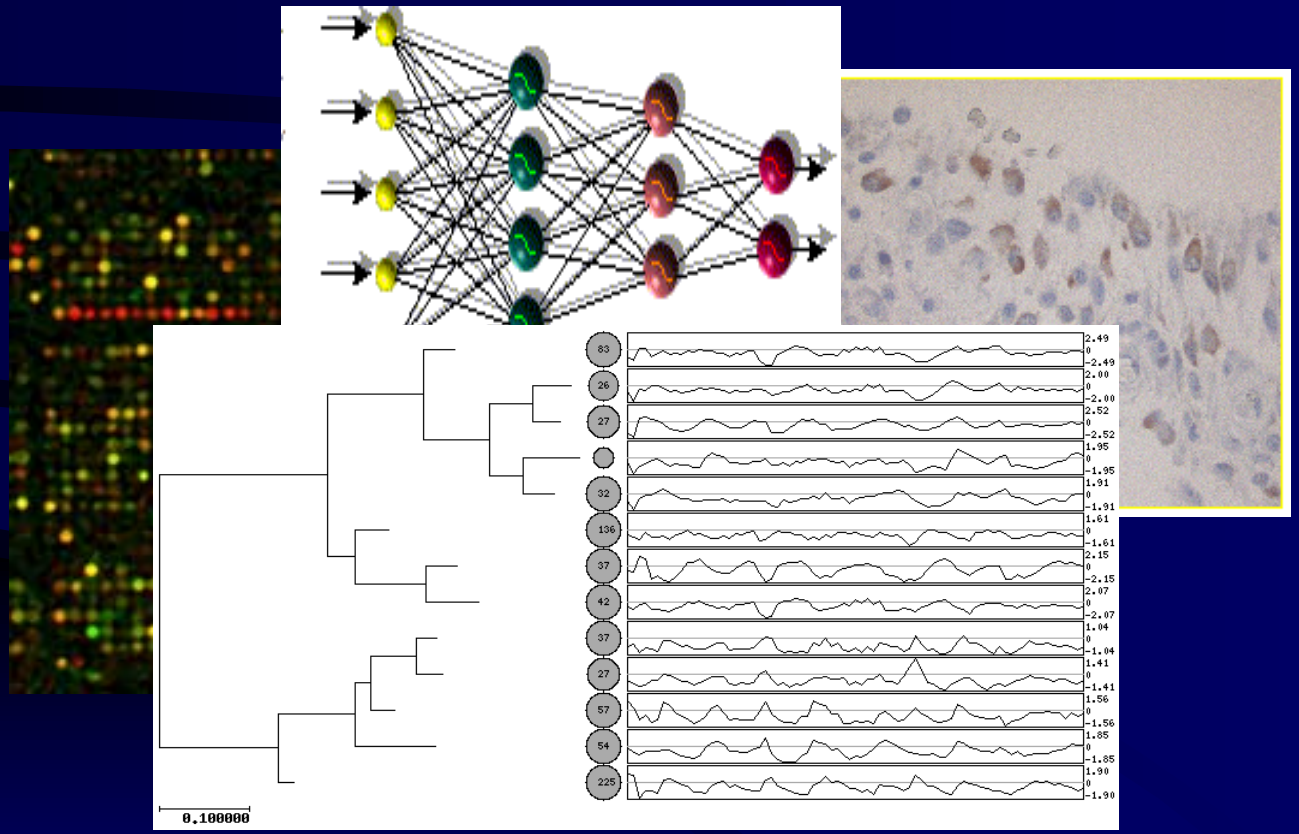
Supervised and hierarchical unsupervised neural networks for clustering both gene expression profiles and samples

A. Mateos, J. Herrero, J. Tamames & J. Dopazo



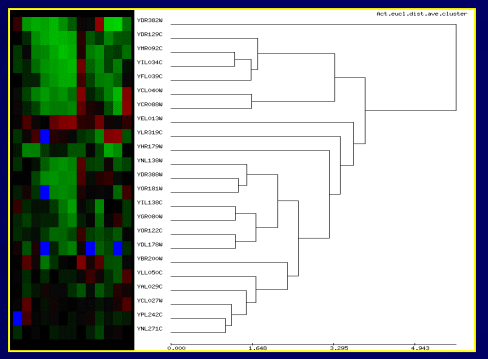
Joaquín Dopazo. Bioinformatics Unit, CNIO.

