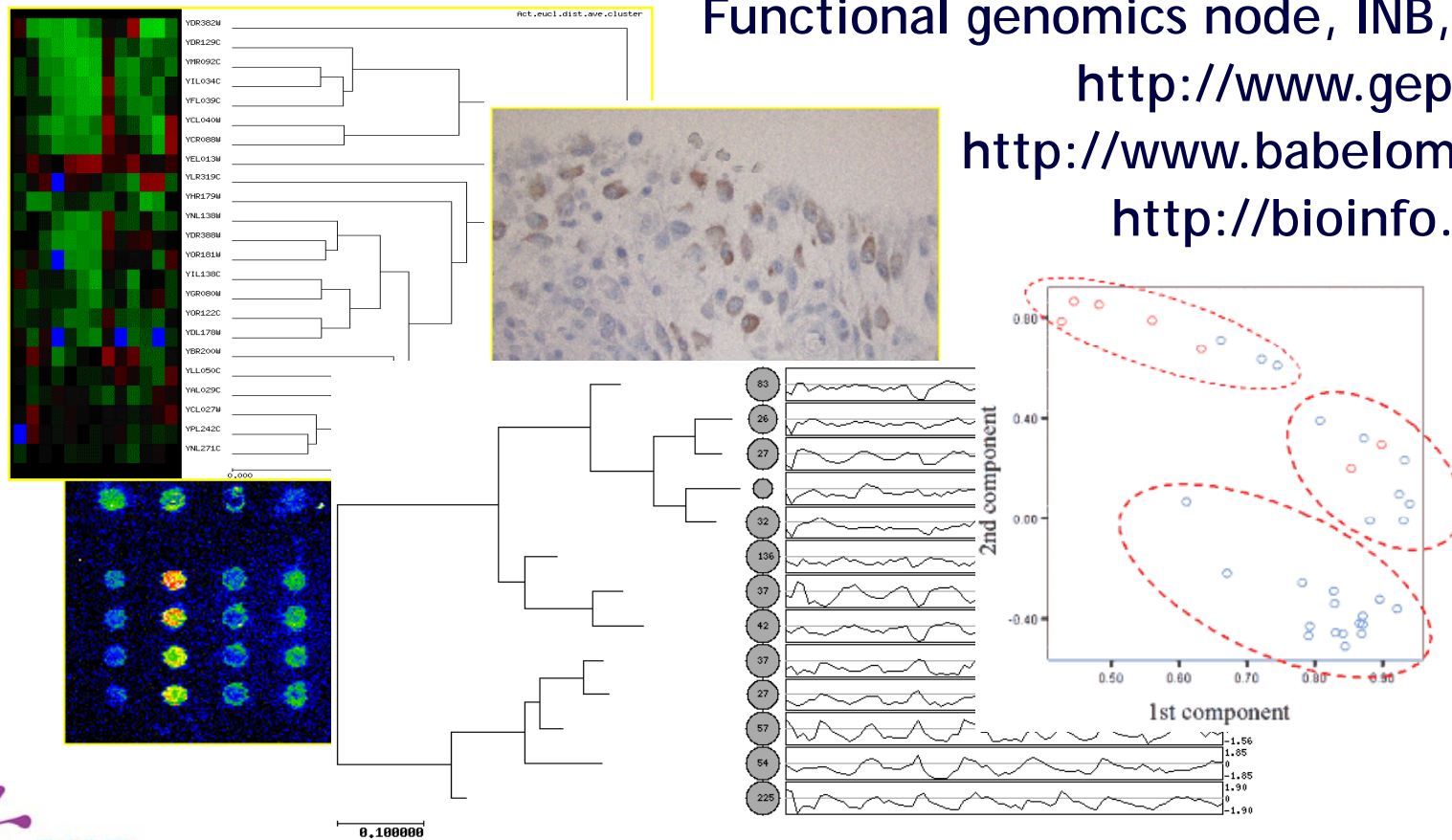


Clustering of DNA microarray data

Department of Bioinformatics and Genomics,
Centro de Investigación Príncipe Felipe, and
Functional genomics node, INB, Spain.

<http://www.gepas.org>
<http://www.babelomics.org>
<http://bioinfo.cipf.es>



INSTITUTO NACIONAL
DE BIOINFORMÁTICA



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Studies must be hypothesis driven.

Can we find groups of experiments with similar gene expression profiles?

What is our aim? Class discovery?
sample classification? gene selection?

...

Unsupervised
Supervised

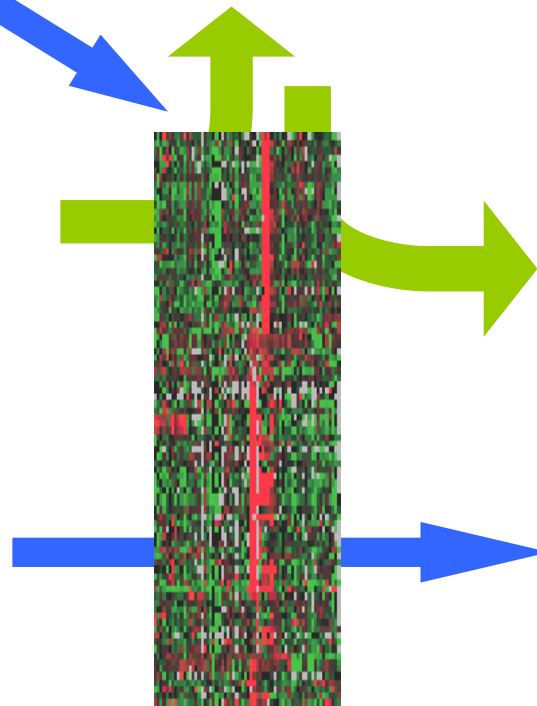
Different classes...

Molecular classification of samples

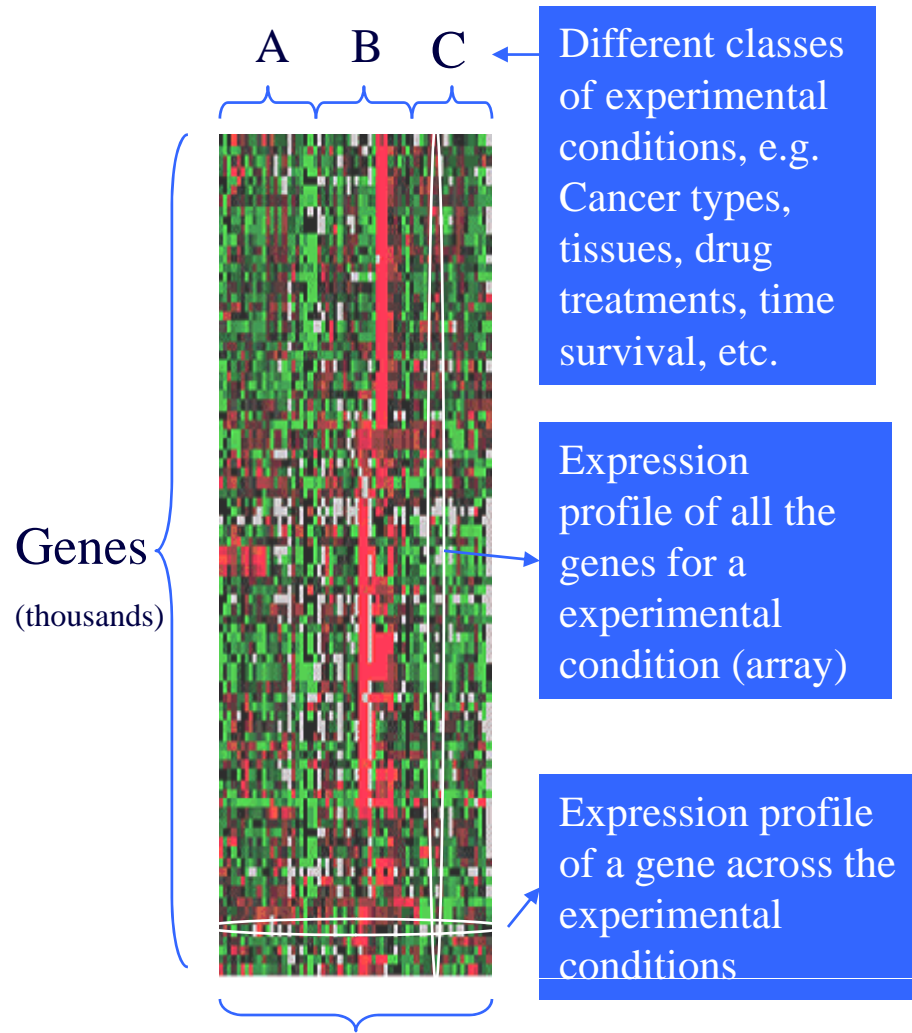
What genes are responsible for?

Co-expressing genes...

What do they have in common?



The data



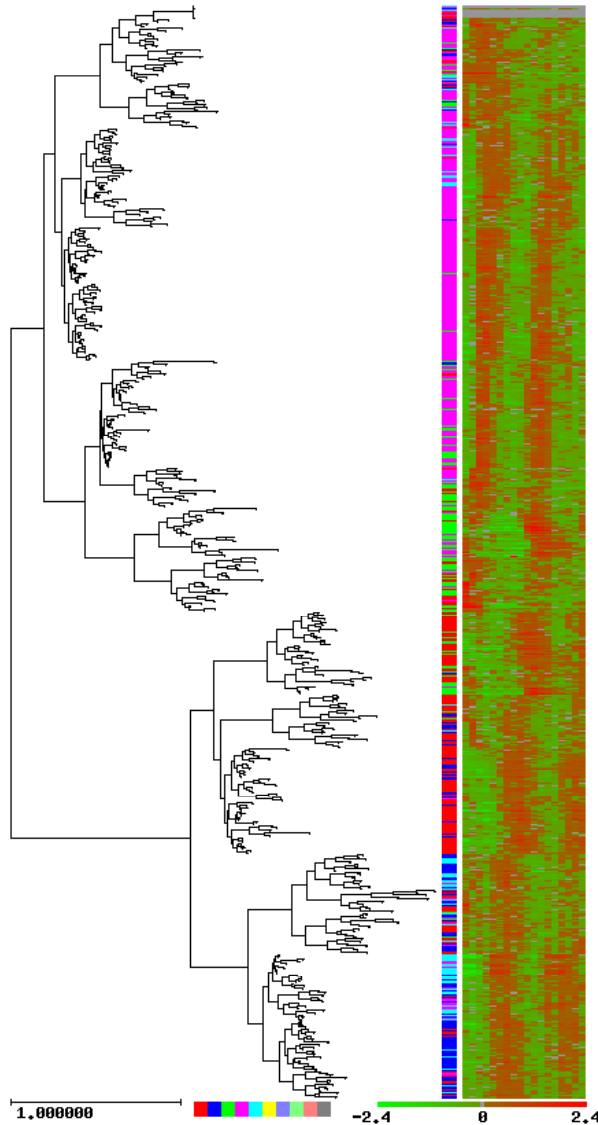
Experimental conditions

(from tens up to no more than a few hundreds)

Characteristics of the data:

- Most of the genes are not informative with respect to the trait we are studying (account for unrelated physiological conditions, etc.)
- Number of variables (genes) is several orders of magnitude larger than the number of experiments
- Low signal to noise ratio

An unsupervised problem: clustering of genes.



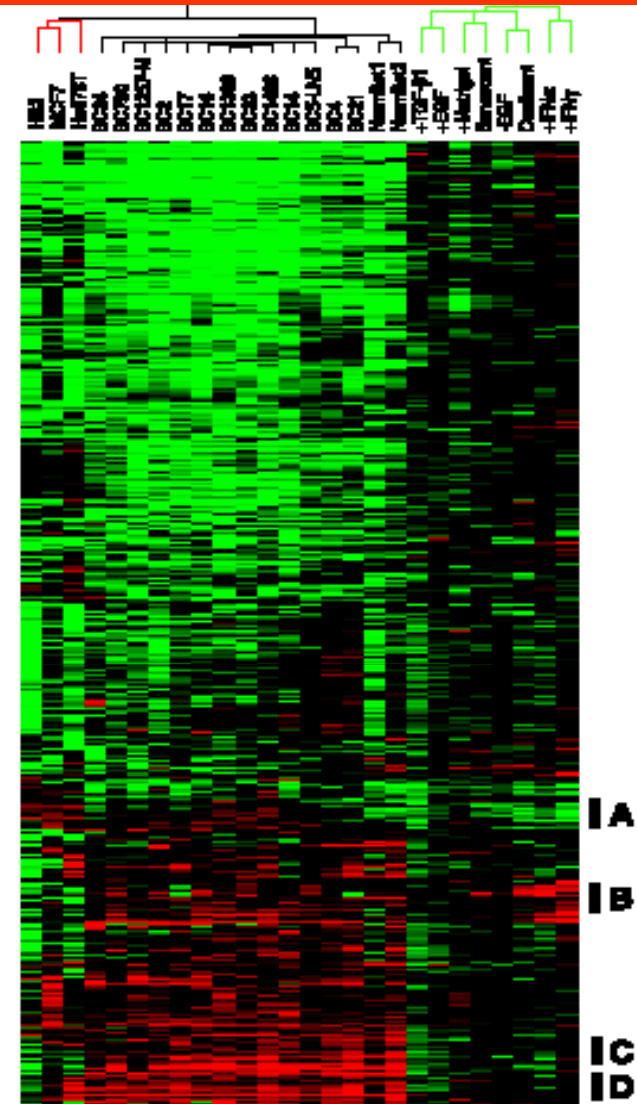
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

