

# Exploratory data analysis for microarrays

Adrian Alexa

Computational Biology and Applied Algorithms  
Max Planck Institute for Informatics  
D-66123 Saarbrücken

slides by Jörg Rahnenführer

7<sup>th</sup> BioSapiens European School in Bioinformatics  
27. - 31. August 2007, Basel, Switzerland

## Overview

- **Classification tasks for microarrays**
- **Cluster analysis**
  - Time series example
  - Distance measures
  - Cluster algorithms
- **Comparisons and recommendations**
  - Estimating the number of clusters
  - Assessment of cluster validity
  - Comparative study for tumor classification
  - Gene selection



## Classification tasks for microarrays

- **Classification of SAMPLES**

Generate gene expression profiles that can

  - discriminate between different **known** cell types or conditions, e.g. between tumor and normal tissue,
  - identify different and previously **unknown** cell types or conditions, e.g. new subclasses of an existing class of tumors.
- **Classification of GENES**
  - Assign an unknown cDNA sequence to one of a set of **known** gene classes.
  - Partition a set of genes into new (**unknown**) functional classes on the basis of their expression patterns across a number of samples.

Cancer classification	Class discovery	Class prediction
Machine learning	Unsupervised learning	Supervised learning
Statistics	Cluster analysis	Discriminant analysis



## Classification

### MESSAGE 1

**Discriminant analysis: CLASSES KNOWN**

**Cluster analysis: CLASSES NOT KNOWN**



## Classification

- Difference between **discriminant analysis** (supervised learning) and **cluster analysis** (unsupervised learning) is important:
- If the class labels are **known**, many different **supervised learning** methods are available. They can be used for prediction of the outcome of future objects.
- If the class labels are **unknown**, **unsupervised learning** methods have to be used. For those, it is **difficult to ascertain the validity of inferences** drawn from the output.



## Cluster analysis

### Goal in cluster analysis:

Grouping a collection of objects into subsets or “clusters”, such that those within each cluster are more closely related to one another than objects assigned to different clusters.



## Cluster analysis

### Goal in cluster analysis:

Grouping a collection of objects into subsets or “clusters”, such that those within each cluster are more closely related to one another than objects assigned to different clusters.

### Two ingredients are needed to group objects:

#### Distance measure

A notion of distance or similarity of two objects: **When are two objects close to each other?**

#### Cluster algorithm

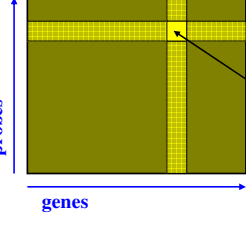
A procedure to minimize distances of objects within groups and/or maximize distances between groups.



## Cluster analysis

- Clustering columns: **grouping similar samples**
- Clustering rows: **grouping genes with similar trajectories**

The gene expression matrix  
probes



$L_{i,j}$ : expression level of gene  $i$  in probe  $j$



## Cluster analysis: Bi-Clustering

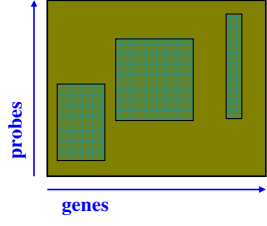
- Clustering columns: **grouping similar samples**
- Clustering rows: **grouping genes with similar trajectories**
- Biclustering: Group genes that have similar partial trajectories in a subset of the samples

### Literature

Tanay, A., Sharan, R., and Shamir, R. (2002): **Discovering Statistically Significant Biclusters in Gene Expression Data**, *Bioinformatics* 18, Suppl.1, 136-144.



The gene expression matrix



## Time series example

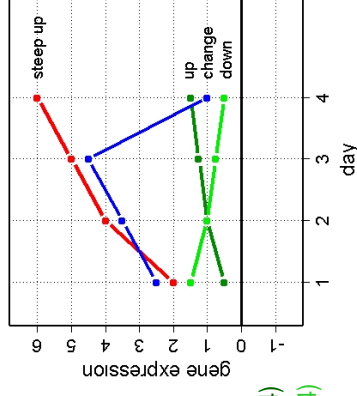
### Biology

Measurements of gene expression on 4 (consecutive) days.

### Statistics

Every gene is coded by a vector of length 4.

