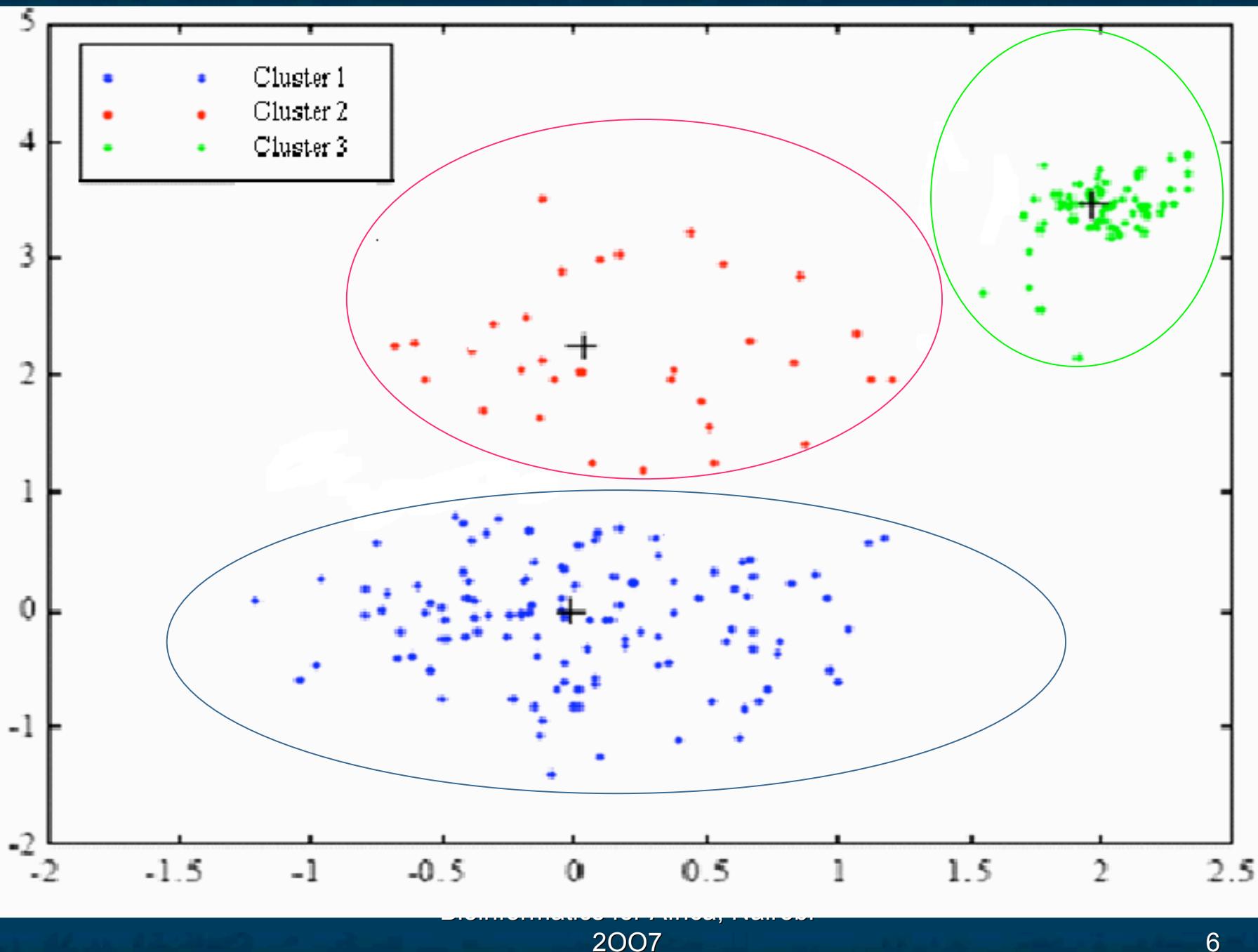
■ Aim



To divide samples into homogeneous groups (clusters) based on their similarities.