Clustering of experiments: The rationale

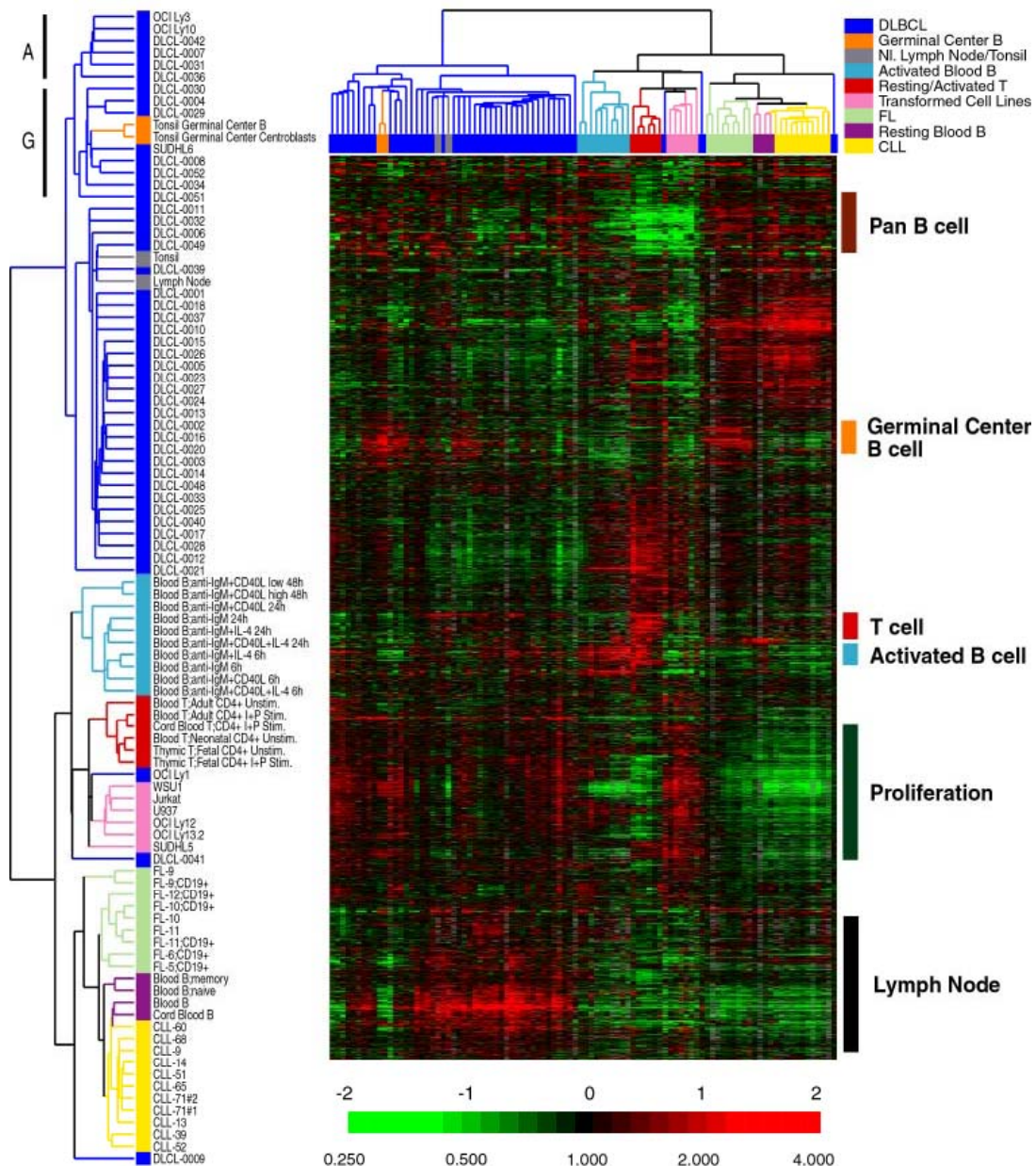
If enough genes have their expression levels altered in the different experiments, we might be able of finding these classes by comparing gene expression profiles.

Distinctive gene expression patterns in human mammary epithelial cells and breast cancers

Overview of the combined *in vitro* and breast tissue specimen cluster diagram. A scaled-down representation of the 1,247-gene cluster diagram. The black bars show the positions of the clusters discussed in the text: (A) proliferation-associated, (B) IFNregulated, (C) B lymphocytes, and (D) stromal cells.



Perou et al., PNAS 96 (1999)



Taxonomic Relationships Between Normal & Malignant Lymphoid Populations

Alizadeh et al.,
Nature 2000
(96 samples)

Clustering: unsupervised classification

- You do not have “external” information on how the data are arranged.
- You only have the values measured in the experiment
- You need a distance, which is a quantitative and non-subjective measure of the “closeness” of a pair of data. Distances use **all the components** of the vectors compared.
- You use the distances within a clustering method for producing groups (clusters) of data related. This relationship depends on the distance and the clustering method.

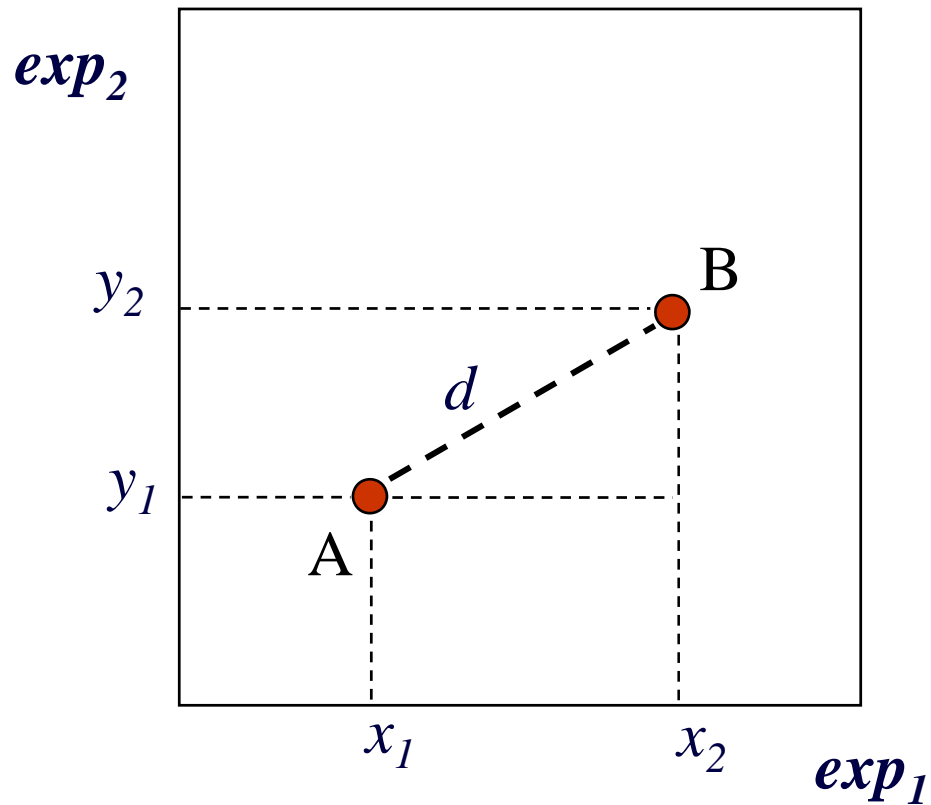
Distances

Distances used in microarray are related to:

- Euclidean (absolute differences)
- Correlation (trends)

Euclidian Distance

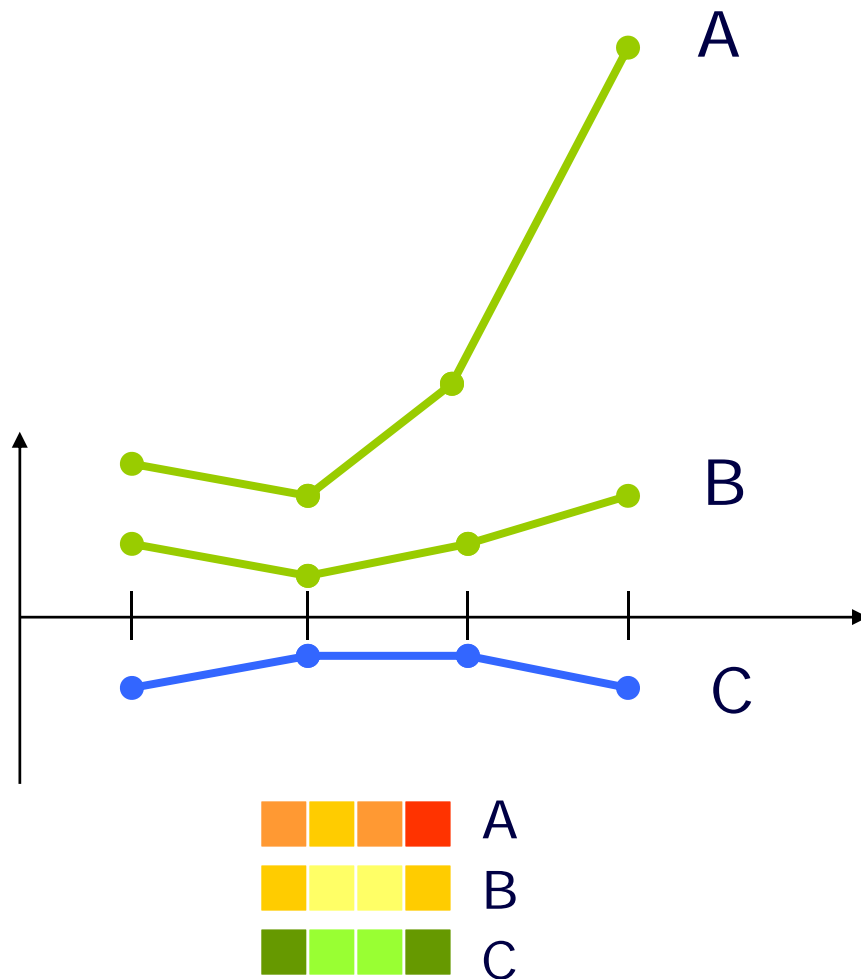
	<i>exp₁</i>	<i>exp₂</i>
A	Red	Yellow
B	Light Yellow	Light Green
C	Orange	Yellow



	<i>exp₁</i>	<i>exp₂</i>
A	x_1	y_1
B	x_2	y_2
C	x_3	y_3

$$d_{x,y} = \sqrt{\sum (x_i - y_i)^2}$$

Distance types



Differences (euclidean)