Clustering methods

Non hierarchical **hierarchical**

K-means, PCA

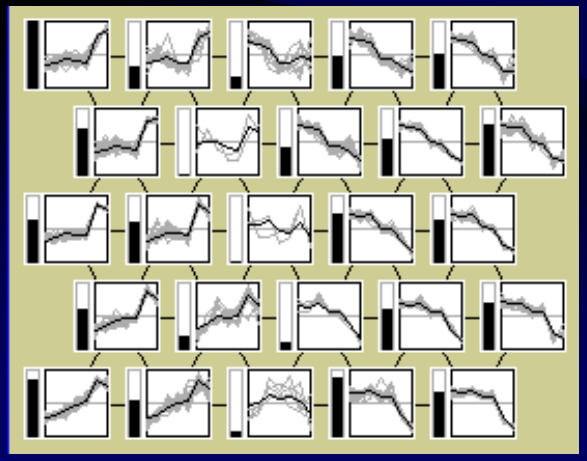
UPGMA



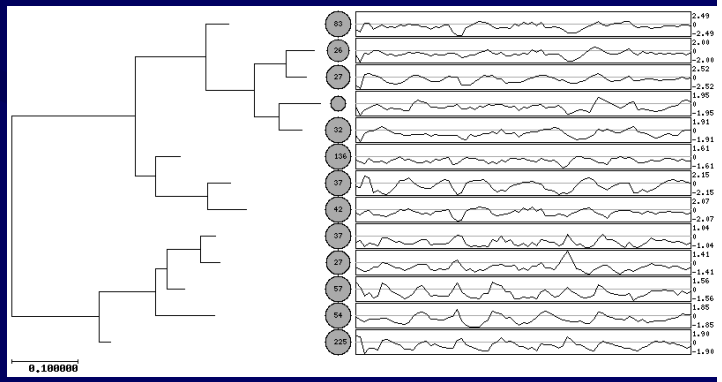
SOM

SOTA

quick and robust

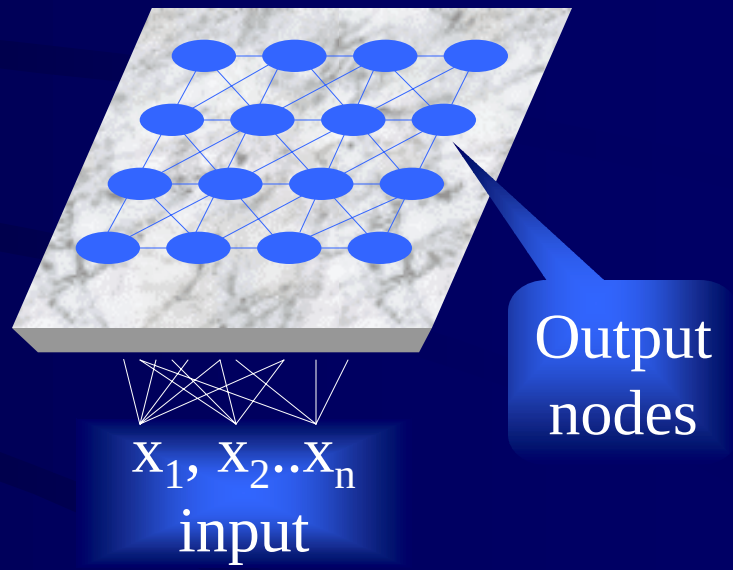


Different levels of information

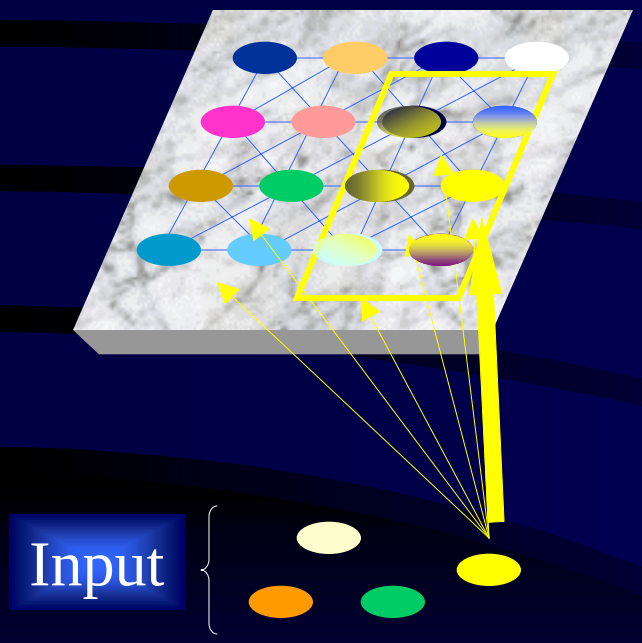


Self Organising Maps: SOM

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:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)); \text{ for } j \in NE_{j^*}(t)$$

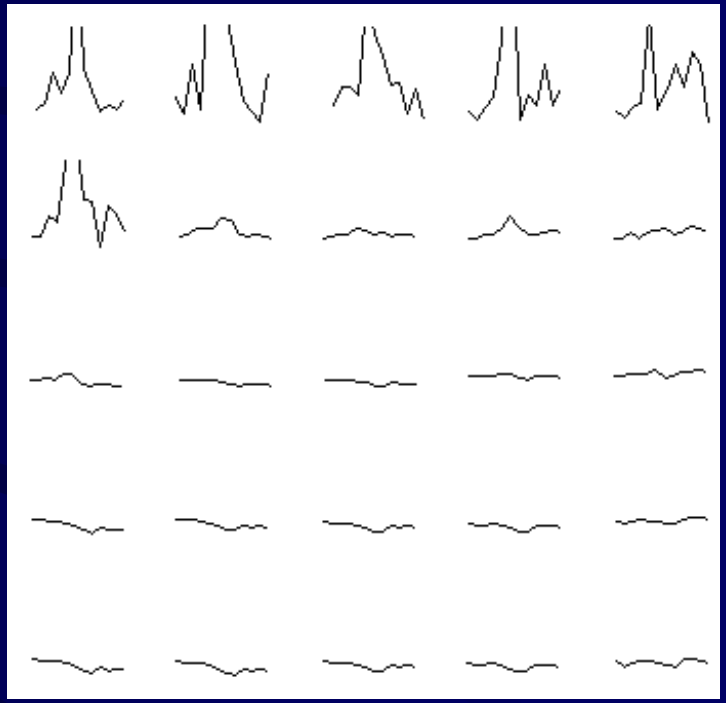
$\eta(t)$ is a gain term that decreases in time.

Step4 Repeat by going to Step 2 until convergence.

SOM:
Example

Response of human fibroblasts to serum

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



SOTA: The algorithm

Step 1. Initialize nodes to random values.

Step 2. Present new input: Compute distances to all **terminal** nodes.

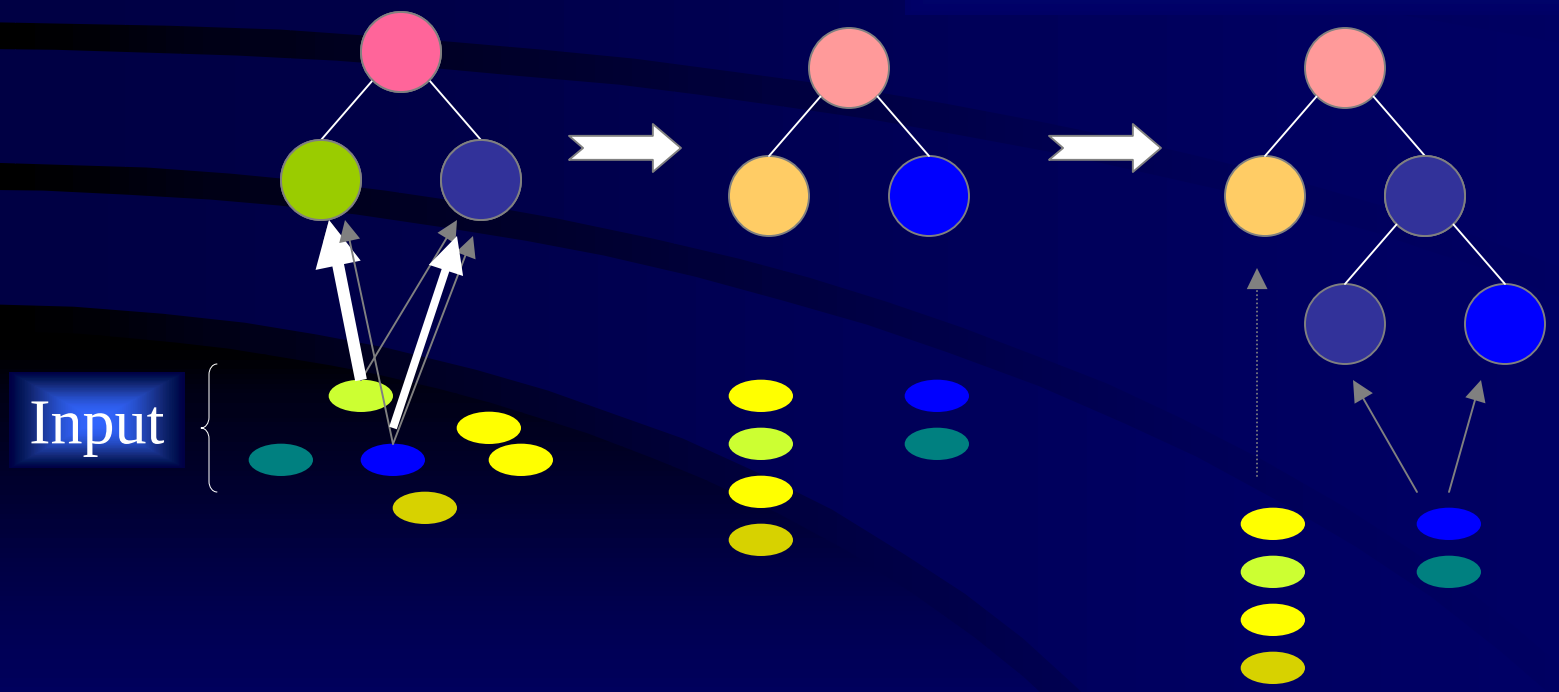
Step 3. Select output node j^* with minimum distance d_j. Update node j^* and neighbors. Nodes updated for the neighborhood $NE_{j^*}(t)$ as:

$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t))$; for $j \in NE_{j^*}(t)$

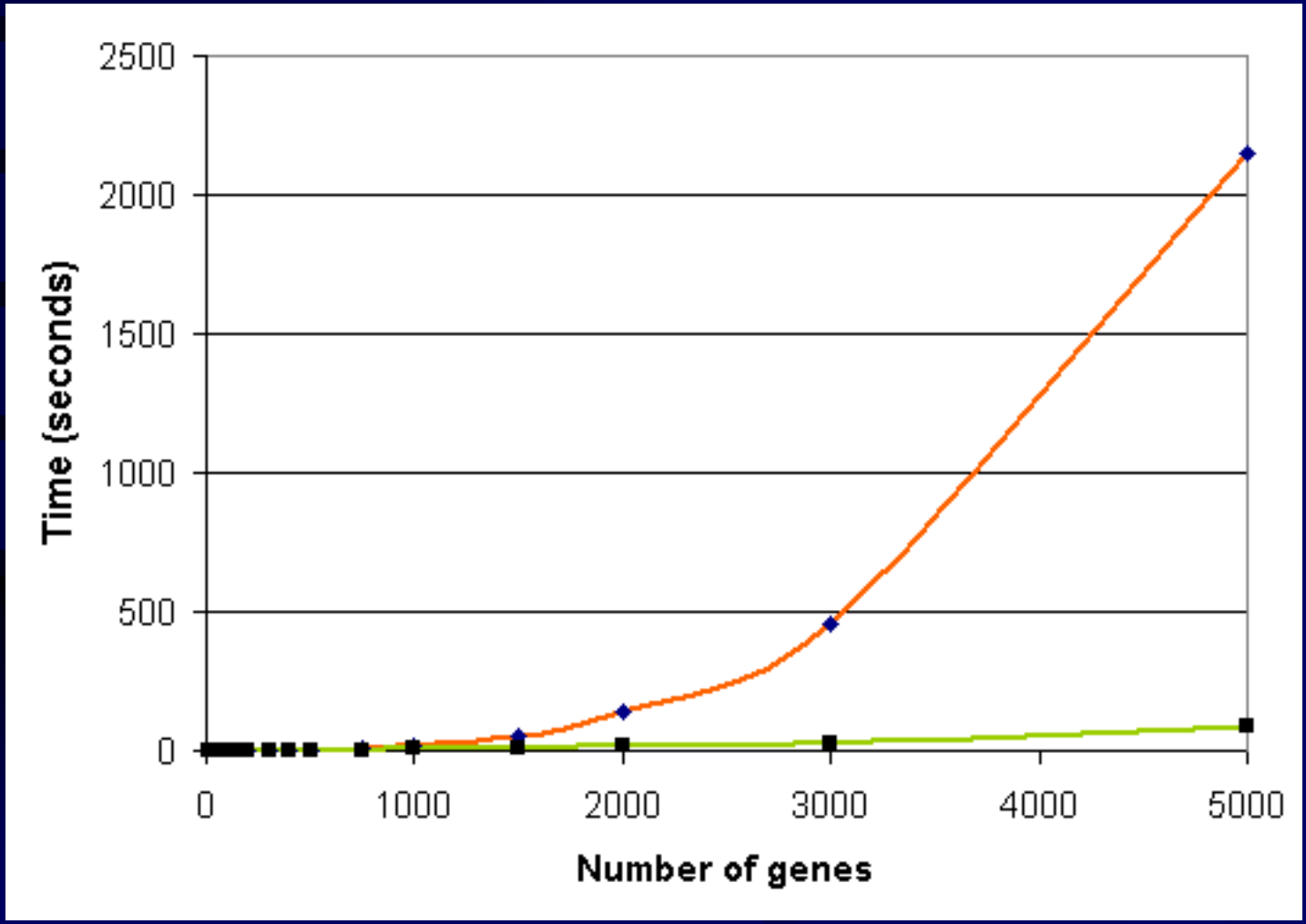
$\eta(t)$ is a gain term that decreases in time.

Step 4 Repeat by going to Step 2 until convergence.

Step 5 Reproduce the node with highest variability.

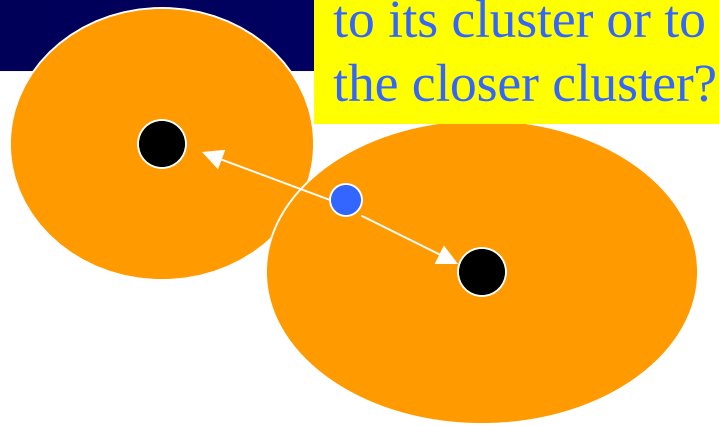


SOTA/SOM vs classical clustering (UPGMA)



Acuracy: the silhouette

Is the object closer to its cluster or to the closer cluster?