- steep up:**  $x_1 = (2, 4, 5, 6)$
- up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down:**  $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change:**  $x_4 = (2.5, 3.5, 4.5, 1)$



## Distance measures - Time series example

### Euclidean distance

The distance between two vectors is the square root of the sum of the squared differences over all coordinates.

$$d_E(x_1, x_2) = \sqrt{(2-2/4)^2 + (4-4/4)^2 + (5-5/4)^2 + (6-6/4)^2} = 3\sqrt{3/4} \approx 2.598$$

- steep up:**  $x_1 = (2, 4, 5, 6)$
- up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$



## Distance measures - Time series example

### Euclidean distance

The distance between two vectors is the square root of the sum of the squared differences over all coordinates.

$$d_E(x_1, x_2) = \sqrt{(2-2/4)^2 + (4-4/4)^2 + (5-5/4)^2 + (6-6/4)^2} = 3\sqrt{3/4} \approx 2.598$$

- steep up:**  $x_1 = (2, 4, 5, 6)$
- up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down:**  $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change:**  $x_4 = (2.5, 3.5, 4.5, 1)$

0	2.60	2.75	2.25
2.60	0	1.23	2.14
2.75	1.23	0	2.15
2.25	2.14	2.15	0

Matrix of pairwise distances



**Distance measures - Time series example**

Manhattan distance

The distance between two vectors is the sum of the absolute (unsquared) differences over all coordinates.

$$d_M(x_1, x_2) = |2-2/4| + |4-4/4| + |5-5/4| + |6-6/4| = 51/4 = 12.75$$

- **steep up:**  $x_1 = (2, 4, 5, 6)$
- **up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$

**Distance measures - Time series example**

Manhattan distance

The distance between two vectors is the sum of the absolute (unsquared) differences over all coordinates.

$$d_M(x_1, x_2) = |2-2/4| + |4-4/4| + |5-5/4| + |6-6/4| = 51/4 = 12.75$$

- **steep up:**  $x_1 = (2, 4, 5, 6)$
- **up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$
- **down:**  $x_3 = (6/4, 4/4, 3/4, 2/4)$
- **change:**  $x_4 = (2.5, 3.5, 4.5, 1)$

0	12.75	13.25	6.50
12.75	0	2.50	8.25
13.25	2.50	0	7.75
6.50	8.25	7.75	0

Matrix of pairwise distances

**Distance measures - Time series example**

Correlation distance

Distance between two vectors is  $1-\rho$ , where  $\rho$  is the Pearson correlation of the two vectors.

$$d_C(x_1, x_2) = 1 - \frac{(2-\frac{2}{4})(\frac{4}{4}-\frac{1}{4}) + (4-\frac{4}{4})(\frac{4}{4}-\frac{1}{4}) + (5-\frac{5}{4})(\frac{4}{4}-\frac{1}{4}) + (6-\frac{6}{4})(\frac{4}{4}-\frac{1}{4})}{\sqrt{(2-\frac{2}{4})^2 + (4-\frac{4}{4})^2 + (5-\frac{5}{4})^2 + (6-\frac{6}{4})^2} \sqrt{(\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2}}$$

- **steep up:**  $x_1 = (2, 4, 5, 6)$
- **up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$

**Distance measures - Time series example**

Correlation distance

Distance between two vectors is  $1-\rho$ , where  $\rho$  is the Pearson correlation of the two vectors.

$$d_C(x_1, x_2) = 1 - \frac{(2-\frac{2}{4})(\frac{4}{4}-\frac{1}{4}) + (4-\frac{4}{4})(\frac{4}{4}-\frac{1}{4}) + (5-\frac{5}{4})(\frac{4}{4}-\frac{1}{4}) + (6-\frac{6}{4})(\frac{4}{4}-\frac{1}{4})}{\sqrt{(2-\frac{2}{4})^2 + (4-\frac{4}{4})^2 + (5-\frac{5}{4})^2 + (6-\frac{6}{4})^2} \sqrt{(\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2 + (\frac{4}{4}-\frac{1}{4})^2}}$$

- **steep up:**  $x_1 = (2, 4, 5, 6)$
- **up:**  $x_2 = (2/4, 4/4, 5/4, 6/4)$
- **down:**  $x_3 = (6/4, 4/4, 3/4, 2/4)$
- **change:**  $x_4 = (2.5, 3.5, 4.5, 1)$

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Matrix of pairwise distances

## Distance measures - Time series example

### Summary

- **Euclidean** distance measures average difference across coordinates.
- **Manhattan** distance measures average difference across coordinates, in a robust way.
- **Correlation** distance measures difference with respect to trends.