$$d_{x,y} = \sqrt{\sum (x_i - y_i)^2}$$

$B < => C$

Correlation

• Pearson Correlation Coefficient

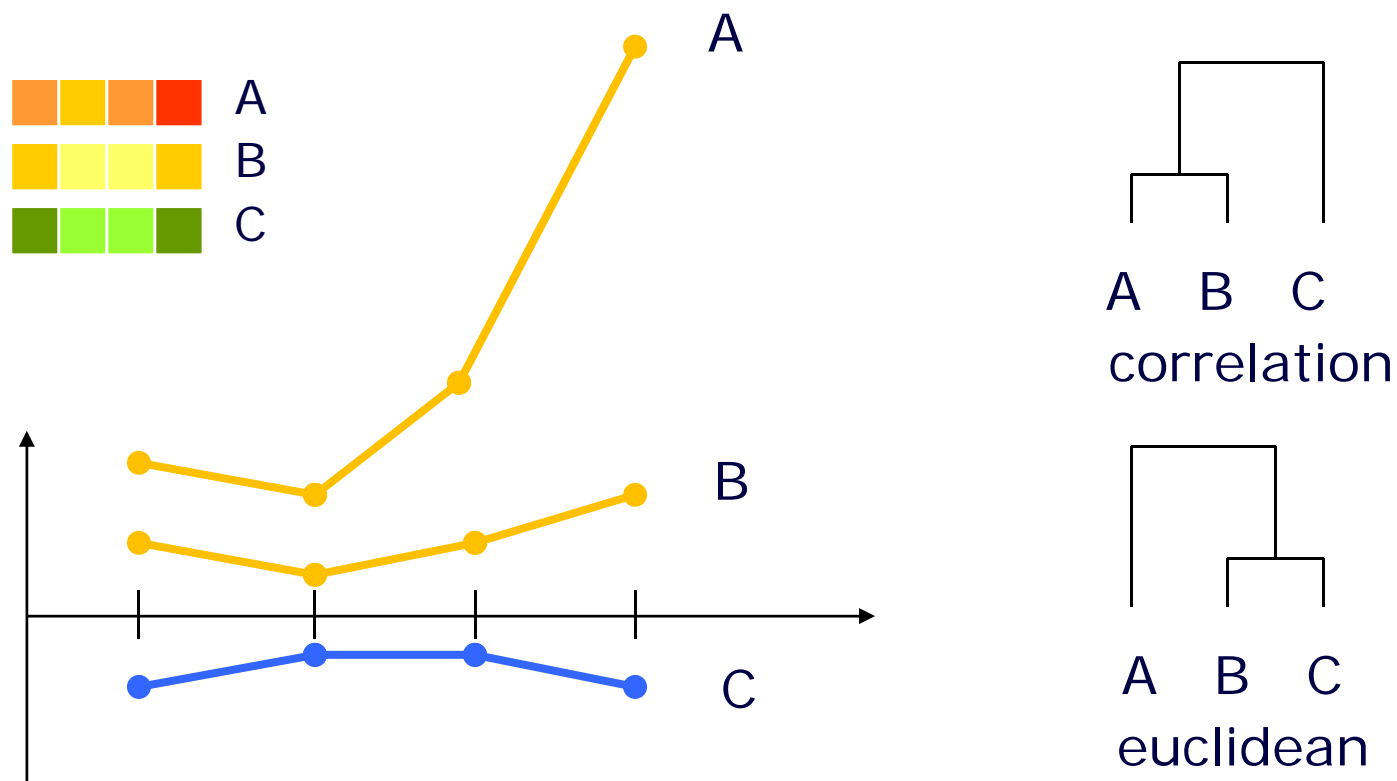
S_x = Standard deviation of x

S_y = Standard deviation of y

$$\frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

$A < => B$

Different distances account for different properties

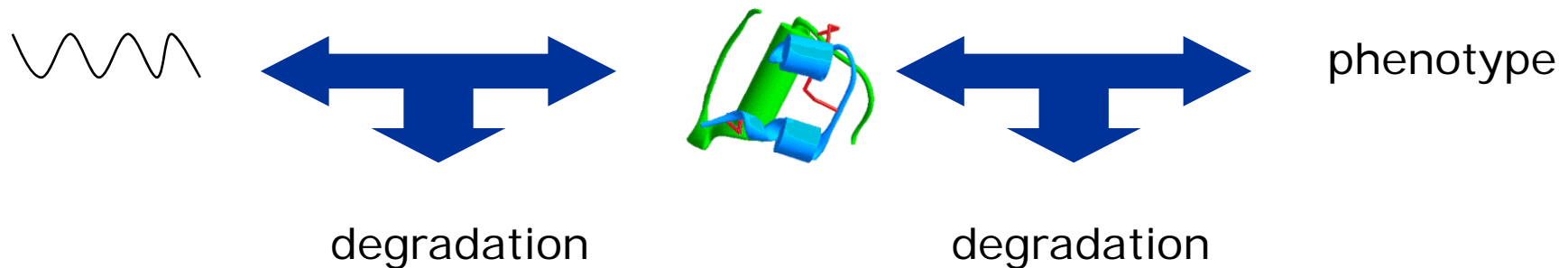


Correlation: tendencies; euclidean: global similarity

What distance should I use?

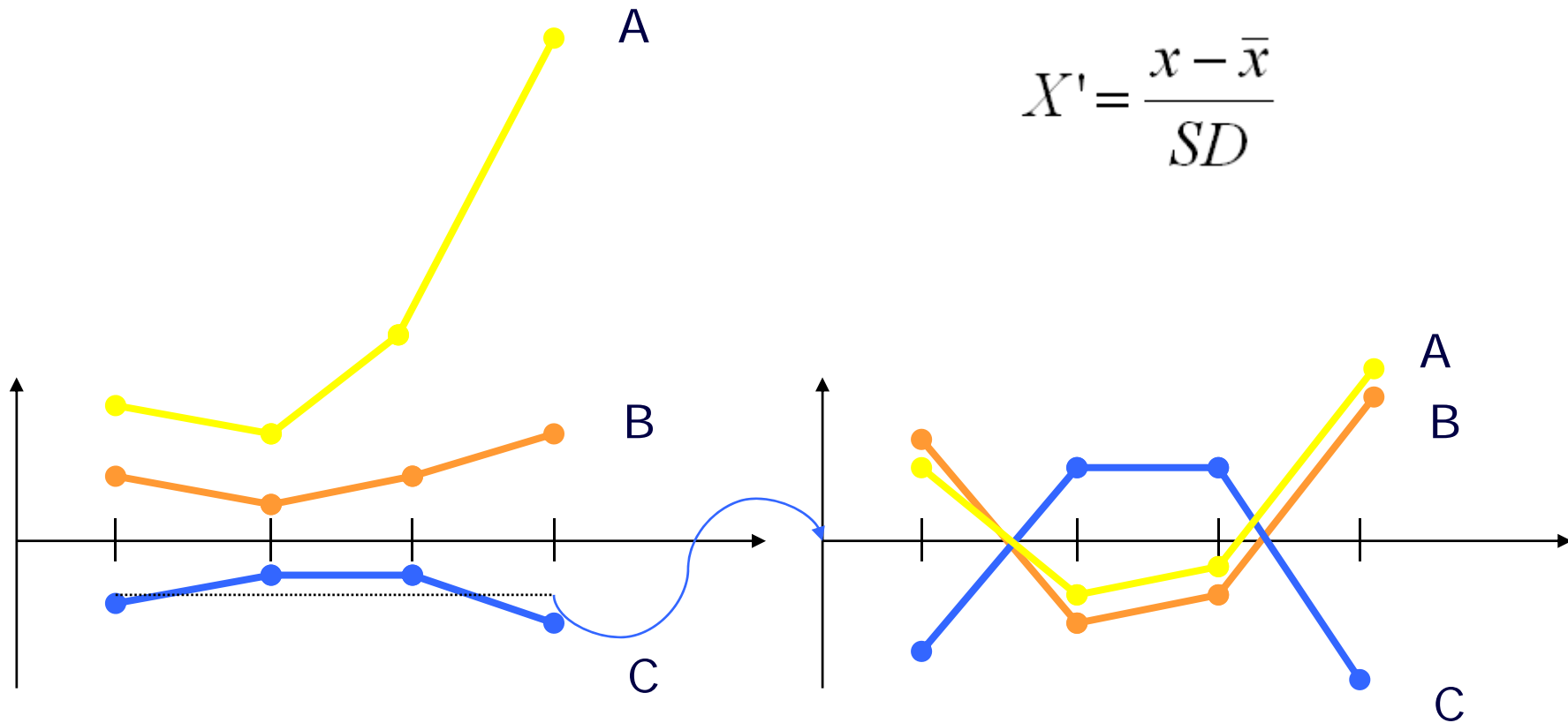
This is not the proper question.

The question is: what do I want to measure?



Correlation accounts for
coordinate changes

Standardising profiles



Unsupervised clustering methods

Non hierarchical

hierarchical

K-means, PCA

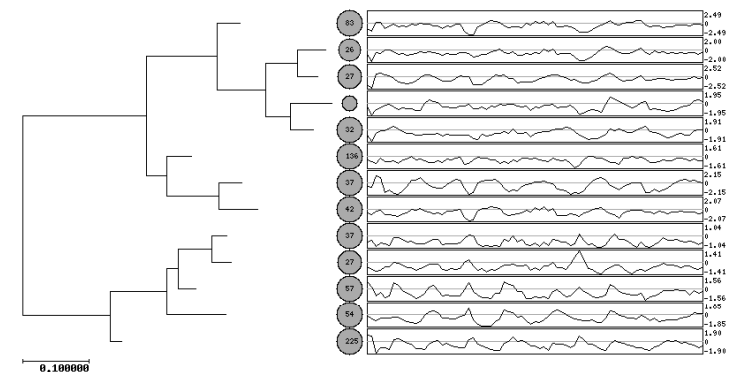
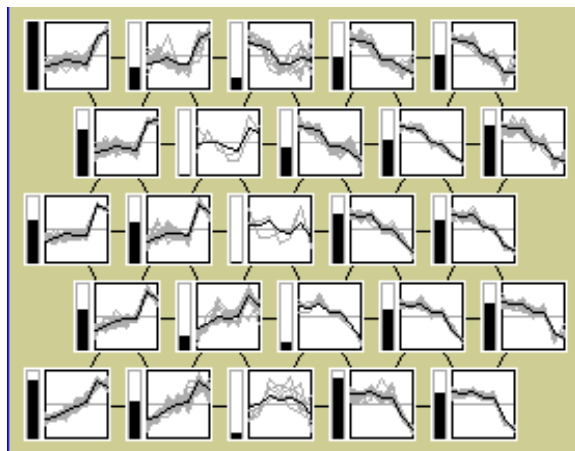
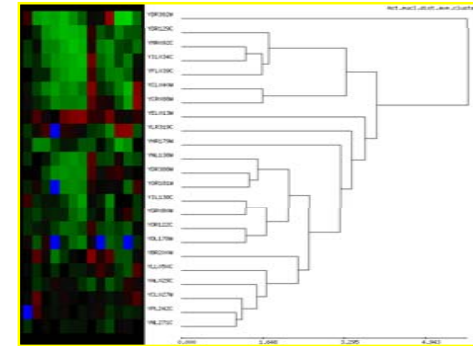
UPGMA

SOM

SOTA

quick and robust

Different levels of information

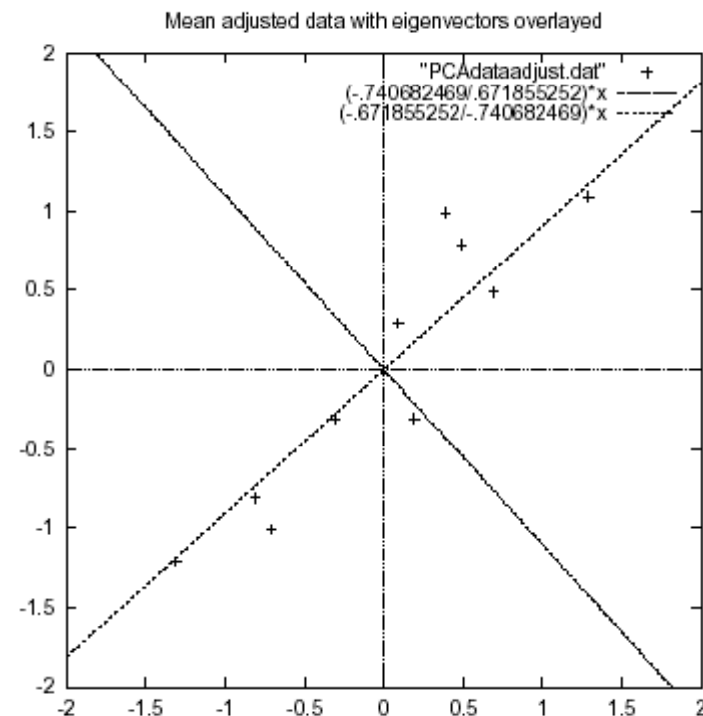
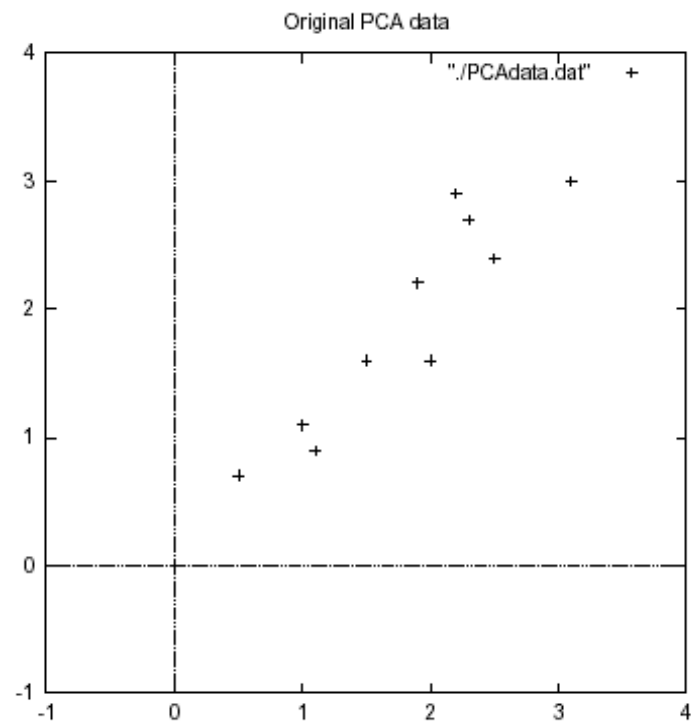