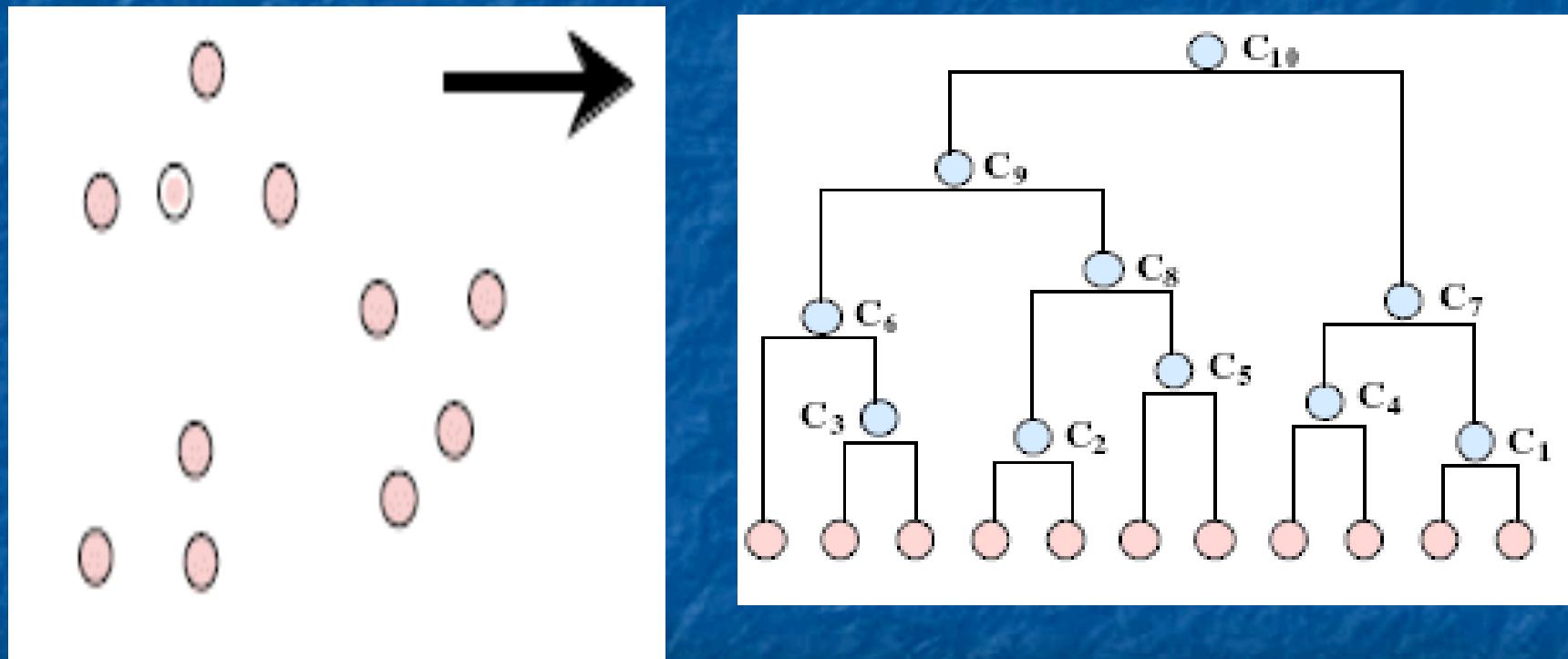
A clustering algorithm must guarantee good separation between clusters as well as intra-cluster homogeneity



Clustering algorithms

- Hierarchical clustering
- K-means clustering
- Self Organising maps

Hierarchical clustering



Hierarchical clustering (2)

- Analysis of expression profile of macrophage infected with different Leishmania species (Chaussabel and al., 2003)
- Analysis of expression profile macrophage exposed to bacterial pathogens (Nau and al., 2002)
 - ↗ Do not require many parameters
 - ↗ Easy to apply

Limits

- ☞ Difficulty to delimitate the cluster boundaries
- ☞ Analysis is based on visual inspection of the hierarchical tree
- ☞ No formal rules

Self Organizing Maps (SOM)

- ➲ One of the most commonly used artificial neural network
- ➲ The reduction of the complexity of the data space
- ➲ Very useful and robust approach to the clustering of large amount of data

Self Organizing Maps (SOM)