## Distance measures - standardization

### Standardization

- Data points are normalized with respect to mean and variance:  
Apply transformation  $x \mapsto \frac{x - \hat{\mu}}{\hat{\sigma}}$ , where  $\hat{\mu}$  is an estimator of the mean (usually average across coordinates) and  $\hat{\sigma}$  is an estimator of the variation (usually empirical standard deviation).
- After standardization, Euclidean distance and Correlation distance are equivalent(!):  $d_E(x_1, x_2)^2 = 2nd_C(x_1, x_2)$
- Standardization makes sense, if you are not interested in the magnitude of the effects, but in the effect itself. Results can be misleading for noisy data.



## Distance measures

### MESSAGE 2

**Appropriate choice of distance measure depends on your intention!**



## Cluster algorithms

There are three different types of cluster algorithms:

**combinatorial algorithms**, *mixture modeling* and *mode seeking*

Most popular cluster algorithms:

- Hierarchical clustering
  - K-means
  - PAM (Partitioning around medoids)
  - SOM's (Self-Organizing Maps)
- K-means and SOM's take original data directly as input: the attributes are assume to live into an Euclidean space.
  - Hierarchical cluster algorithms and PAM allow the choice of a dissimilarity matrix  $\mathbf{d}$ , that assigns to each pair of objects  $x_i$  and  $x_j$  a value  $d(x_i, x_j)$  as their distance.



## Hierarchical cluster algorithms

- **Hierarchical clustering** was the first algorithm used in microarray research to cluster genes (Eisen et al. (1998)).
1. First, each object is assigned to its own cluster.
  2. **Iteratively:**
    - **the two most similar clusters are joined**, representing a new node of the clustering tree. The node is computed as **average of all objects of the joined clusters**.
    - the similarity matrix is updated with this new node replacing the two joined clusters.
  3. Step 2 is repeated until only one single cluster remains.



Adrian Alexa

Basel, August 31, 2007

## Hierarchical cluster algorithms

- **Calculation** of distance  $d(G, H)$  between two clusters **G** and **H** is based on object dissimilarity between the objects from the two clusters:
  - Single linkage uses the **smallest distance**:  $d_S(G, H) = \min_{i \in G, j \in H} d_{ij}$
  - Complete linkage uses the **largest distance**:  $d_C(G, H) = \max_{i \in G, j \in H} d_{ij}$
  - Average linkage uses the **average distance**:  $d_A(G, H) = \frac{1}{N_G N_H} \sum_{i \in G, j \in H} d_{ij}$
- Instead of agglomerative clustering, sometimes **divisive clustering** is used: **Iteratively, best possible splits are calculated.**



Adrian Alexa

Basel, August 31, 2007

## Hierarchical cluster algorithms

- **Visualization** of hierarchical clustering through **dendrogram**:
  - Clusters that are joined are combined by a line.
  - Height of line is **average** distance between clusters.
  - Cluster with smaller variation is plotted on left side.
- The procedure provides a **hierarchy of clusterings**, with the number of clusters ranging from 1 to the number of objects.
- **BUT:**
  - Parameters for distance matrix:  **$n(n-1)/2$**
  - Parameters for dendrogram:  **$n-1$** .
  - **Hierarchical clustering does not show the full picture!**

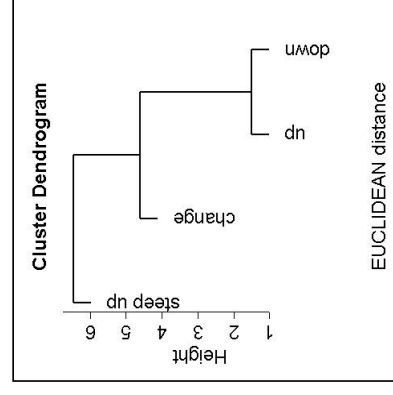
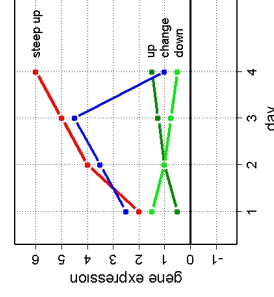