Clustering methods

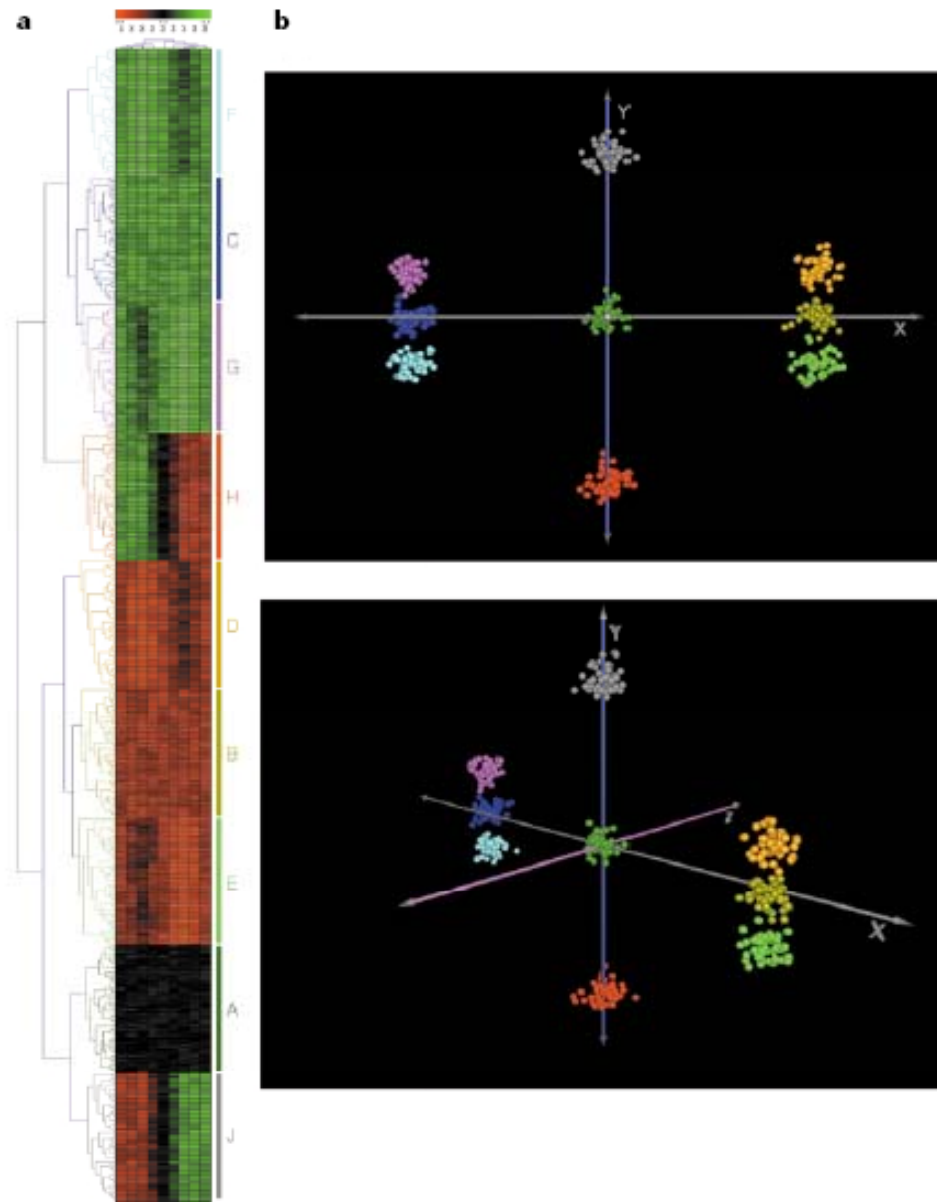
	Non hierarchical	Hierarchical	
deterministic	K-means, PCA	UPGMA	
NN	SOM	SOTA	Robust
		Provides different levels of information	Properties

PCA: an exploratory technique

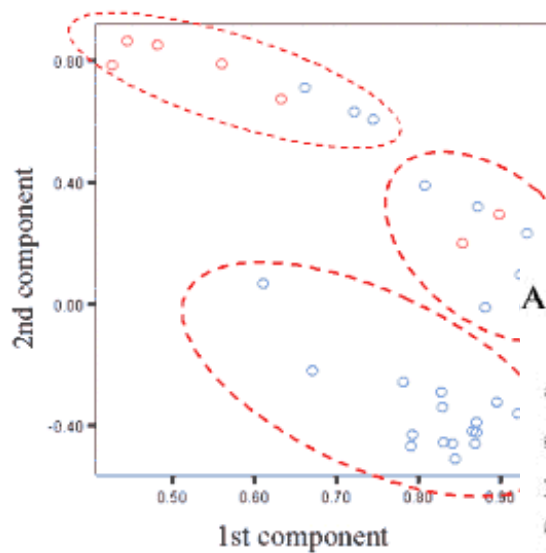
A common method from statistics for analysing data is **principal component analysis** (PCA). The aim is to find a set of M orthogonal vectors in data space that account for as much as possible of the data's variance. Projecting the data from their original N -dimensional space onto the M -dimensional subspace spanned by these vectors then performs a **dimensionality reduction** that often retains most of the intrinsic information in the data. See: <http://diwww.epfl.ch/mantra/tutorial/english/pca/html/>



PCA: an exploratory technique

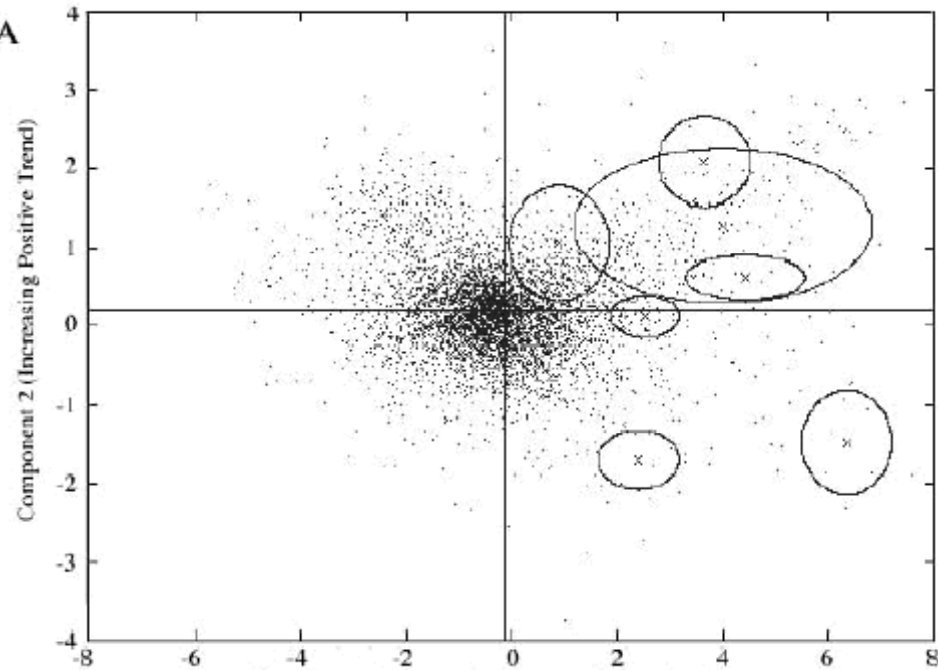


PCA: an exploratory technique



The example for the presentation...

...and the real,
cruel world:
subjective cluster
definition



K-Means clustering

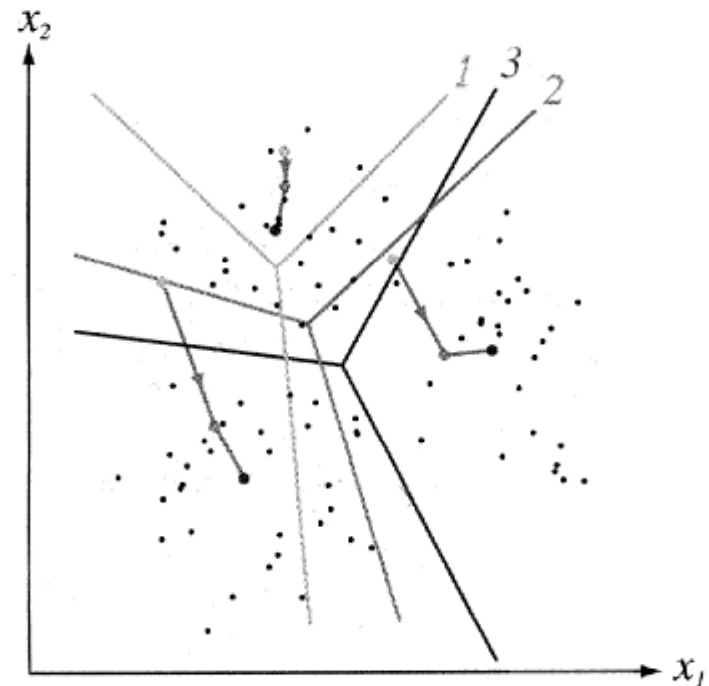
The idea is to find the best division of N samples by K clusters C_i such that the total distance between the clustered samples and their respective centers (that is, the total variance) is minimized.

This criterion is expressed like this $J = \sum_{i=1}^K \sum_{x \in C_i} |x_n - \gamma_i|^2$


where γ_i is the center of class i . Analogy to linear regression can be seen: there the residuals are the distance from each point to the regression line. In clustering, the residuals are the distance between each point and its cluster center. The k -means algorithm starts by randomly assigning instances to the classes, computes the centers according to

$$\gamma_i = \frac{1}{N_i} \sum_{x \in C_i} x_n$$

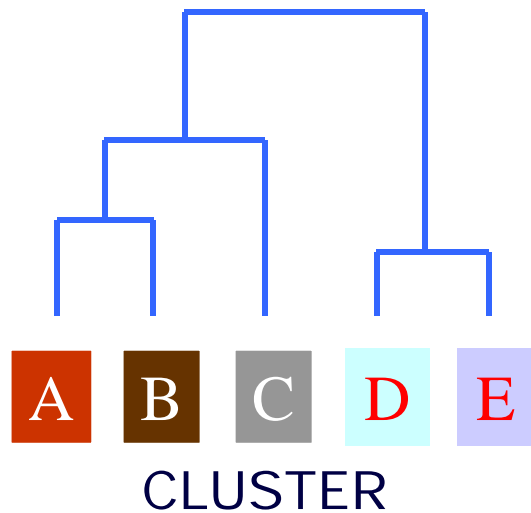
then reassigns the instances to the nearest cluster's center, recalculates centers, reassigns the instances, etc. until J stops decreasing (or centers stop to move). Here is a two-dimensional example of clustering



Clustering methods

	Non hierarchical	Hierarchical	
deterministic	K-means, PCA	 UPGMA	
NN	SOM	SOTA	Robust
		Provides different levels of information	Properties

Aggregative hierarchical clustering



	A	B	C	D	E
A					
B	2				
C	3	3			
D	5	5	5		
E	5	5	5	1	



	A	B	C	DE
A				
B	2			
C	3	3		
DE	5	5	5	



	AB	C	DE
AB			
C	3		
DE	5	5	



	ABC	DE
ABC		
DE	5	

Relationships among profiles are represented by branch lengths.

The closest pair of profiles are recursively linked until the complete hierarchy is reconstructed

Aggregative hierarchical clustering

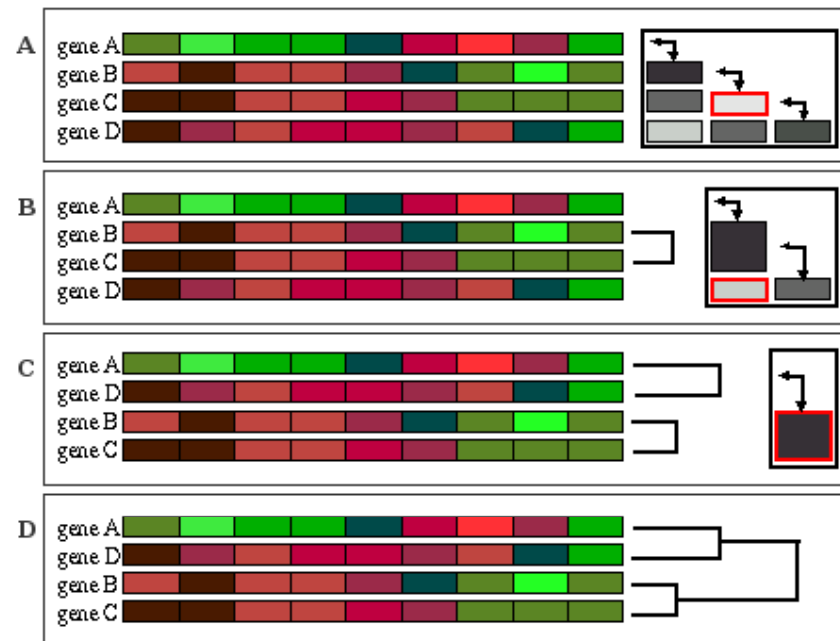
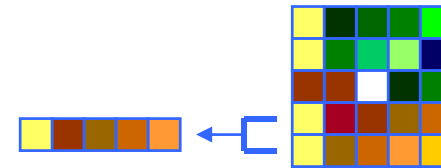
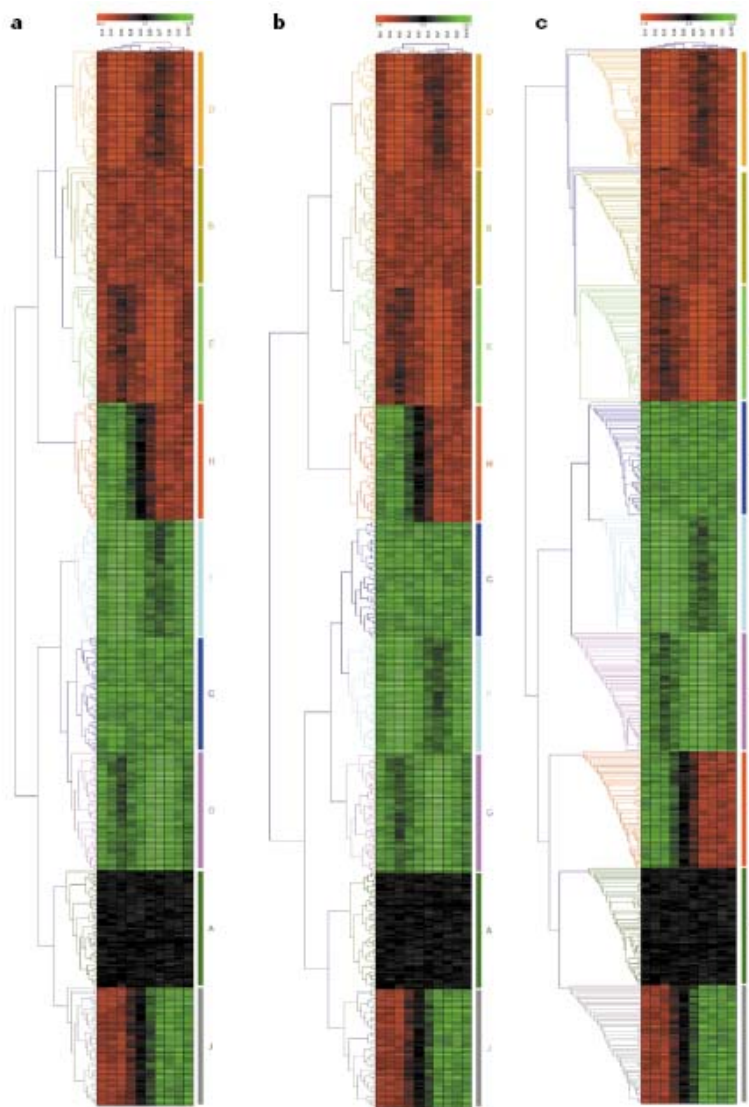


Fig. 1.5.1 UPGMA: the all-to-all distance matrix (black box) is calculated and two closest elements (red box) are merged. **(A)** The two closest elements are genes B and C. **(B)** The genes B and C are merged and the all-to-all distance matrix is calculated again using the new cluster instead of the genes B and C. Now, the two closest elements are genes A and D. **(C)** The genes A and D are also merged. The elements must be reordered to fit the topology of the tree. The all-to-all distance matrix is calculated again with the two remaining elements. **(D)** The process ends when all the complete dendrogram is built.