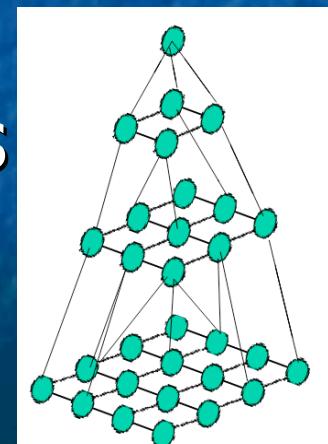
- ☞ Finding clusters from the SOM grid is still a crucial task to tackle
- ☞ Difficulty to decide about the number of the grid units
- ☞ Similar neurons need to be grouped

Proposed solutions

- Start with a large grid to obtain homogenous clusters
- Gradually decrease the number of clusters by grouping similar units
- Introduce statistical indexes to better understand data characteristics

Development of Multi-SOM

- Based on Self Organizing Maps
- Data is first clustered by SOM
- SOM grid is then clustered
- Build an hierarchy of SOM grids
- Each grid aims to group similar units within the previous one
- Integrate the use of validity indices guide the cluster delimitation



Validation on labeled data

- Labeled data sets
 - Iris data set
 - Pima Indians for diabetes
- The application of Multi-SOM:
 - Identification of the correct number of classes
 - A better performance was obtained (Smaller error values)
 - Better classification

Cluster validation

■ Statistical validation :

- Based on the gene expression levels
- Assess the cluster separation
- Assess the distribution of different samples among clusters

Cluster validation (2)

- Biological validation:
 - Visualisation tools to understand gene functions within a cluster
 - Based on the use of Gene Ontology

Multi-SOM testing

1720 transcripts hybridised on Affymetrix HGU133A (22,000 transcripts)

1030 probes representing 978 non redundant over expressed genes

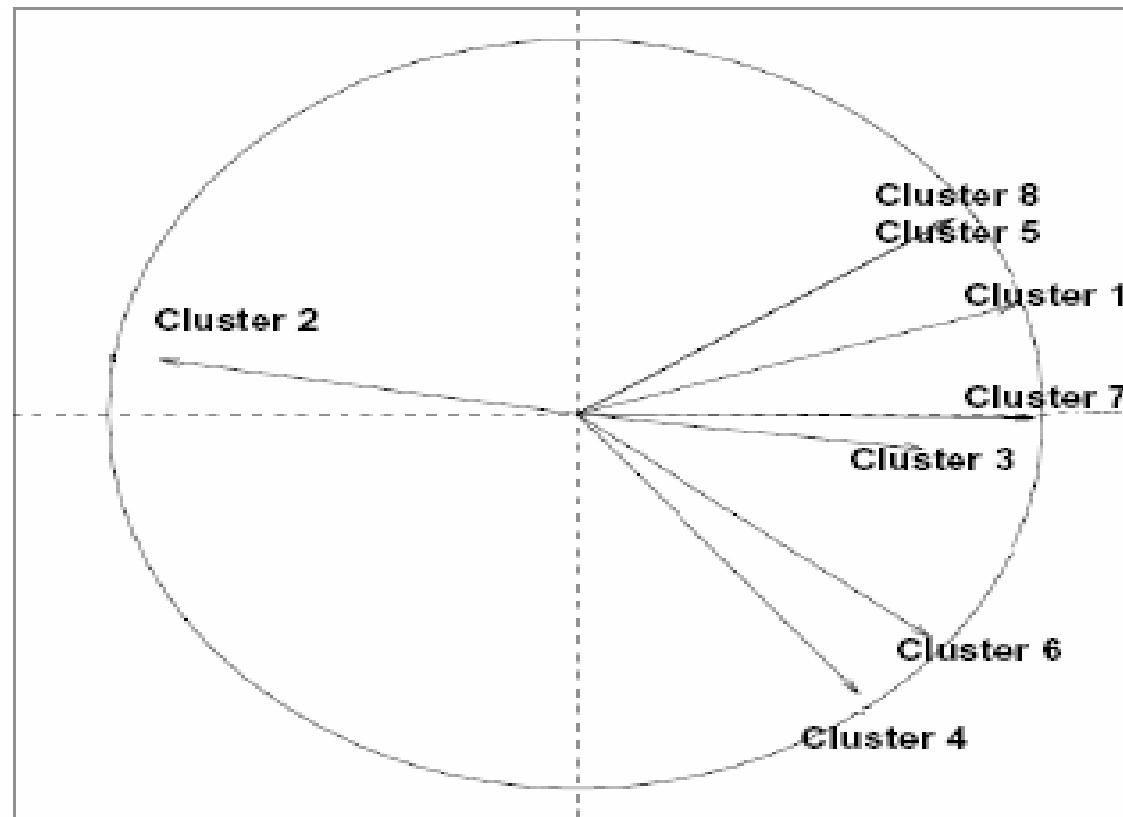
8 Clusters potentially regulated having similar gene expression profiles