$$a(i) = \frac{1}{|A|-1} \sum_{x_i, x_j \in A} d(x_i, x_j) \quad x_i \in A$$

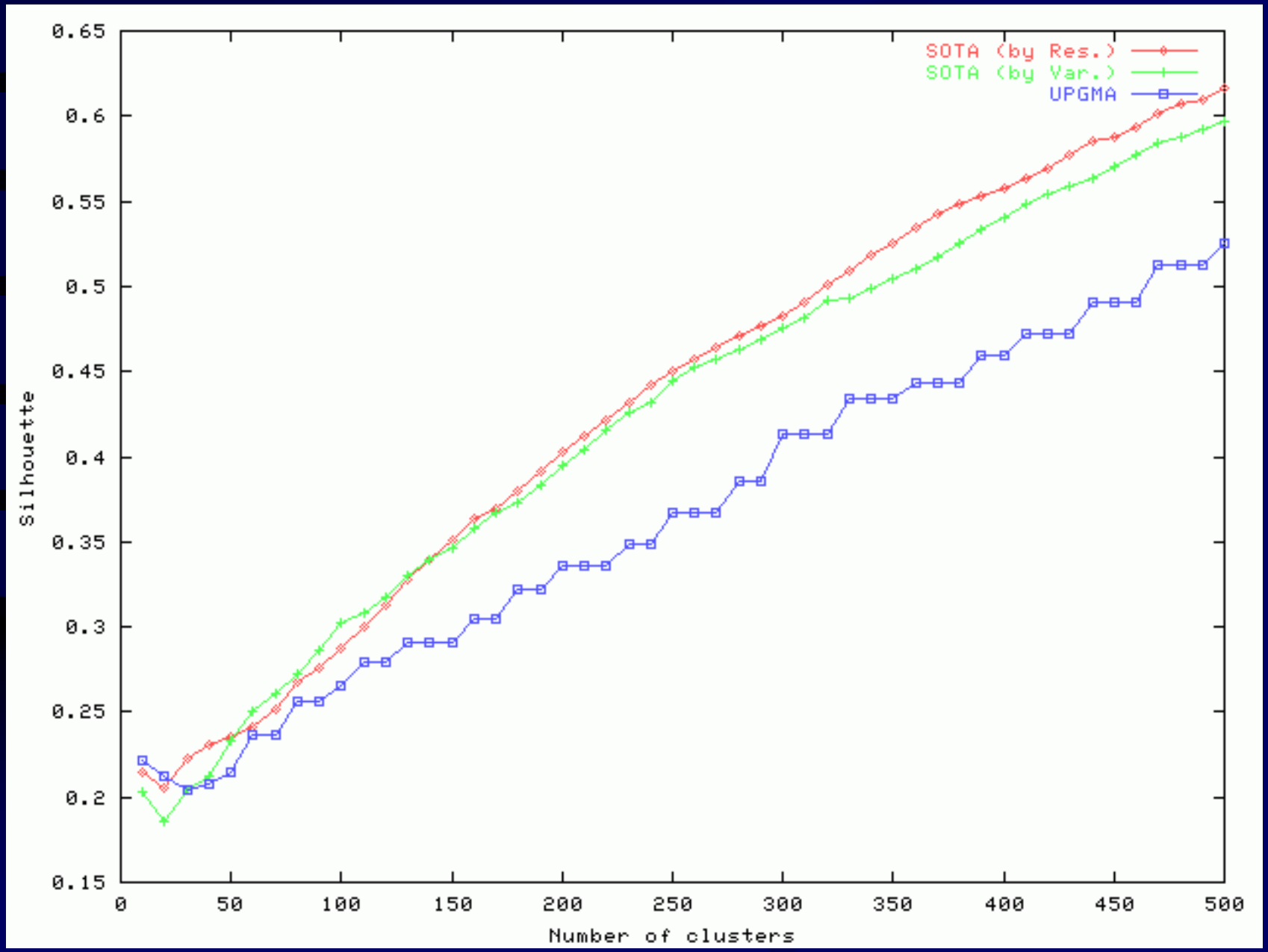
$$d(x_i, C) = \frac{1}{|C|} \sum_{x_j \in C} d(x_i, x_j) \qquad C \neq A$$

$$b(i) = \min_{C \neq A} d(x_i, C)$$

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

$$\left\{ \begin{array}{l} x_i \in A \;\; a(i) << b(i) \Rightarrow \;\; \frac{a(i)}{b(i)} \downarrow \quad s(i) = \frac{b(i)-a(i)}{b(i)} = \left(1 - \frac{a(i)}{b(i)}\right) \rightarrow 1(OK) \\ x_i \in B \;\; a(i) >> b(i) \Rightarrow \;\; \frac{a(i)}{b(i)} \uparrow \quad s(i) = \frac{b(i)-a(i)}{a(i)} = -1 + \frac{b(i)}{a(i)} \rightarrow -1(Wrong) \end{array} \right\}$$