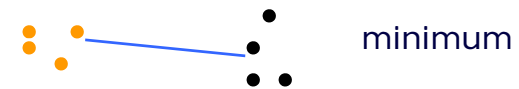
Different aggregative criterion



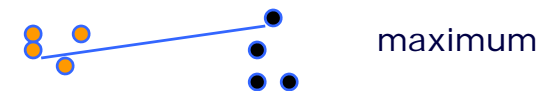
a) Average linkage



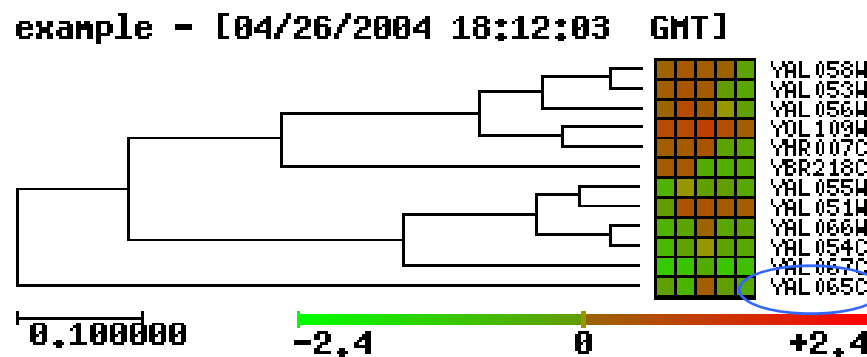
b) Single linkage



c) Complete linkage



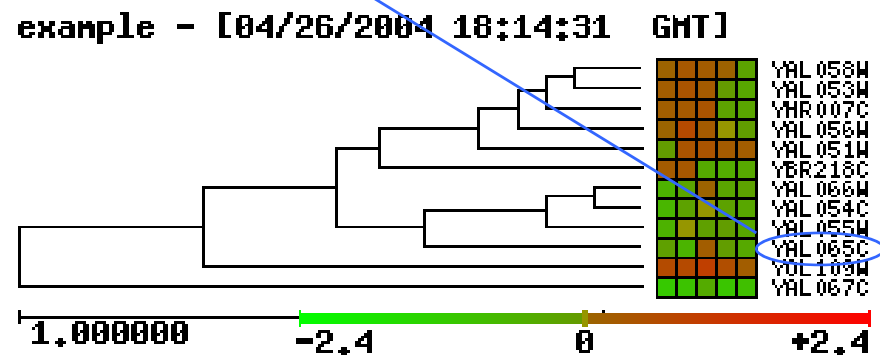
The effect of the distance in aggregative clustering



Correlation

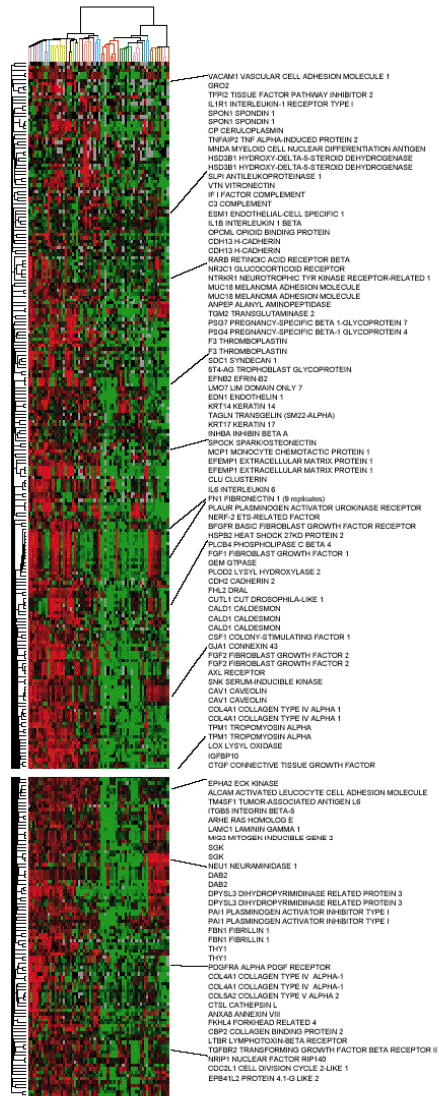
The best correlated is not the most similar...

...and the most similar is not the best correlated



Euclidean

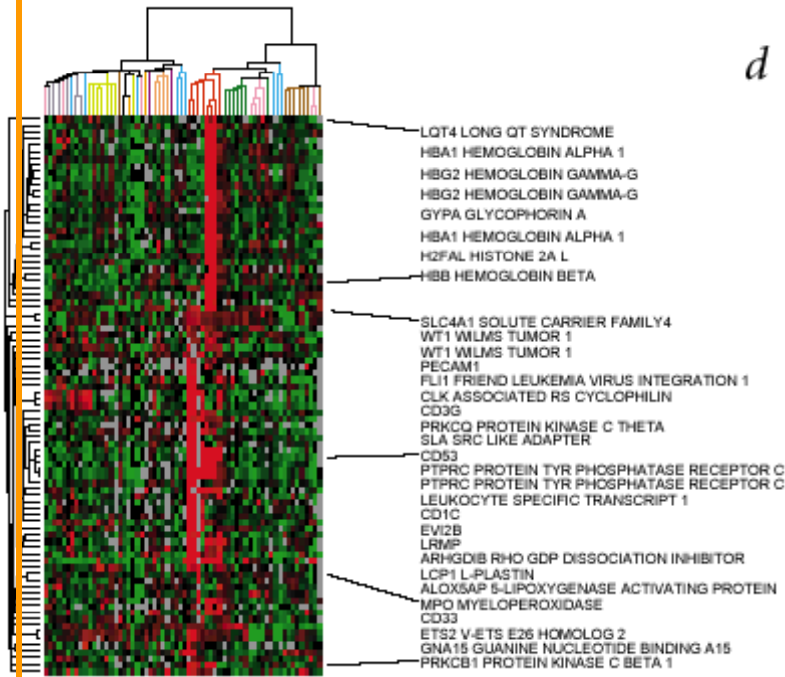
Aggregative hierarchical clustering



mesenchymal cluster (67 ESTs)


Problems:

- lack of robustness
- difficult interpretation
- subjective cluster definition



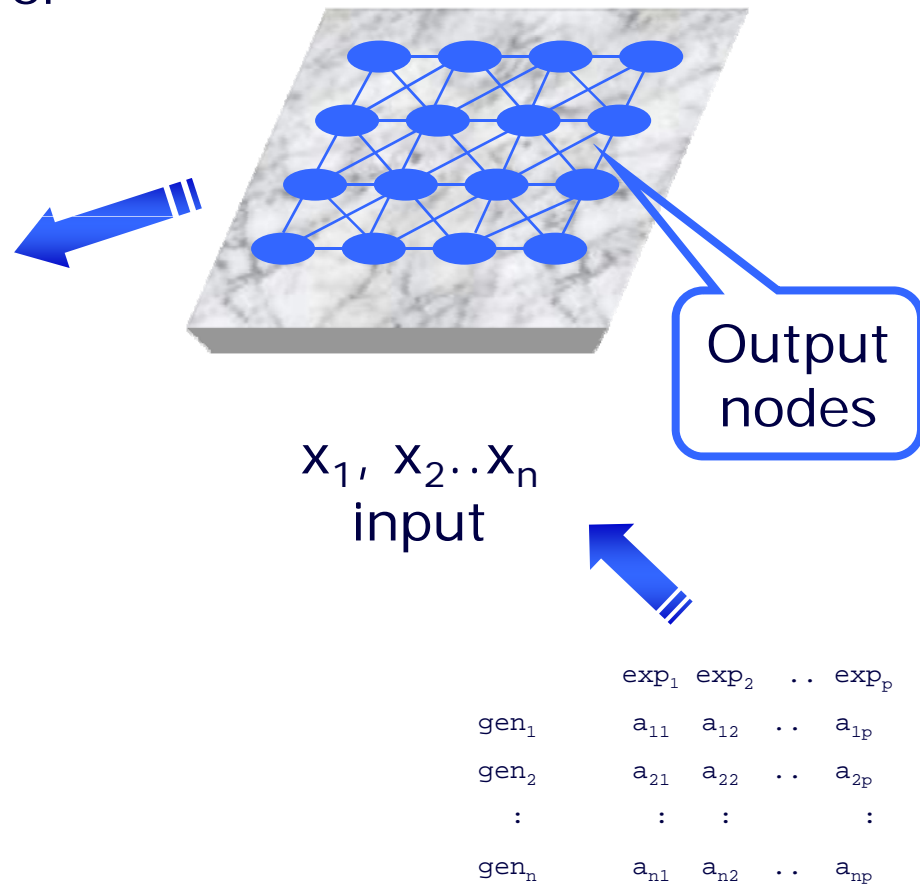
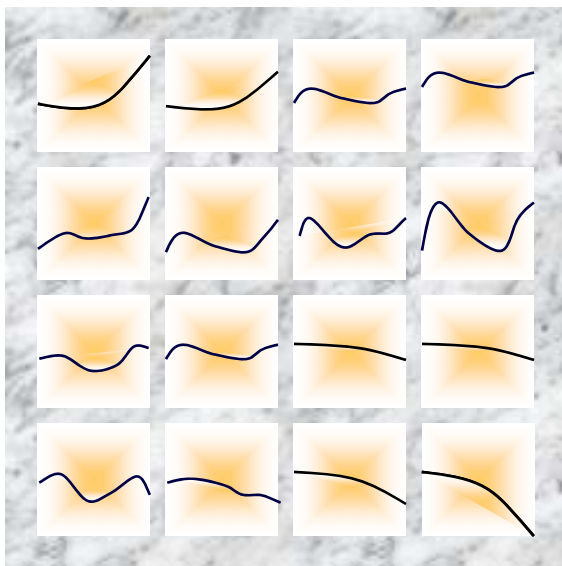
leukaemia cluster (6 ESTs)

Clustering methods

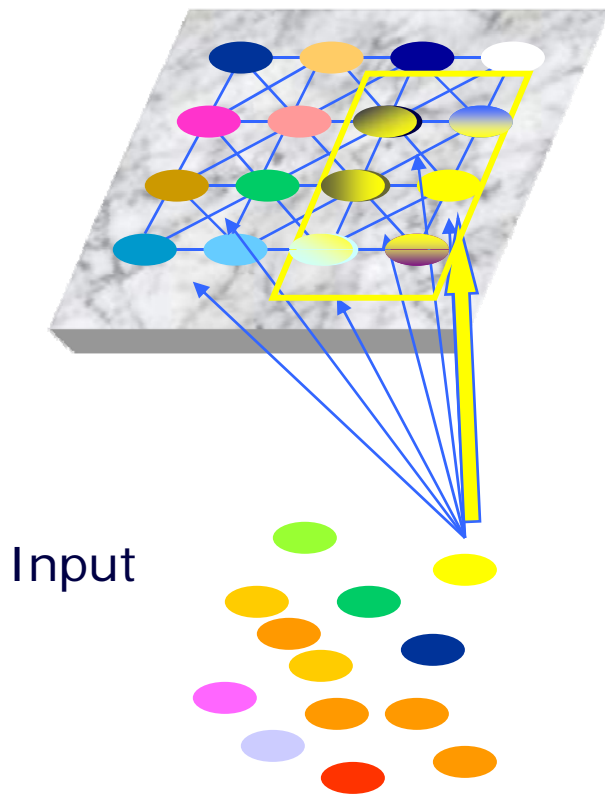
	Non hierarchical	Hierarchical	
	K-means, PCA	UPGMA	
NN	 SOM	SOTA	Robust
		Provides different levels of information	Properties