Cluster validation

Statistical validation:
PCA

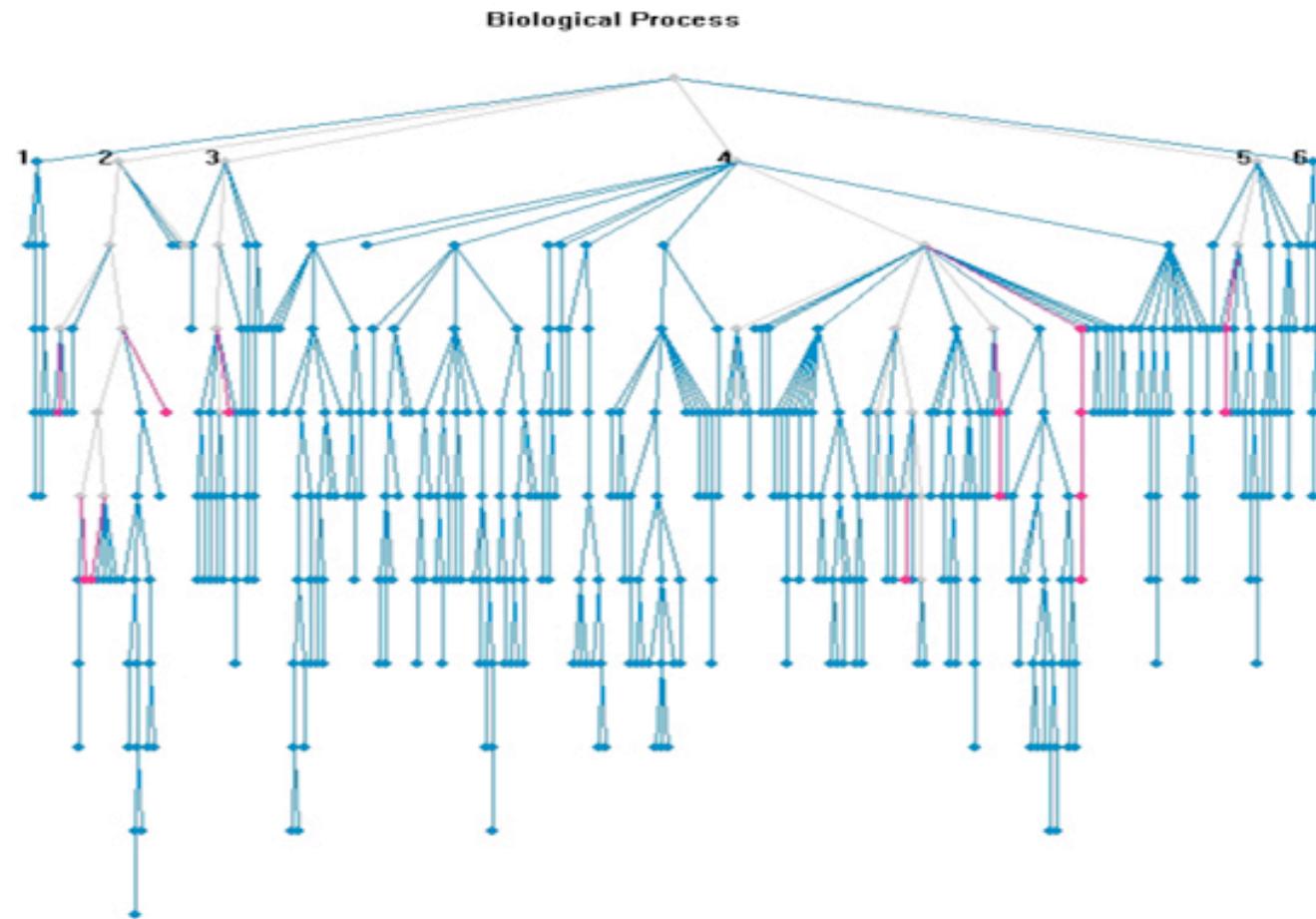
Biological validation:
GO

Principal Component Analysis



Biological Validation

- GOTM
- Onto Express
- GoSurfer
- Fatigo



- **Group 1 only**
- **Group 2 only**
- **both groups**

- 1 **behavior**
- 2 **cellular process**
- 3 **development**
- 4 **physiological process**
- 5 **regulation of biological process**
- 6 **viral life cycle**