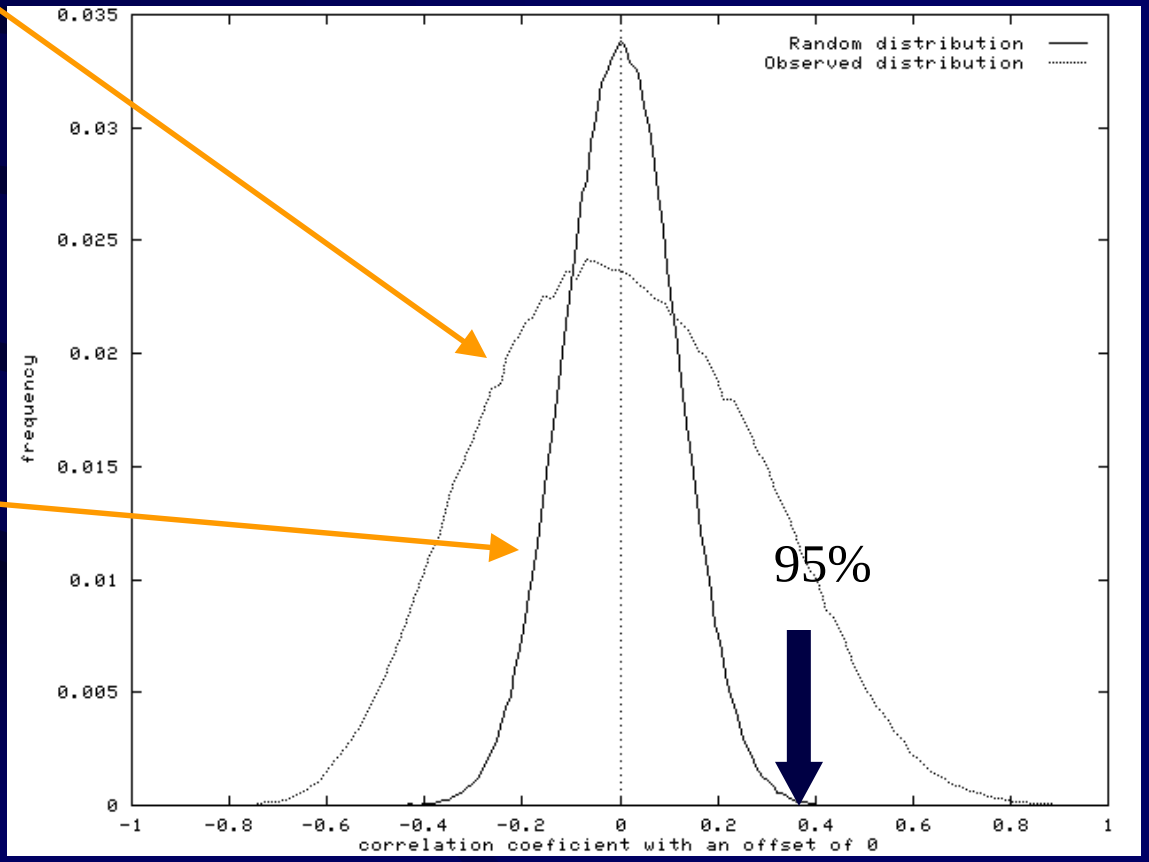
Relative accuracy of SOTA and UPGMA



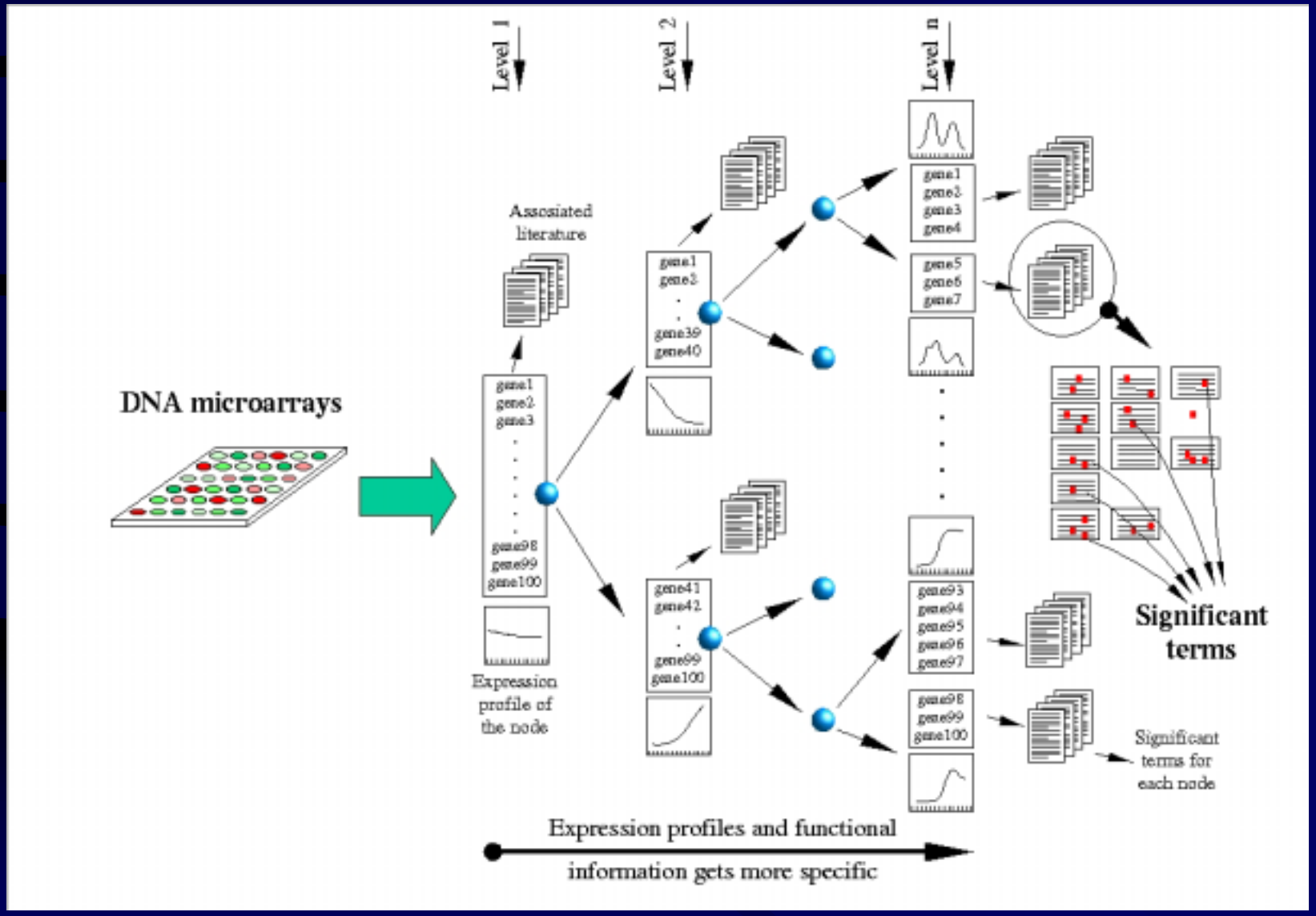
SOTA confidence level with randomisation test

	t_1	t_2	..	t_p
$sample_1$	a_{11}	a_{12}	..	a_{1p}
$sample_2$	a_{21}	a_{22}	..	a_{2p}
:	:	:		:
$sample_n$	a_{n1}	a_{n2}	..	a_{np}

	t_1	t_2	..	t_p
$sample_1$	a_{14}	a_{17}	..	a_{1q}
$sample_2$	a_{23}	a_{21}	..	a_{2r}
:	:	:		:
$sample_n$	a_{n9}	a_{n4}	..	a_{ns}

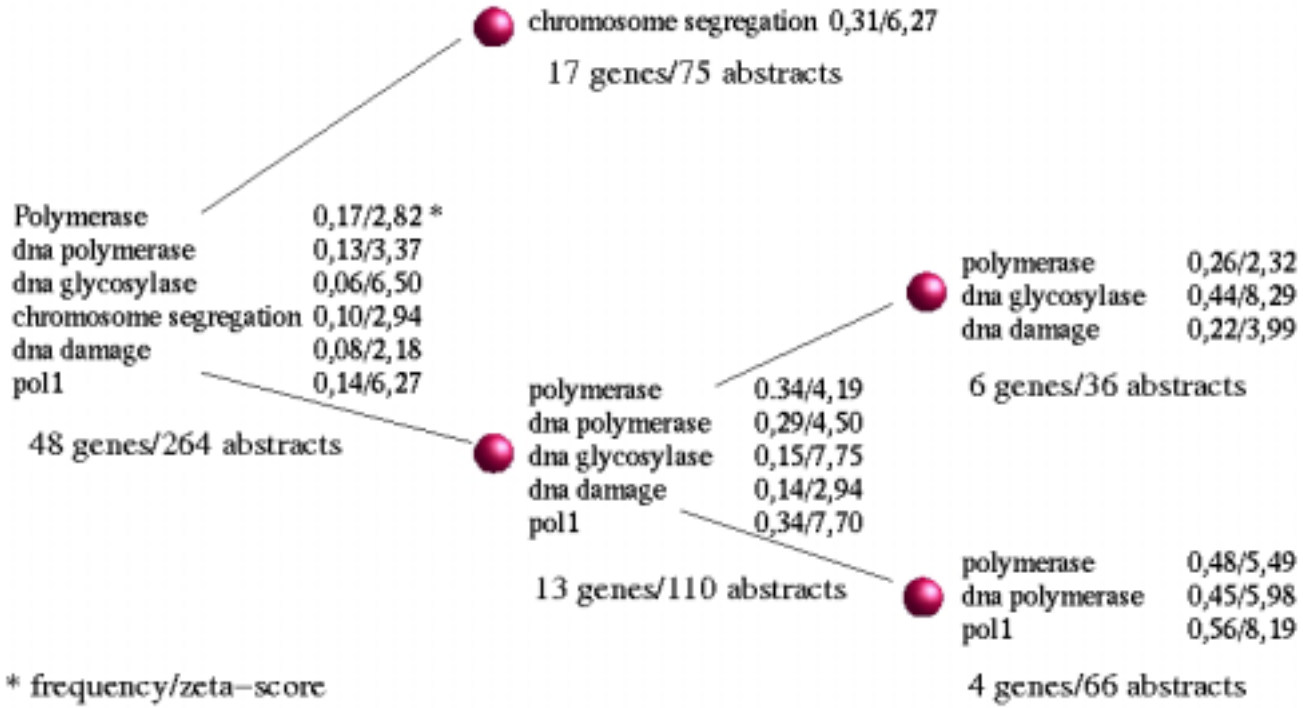


SOTA y almaTM: a simple example.