Self organising maps: SOM

Bidimensional hexagonal or rectangular network

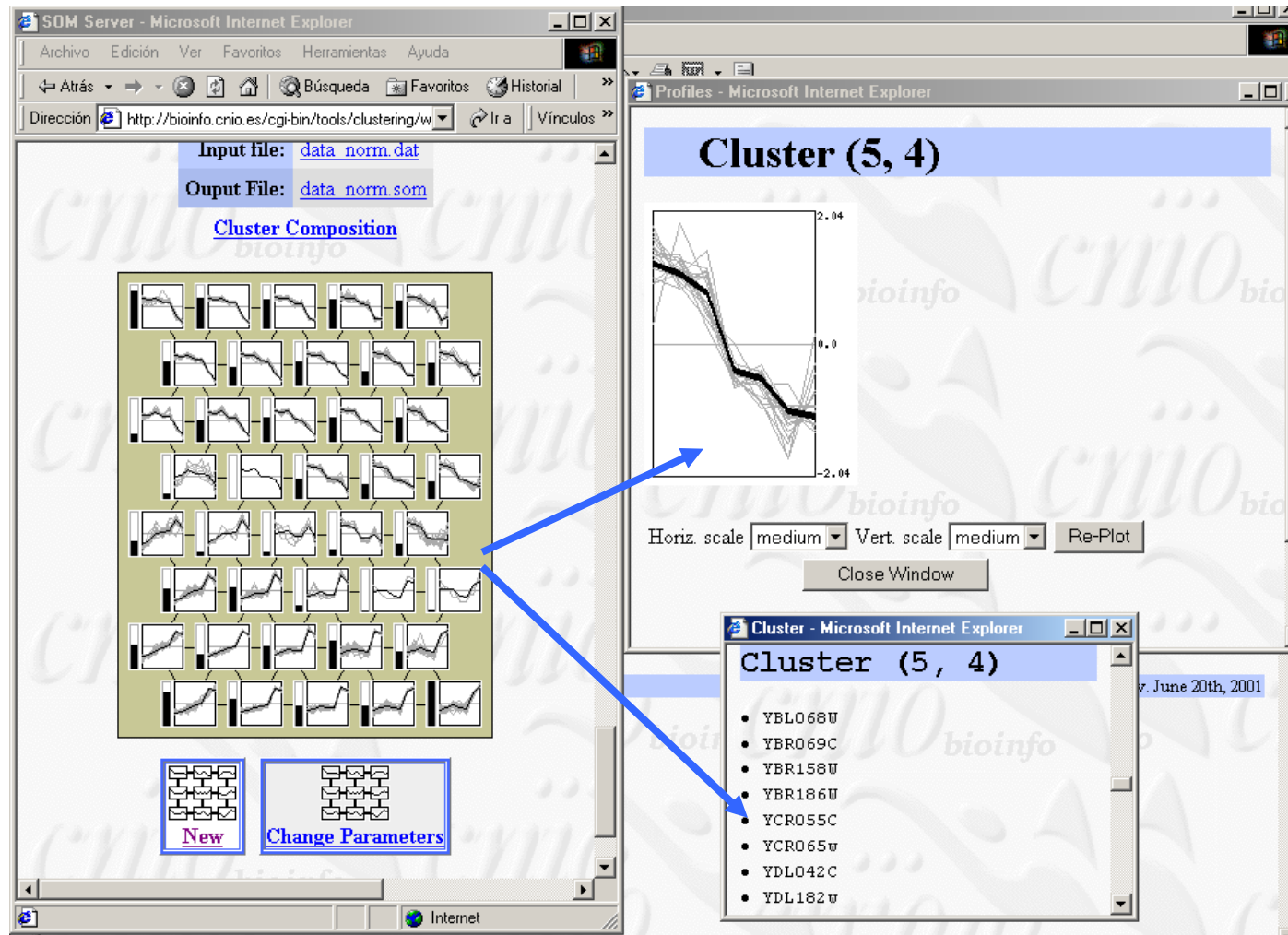


SOM: The algorithm



- Step 1. Initialize nodes to random values. Set the initial radius of the neighborhood.
- Step 2. Present new input: Compute distances to all nodes. Euclidean distances are commonly used
- Step 3. Select output node j^* with minimum distance d_j . Update node j^* and neighbors. Nodes updated for the neighborhood $NE_{j^*}(t)$ as:
$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)); \text{ for } j \in NE_{j^*}(t)$$
$$\eta(t)$$
 is a gain term that decreases in time.
- Step 4. Repeat by going to Step 2 until convergence.

SOM results

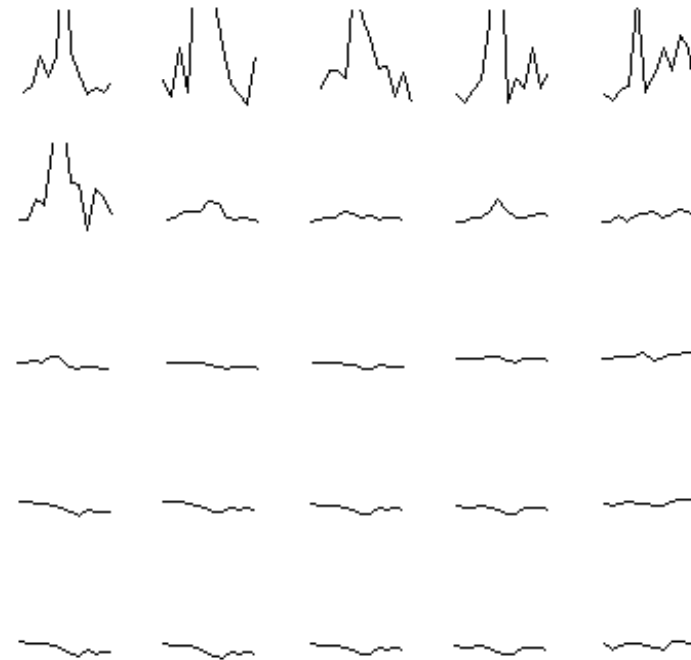


SOM: Example


Response of human
fibroblasts to serum

Iyer et al., 1999 *Science*
283:83-87

If a given class is
overrepresented, it
takes over many
neurons



Clustering methods

	Non hierarchical	Hierarchical	
	K-means, PCA	UPGMA	
NN	SOM	 SOTA	Robust
		Provides different levels of information	Properties

SOTA: The algorithm

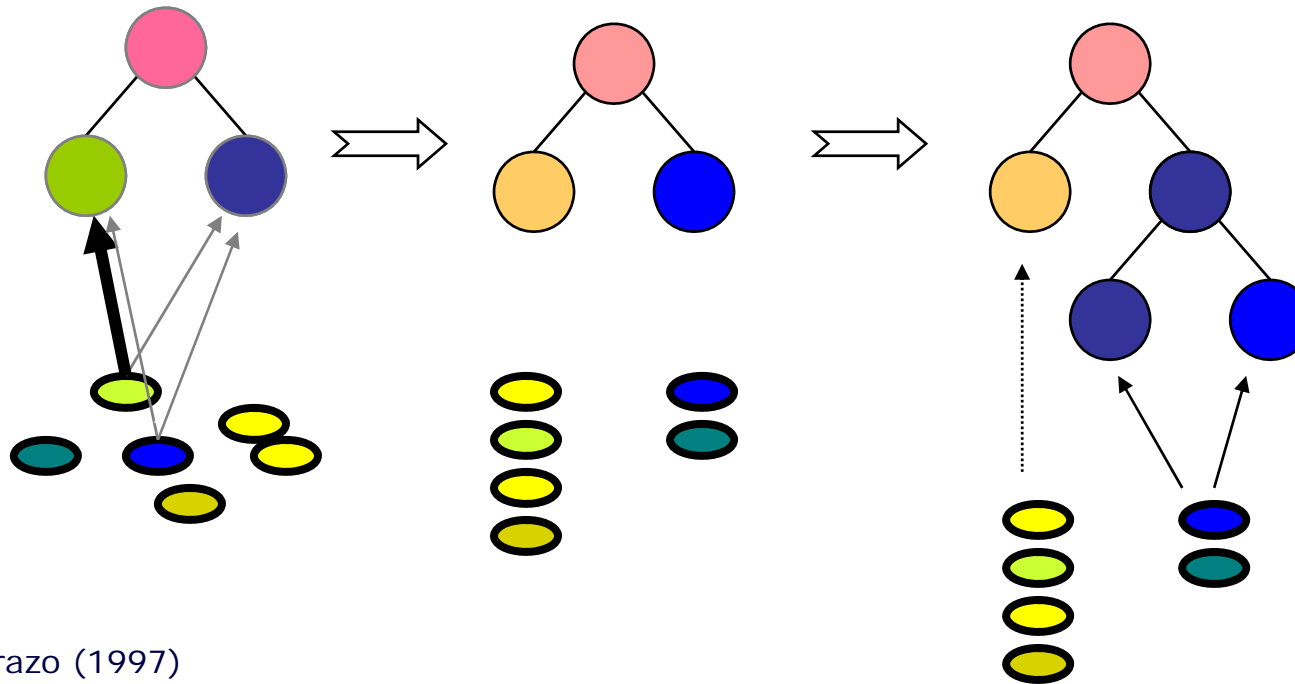
The Self Organising Tree Algorithm (SOTA) is a hierarchical divisive method based on a neural network

SOTA, unlike other hierarchical methods, grows from top to bottom until an appropriate level of variability is reached

- Step 1. Initialize nodes to random values.
- Step 2. Present new input: Compute distances to all **terminal** nodes.
- Step 3. Select output node j^* with minimum distance d_j . Update node j^* and neighbors. Nodes updated for the neighborhood $NE_{j^*}(t)$ as:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)); \text{ for } j \in NE_{j^*}(t)$$
 $\eta(t)$ is a gain term that decreases in time.
- Step 4. Repeat by going to Step 2 until convergence.
- Step 5. Reproduce the node with highest variability.

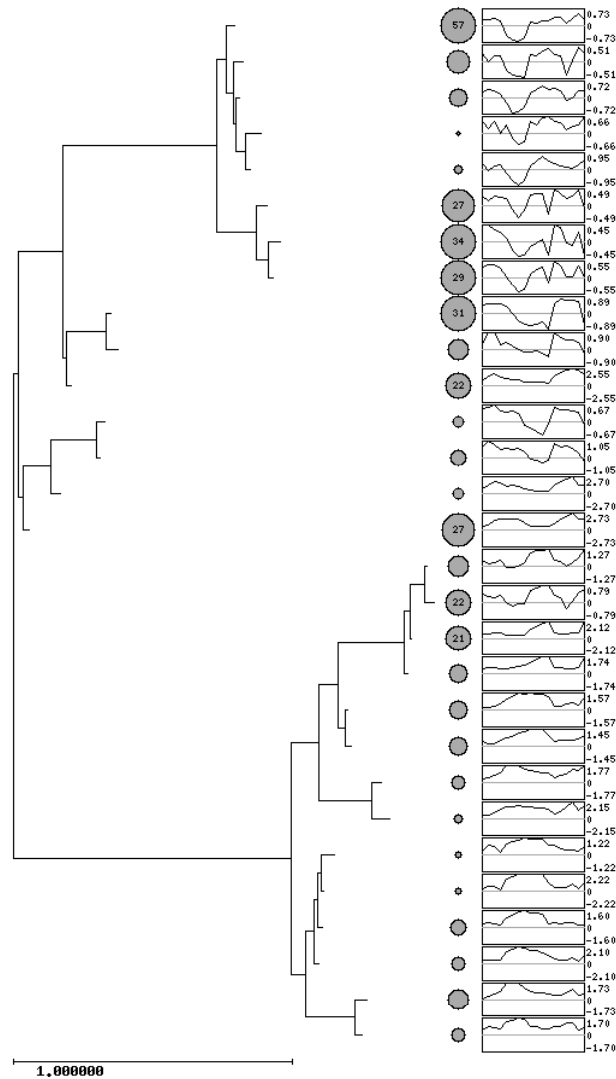
Input



Dopazo, Carazo (1997)

Herrero, Valencia, Dopazo (2001)

Advantages of SOTA



Robustness against noise

Divisive algorithm

SOTA grows from top to bottom: growing can be stopped at any desired level of variability.