Perspectives

- Diseases gene
- Host
- Pathogens
- Influence on gene expression



Acknowledgment

A. Ghouila

S. BenYahia

D. Malouche

H. Jmel

S. Abdelhak

Computer science

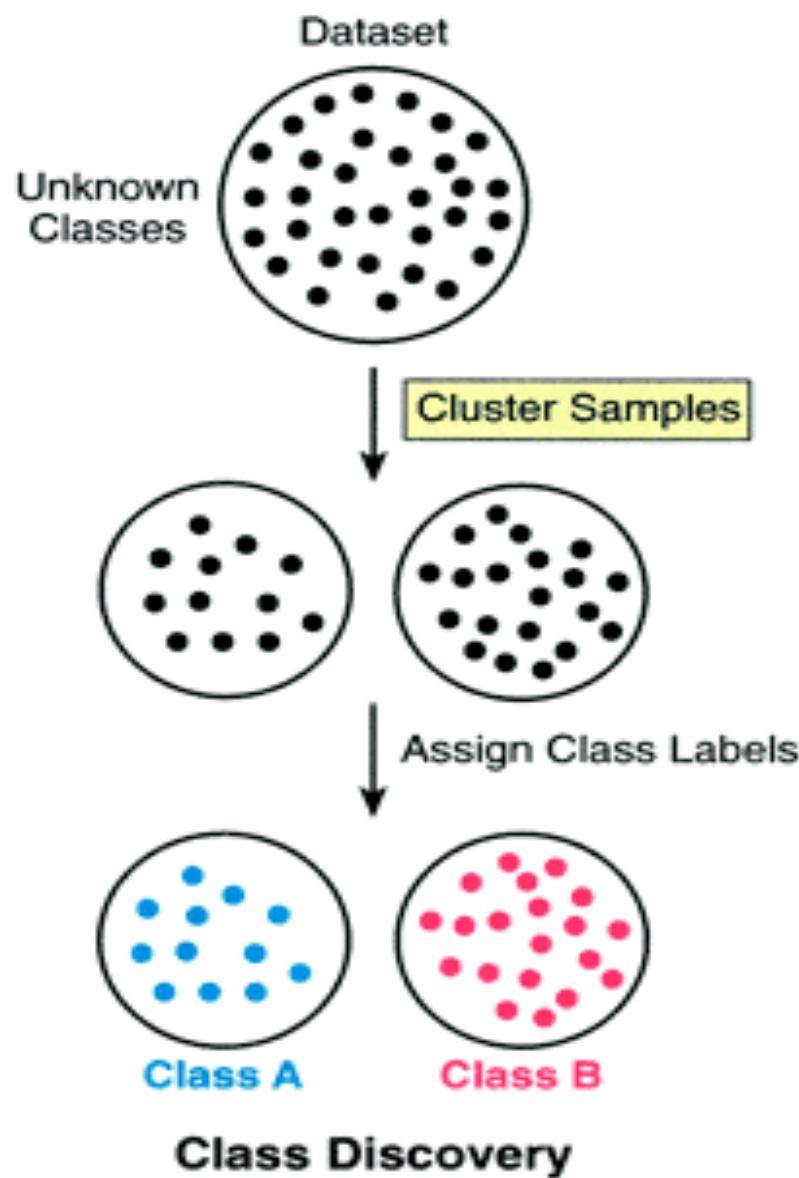
Statistics

Biology

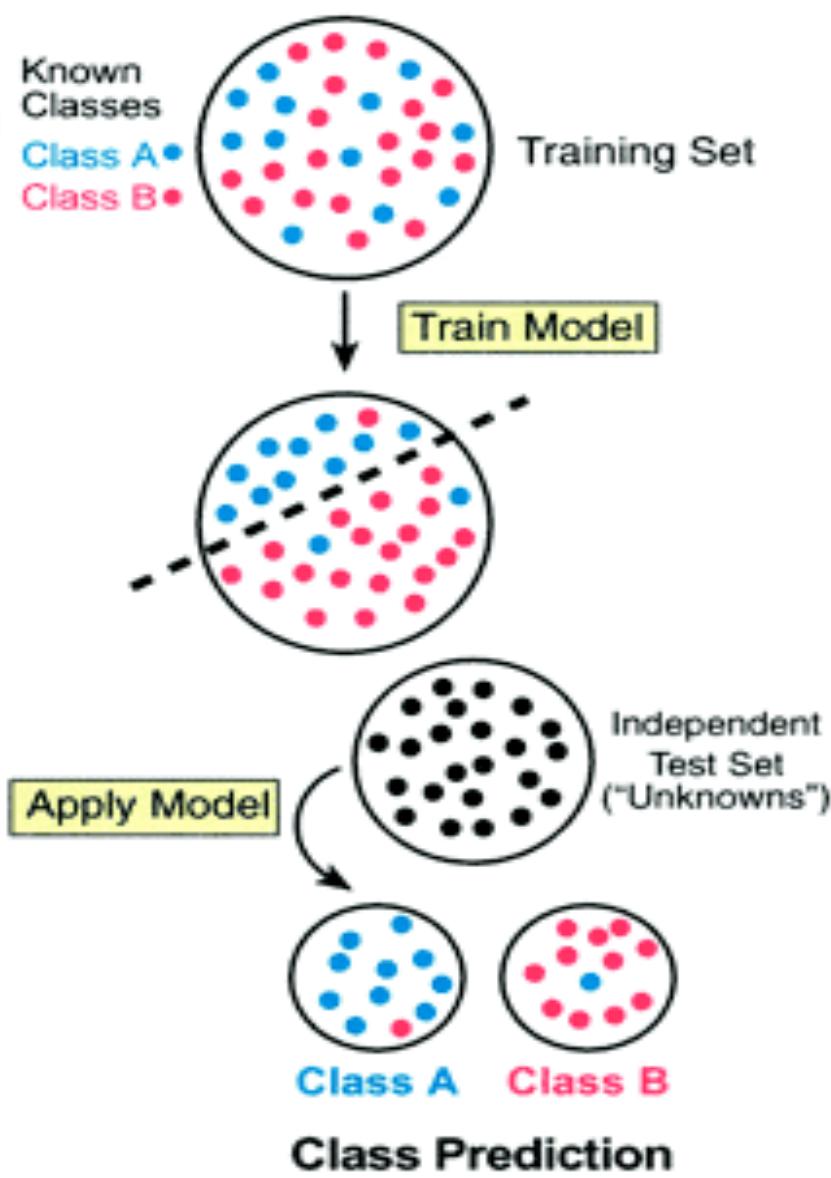
Thank you!!



UNSUPERVISED LEARNING



SUPERVISED LEARNING



Microarray gene expression data

Data preprocessing

Characterisation of gene expression levels of clusters

- Filtering and Normalization
- Clustering (MultiSOM)

8 Clusters potentially regulated having similar gene expression profiles

Cluster characterization

Statistical validation

1030 probes representing 978 non redundant genes over expressed

Cluster validation

Biological validation

K-means clustering

- ↗ A partitionning algorithm
- ↗ Very Simple
- ↖ Requires the cluster number to be initially fixed
- ↖ Depends heavily on the initialization step

Statistical validation

- Carried out using Principal Component Analysis (PCA)
- A first PCA Showed a good separation between clusters
- A second one showed a good separation between genes over-expressed in CSS, NHS
- Keratinocytes and Fibroblasts were merged in the same cluster

Microarray data analysis steps