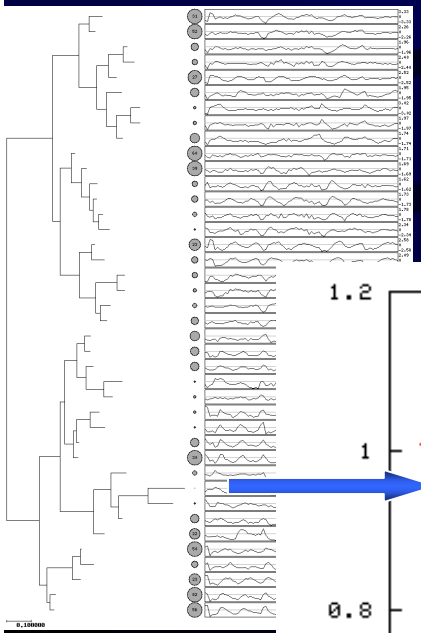
Oliveros, Blaschke, Herrero, Dopazo, Valencia 2000

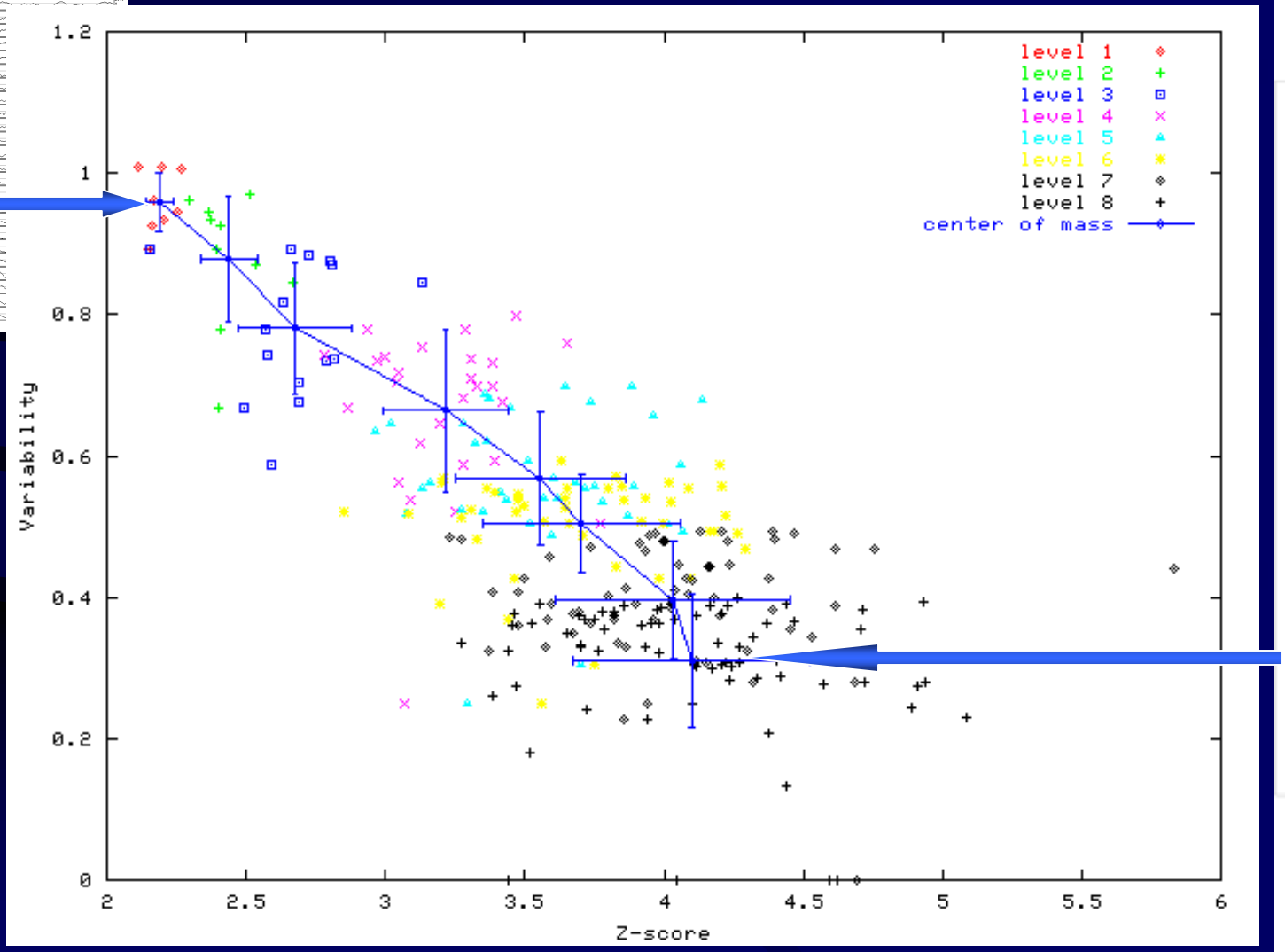
Example of automatic assignation of terms



Oliveros, Blaschke, Herrero, Dopazo, Valencia 2000

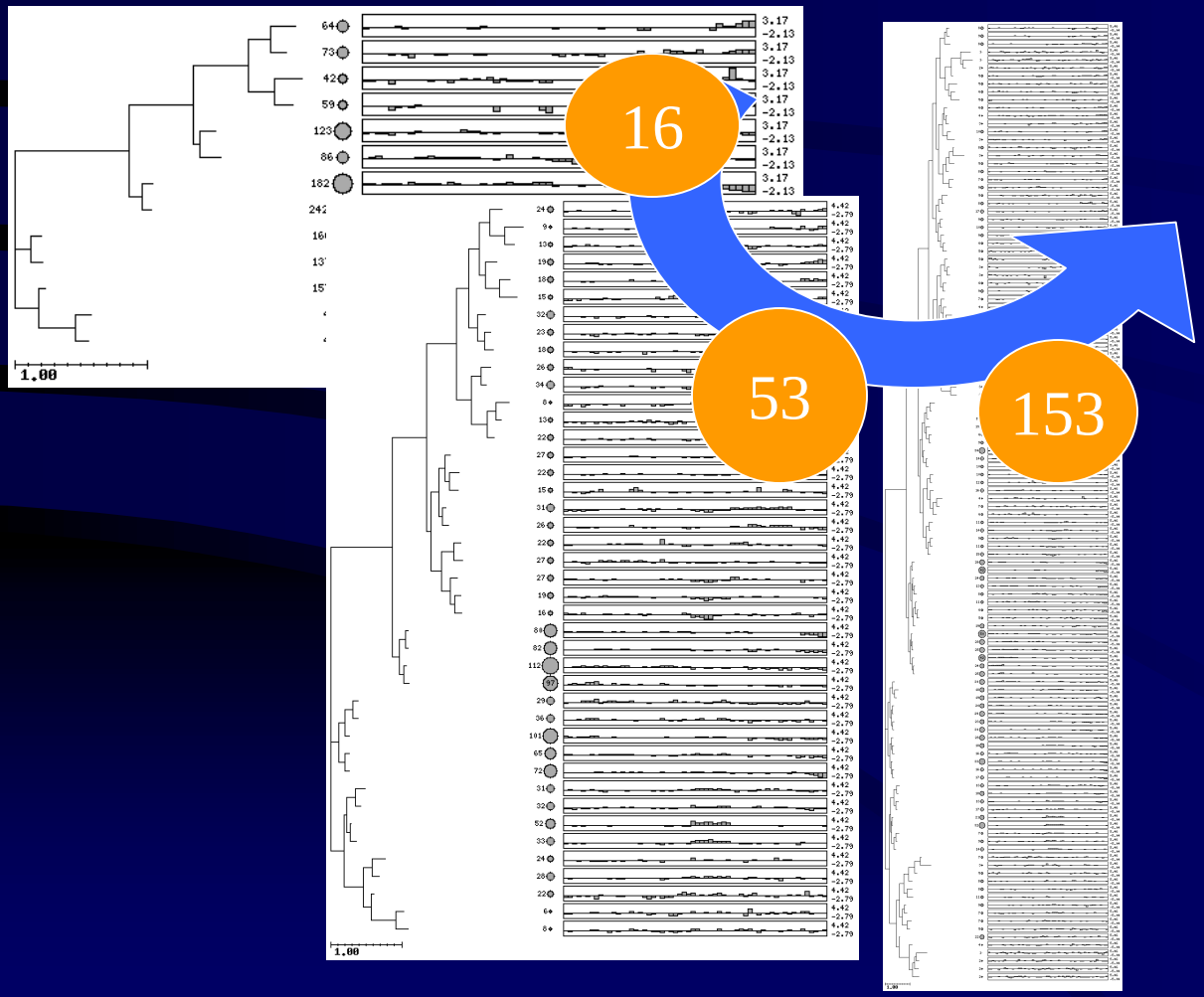
AlmaTM values at different resolution level of a SOTA tree