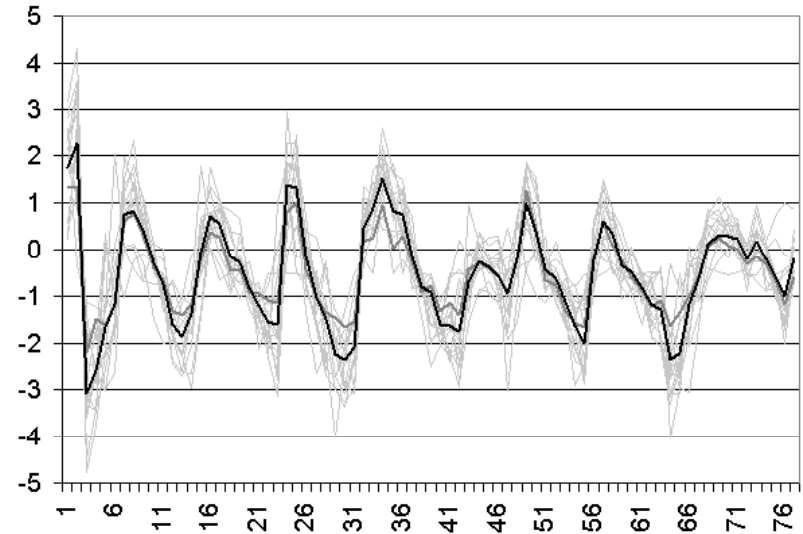
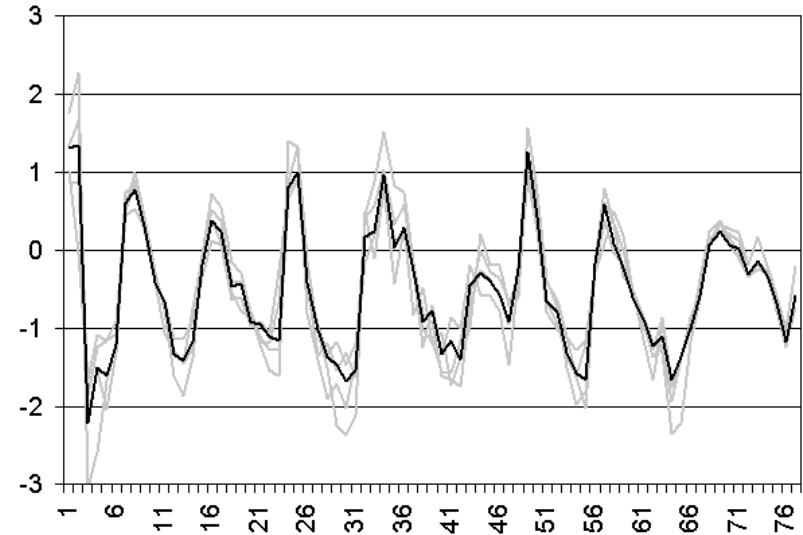
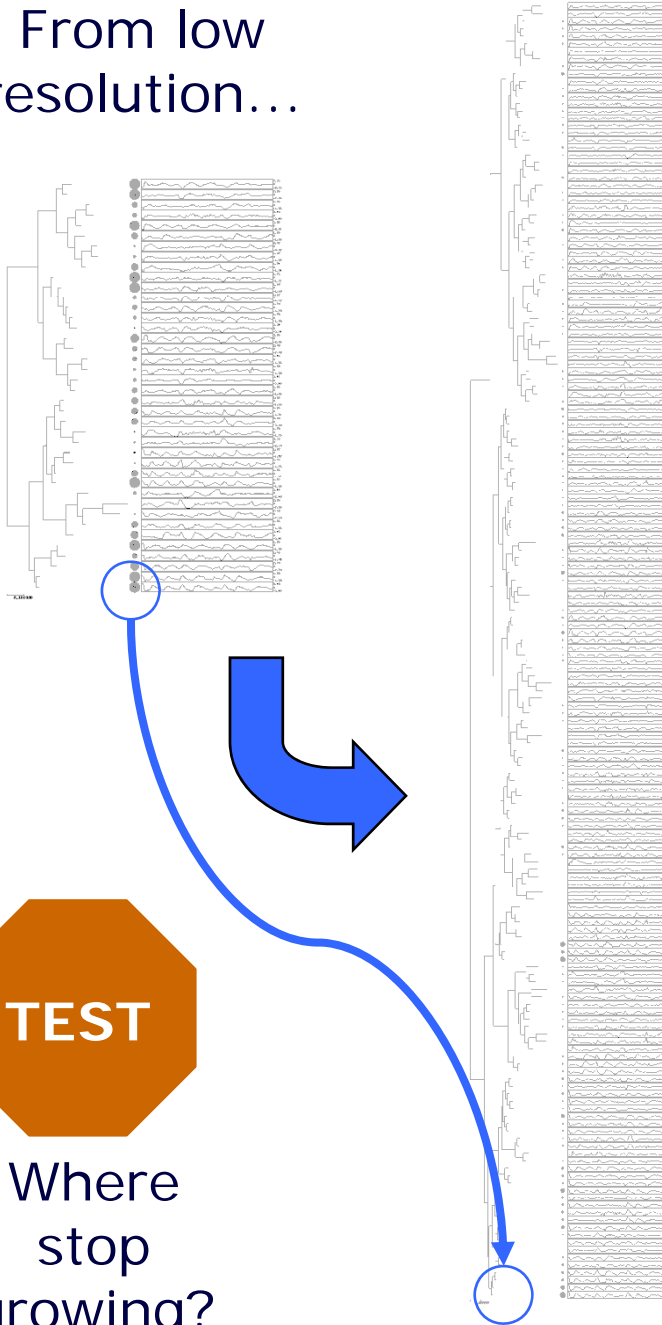
Clusters' patterns

Each node of the tree has a pattern associated with it which corresponds to the cluster under itself.

Distribution preserving

The number of clusters depends on the variability of the data.

From low resolution...



...to high resolution.

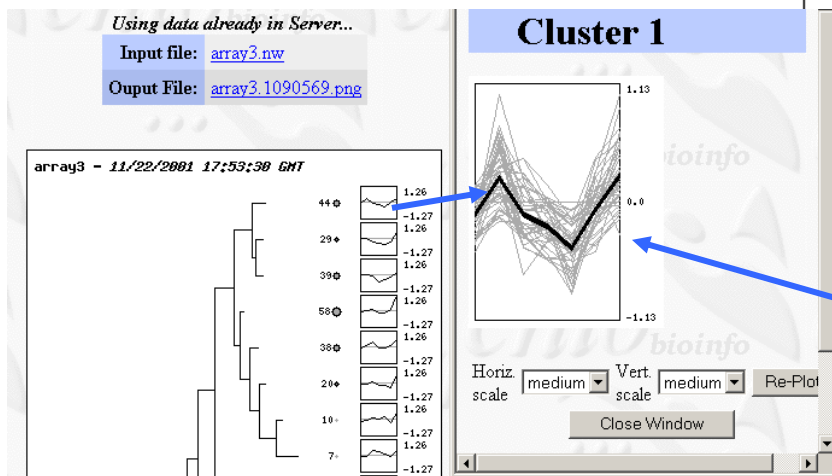
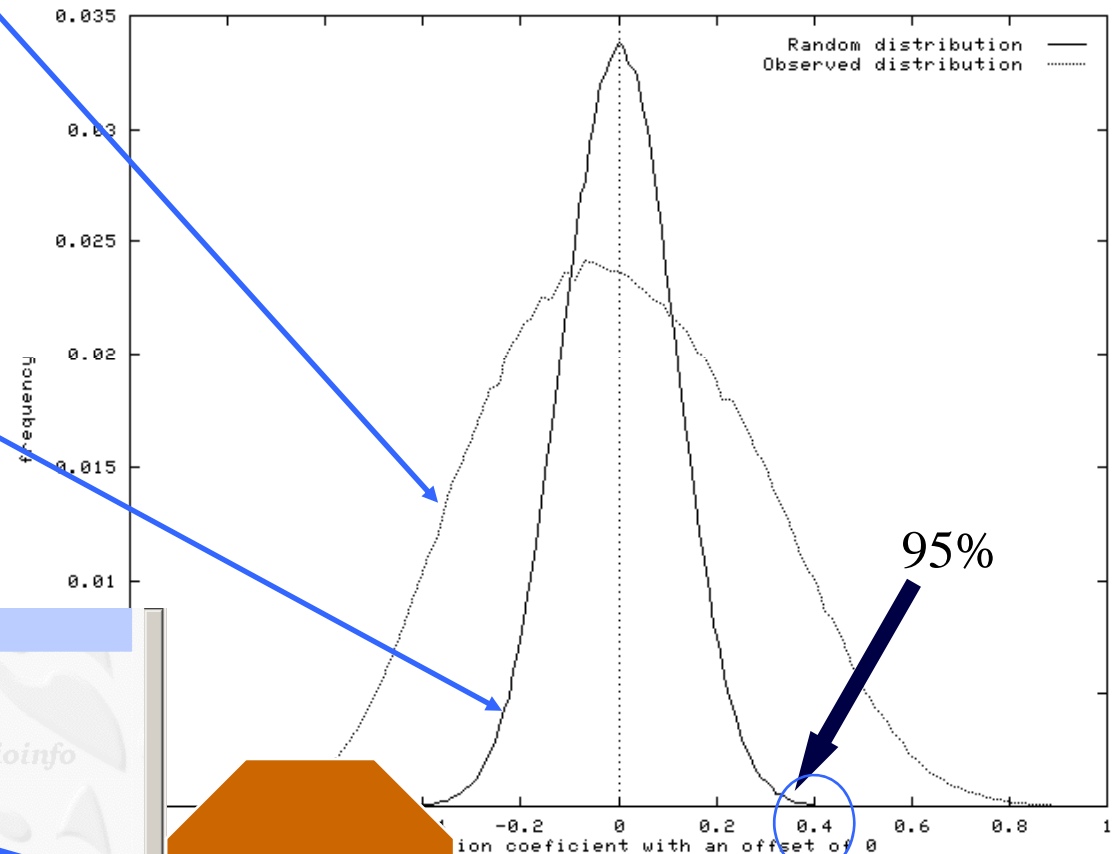
Permutation test for cluster size definition

	exp ₁	exp ₂	...	exp _p
gen ₁	a ₁₁	a ₁₂	...	a _{1p}
gen ₂	a ₂₁	a ₂₂	...	a _{2p}
:	:	:		:
gen _n	a _{n1}	a _{n2}	...	a _{np}



	exp ₁	exp ₂	...	exp _p
gen ₁	a ₁₄	a ₁₇	...	a _{1q}
gen ₂	a ₂₃	a ₂₁	...	a _{2r}
:	:	:		:
gen _n	a _{n9}	a _{n4}	...	a _{ns}

definition



TEST
are $d_{ij} > 0.4$?

Cluster quality measures

It is worth noting that many clustering methods produce partitions even with random data. This is commonly known as the “garbage in garbage out” effect and points out to the necessity of having some criteria in the application of these methods.

There are mainly two classes of quality cluster measures:

Internal

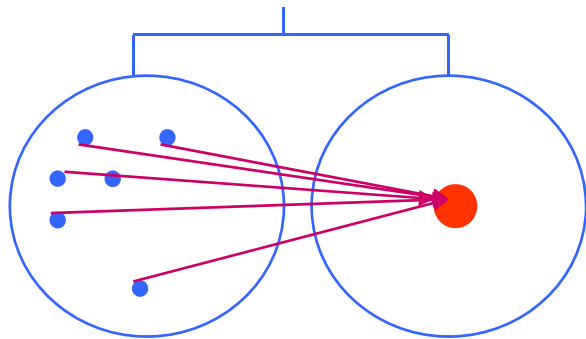
External

Cluster quality internal measures: the silhouette

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

How close are the items within a cluster [intracluster distance a(i)]

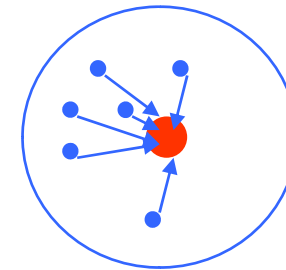
How far are the items among brother clusters [intercluster distance b(i)]



InterCluster Distance method:
average to centroids linkage

$$d(c_1, c_2) = \frac{1}{n_1 + n_2} \left(\sum_{x \in s_1} d(x, v_{c_2}) + \sum_{y \in s_2} d(y, v_{c_1}) \right)$$

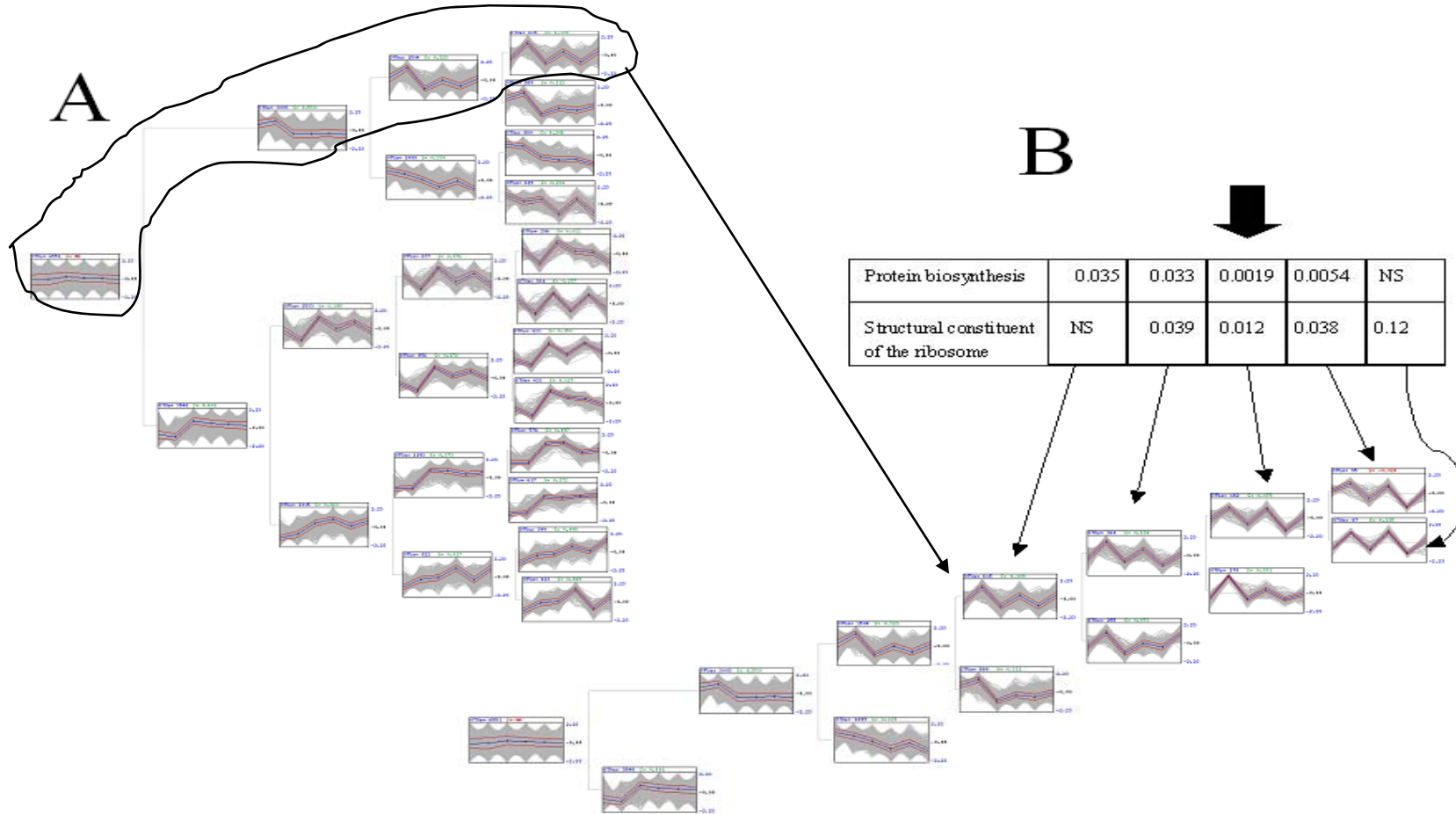
$$v = \frac{1}{n} \sum_{i=1}^n x_i$$



IntraCluster Distance method:
centroid diameter

$$d(c) = 2 \left(\frac{\sum_{x, y} d(x, y)}{n} \right)$$

Cluster quality external measures: Functional interpretation



Relative merits of clustering methods

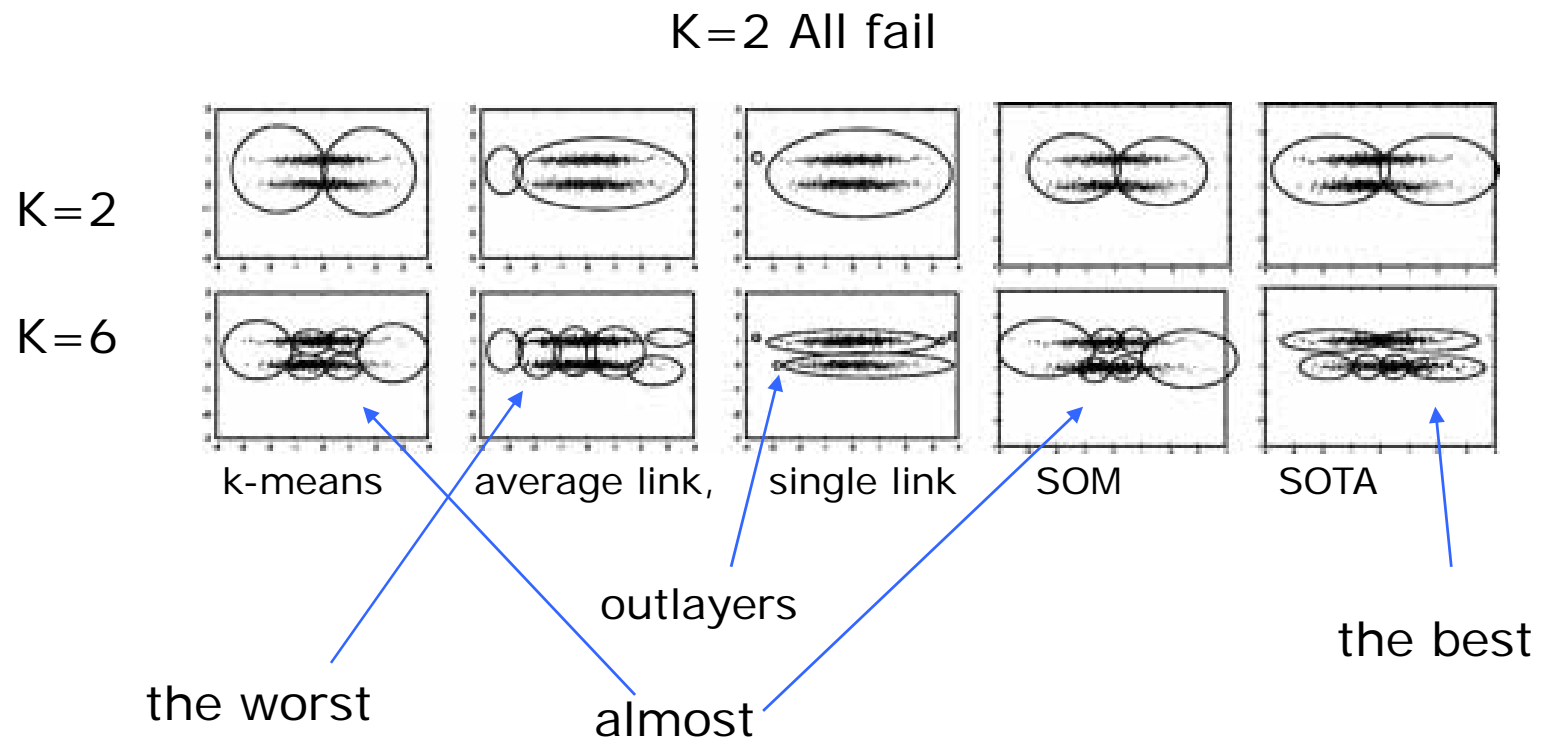
Different benchmarking studies report a **poor performance** of **hierarchical clustering** when **single linkage** was used.

Hierarchical clustering with average or complete linkage seems to work well, and...

SOM, SOTA and k-means seem to be superior according to internal indexes (Silhouette, Dunn, and other) or external criteria (enrichment of functional terms).

A significant problem associated with k-means or SOM algorithms is the arbitrary choice of the number of clusters

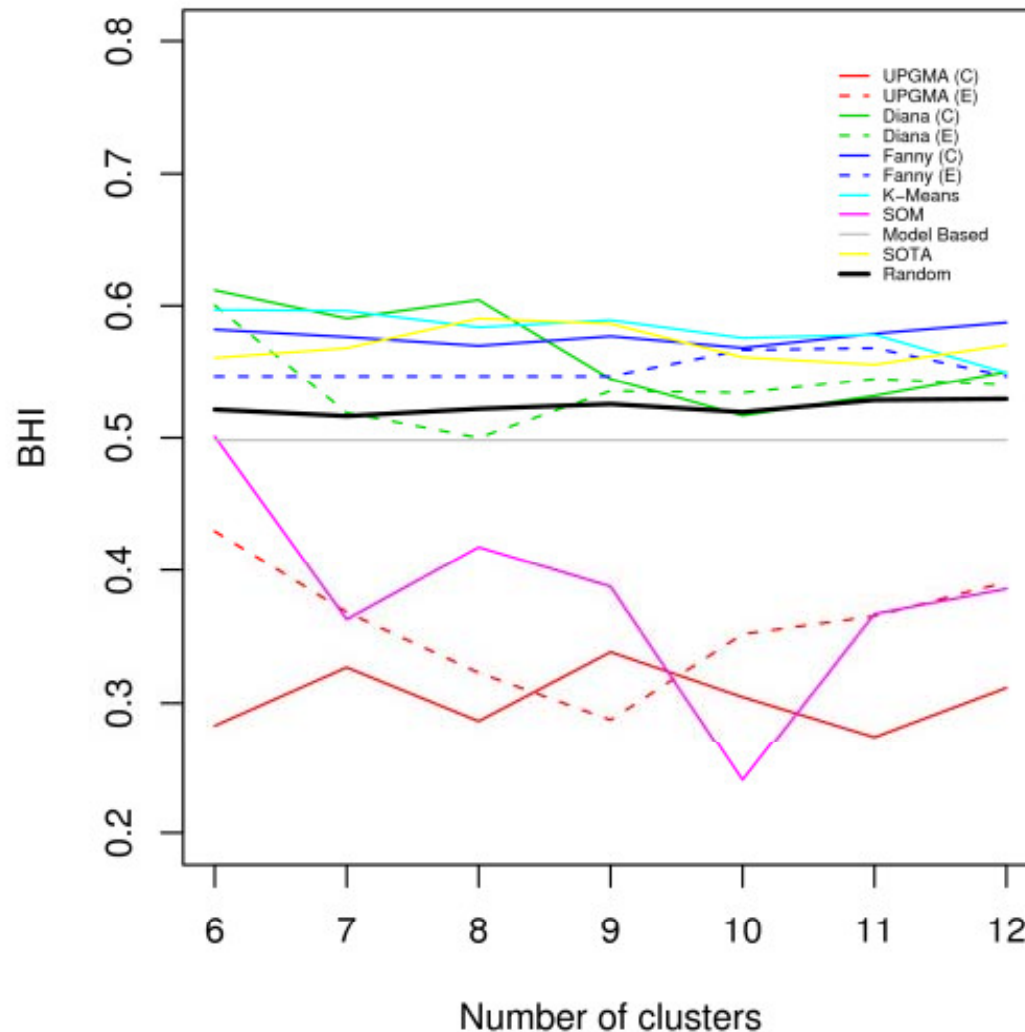
Relative performance of clustering methods (Handl et al., 2005 Bioinformatics)



Non-spherical clusters run methods into troubles.

Relative performance of clustering methods

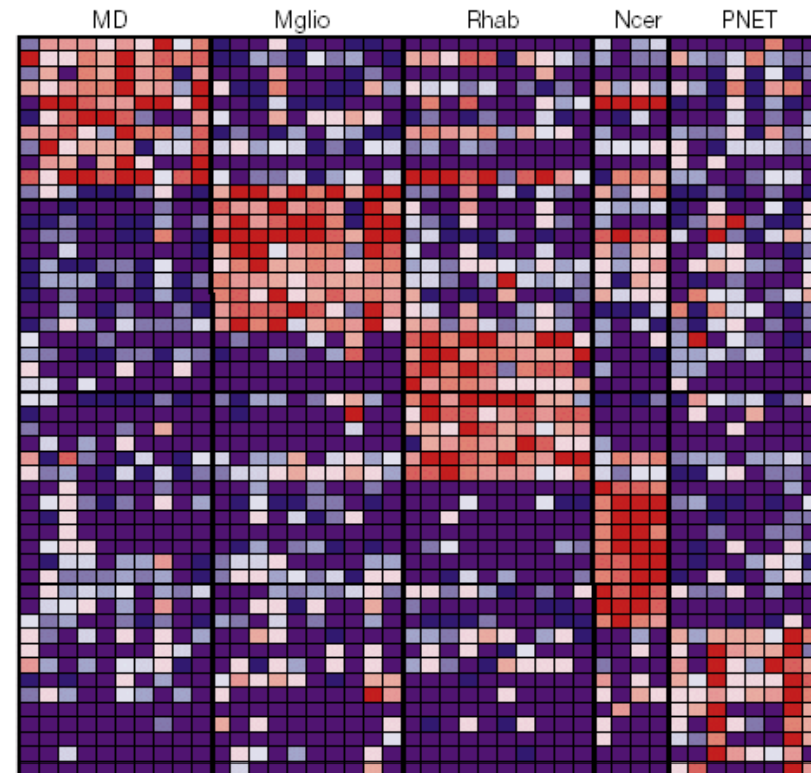
(Datta and Datta, 2006 BMC Bioinformatics)



Biological Homogeneity Index for various clustering algorithms applied to the positively expressed genes in yeast sporulation data with functional classes from FatiGO. The thick black line is the 95th percentile of BHI values under random clustering

Class discovery: biclustering

- Seek biclusters: individual two-way clusters
 - genes may have restricted co-regulation (common change in expression)
- Biclusters can overlap
 - genes may have more than one function
- Biclusters need not cover the data
 - constantly expressed genes/outliers can be ignored



Plaid models, biclustering, clustering on subsets of attributes, feature selection in clustering, other.

Other clustering methods

Quality-based clustering algorithms: QT_Clust, adaptative QT. Greedy algorithms that select for each gene a cluster around that satisfies the quality criteria. The biggest cluster is removed and start again...

Model-based clustering: Mixture models. These assume that data are generated by a finite mixture of probability distributions, where each distribution represents one cluster.

What we have learned? Lessons from the first-generation algorithms and specific demands for clustering microarray data

- Number of clusters. K-means, SOM and hierarchical methods do not provide any method for defining the “true” number of clusters

The wish list

- Methods must be fast
- Robustness and noise tolerance
- Deterministic
- Able to decide the number of clusters automatically