Oliveros, Blaschke, Herrero, Dopazo, Valencia 2000

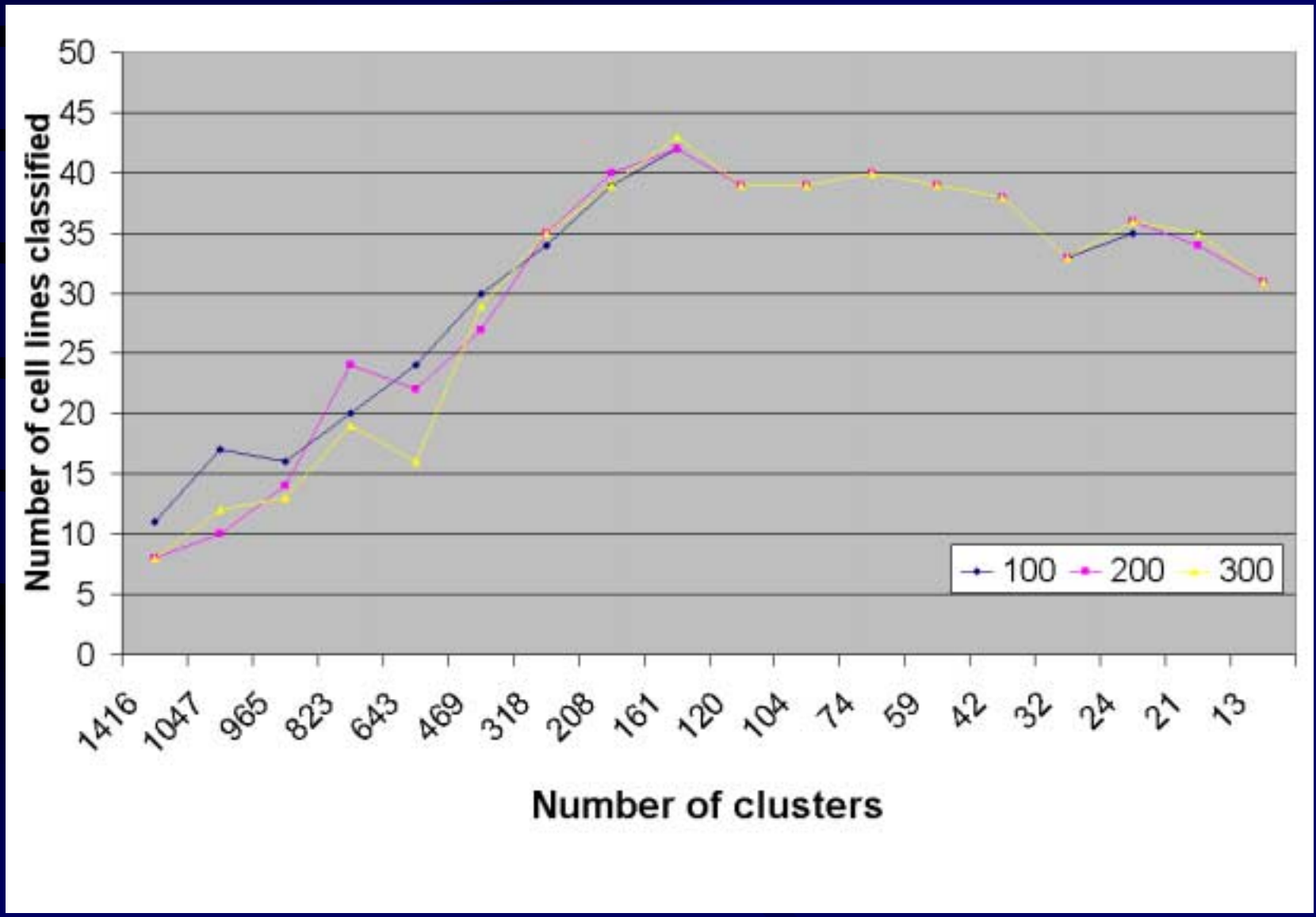
Finding the optimal information level



Finding the optimal relationship between the number of clusters and number of genes in a cluster

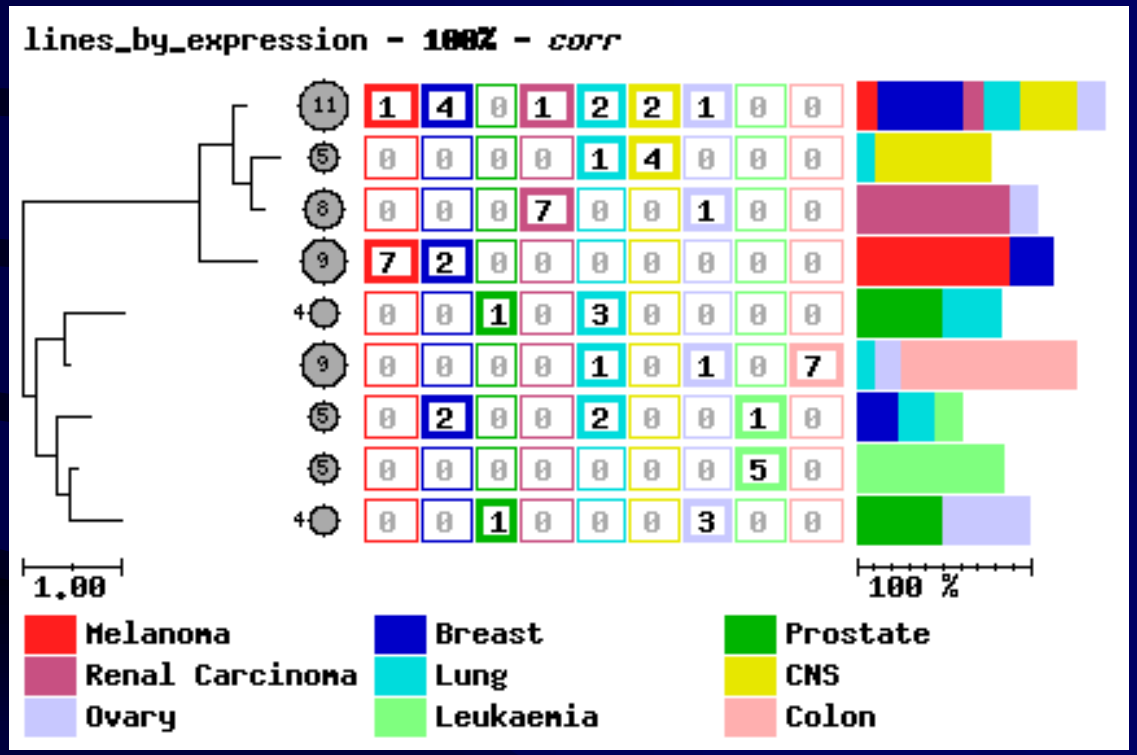
Statistical definition of cluster at 90% gives a tree with 161 clusters

Number of items classified



Efficience in the classification for different levels of information

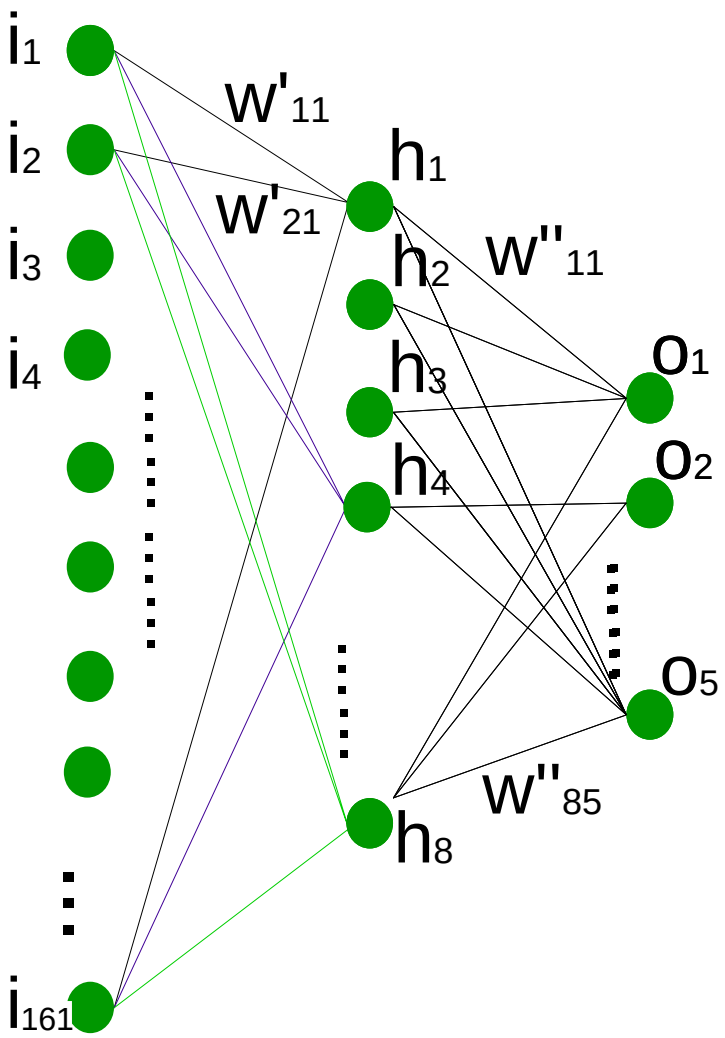
Unsupervised classification of cell lines



Using vectors of 161 average gene expression values as input

Methods: Neural Networks

- Multilayer perceptron.



$$o = \phi(i, w)$$

Training: find w's that minimize errors in example set

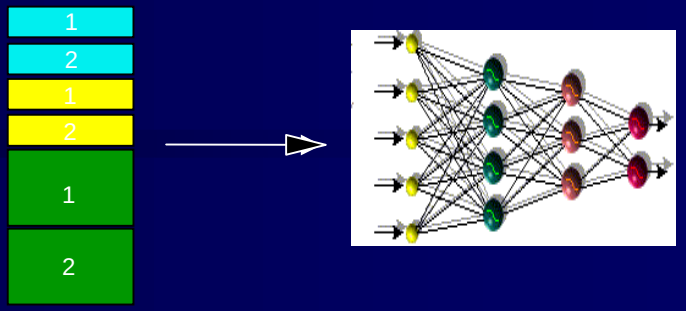
$$|o - \phi(i, w_n)| > |o - \phi(i, w_{n+1})|$$

Generalizing: get a prediction for an input i whose output value is unknown

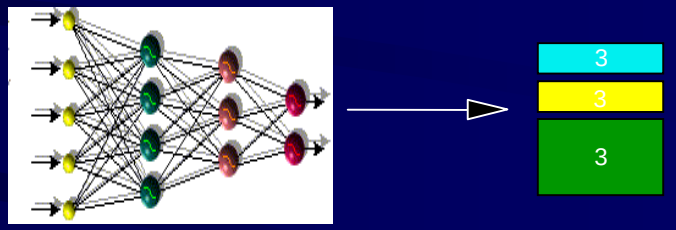
given i, predict: $o = \phi(i, w_{infinity})$

Methods: 3-fold cross validation

Train NN with training set



Predict cell line classes in test set



A pattern in class j will count as a

- ***True Positive:*** if $\phi(i_g, w) > \tau$; τ is a threshold parameter and i_g is the vector of expression levels for class g.
- ***False Negative:*** if $\phi(i_g, w) < \tau$.

A pattern *NOT* inclass j will count as a

- ***True Negative:*** if $\phi(i_g, w) < \tau$.
- ***False Positive:*** if $\phi(i_g, w) > \tau$.

Classification of cancer cell lineages

Cell line	Total	supervised	unsupervised
Breast	8	2	4
Melanoma	8	7	7
Prostate	2	0	0
Renal	8	7	7
Lung	8	5	3
CNS	6	4	4
Ovary	6	5	3
Leukemia	6	6	5
Colon	7	7	7

Conclussions

•SOTA is a clustering method with linear runtime and a superior accuracy than its counterparts UPGMA and SOM.

•SOTA allows obtaining clusters based on a statistical test.

•The biological meaning of these clusters is validated by the fact that their average values are informative enough to be used in the classification of samples.

•SOTA, in combination with a perceptron, is an efficient method for classifying samples based on their gene expression values. Unlike in other approaches the classification thus obtained depends upon groups of co-expressing genes, probably playing a related role in the cell.