Adrian Alexa

Basel, August 31, 2007

## Time series example

- Euclidean distance  
Similar values are clustered together

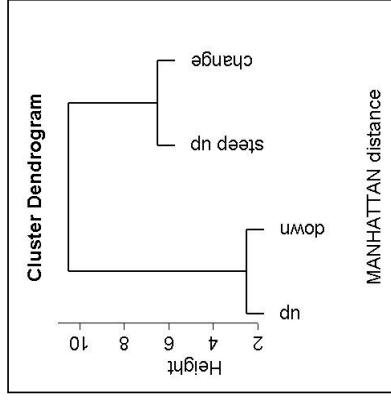
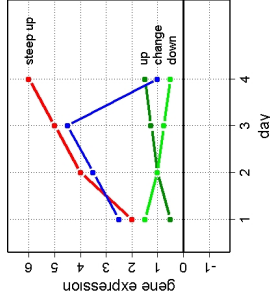


Adrian Alexa

Basel, August 31, 2007

## Time series example

- Manhattan distance  
Similar values are clustered together (robust)

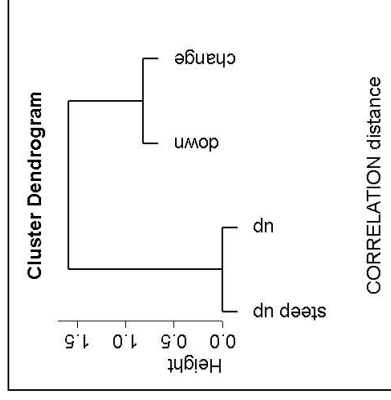
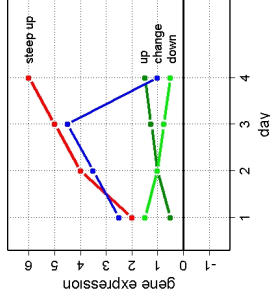


Adrian Alexa

Basel, August 31, 2007

## Time series example

- Correlation distance  
Similar trends are clustered together



Adrian Alexa

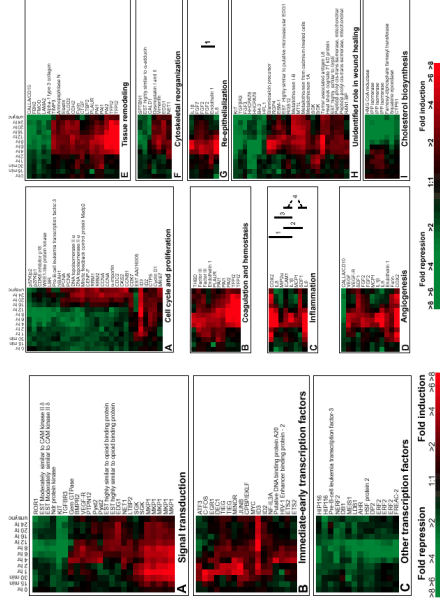
Basel, August 31, 2007

## Clustering time series data – literature examples

Iyer et al., Science, Jan 1999:

Genes from functional classes are clustered together (sometimes!).

Careful interpretation necessary!



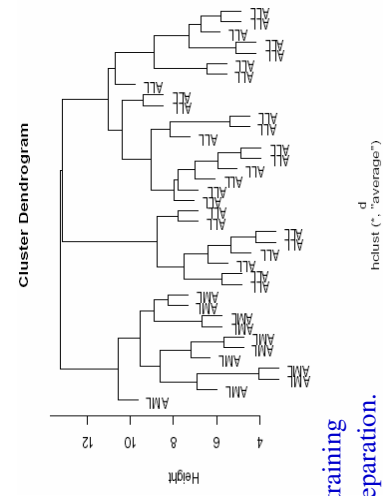
Adrian Alexa

Basel, August 31, 2007

## Clustering time series data – literature examples

Golub et al.: Leukemia dataset, <http://www.genome.wi.mit.edu/MPR>

3 cancer classes:  
25 acute myeloid leukemia (AML),  
47 acute lymphoblastic leukemia (ALL), the latter 9 T-cell and 38 B-cell.



Dendrogram for 38 training data shows perfect separation.

Adrian Alexa

Basel, August 31, 2007

## Cluster algorithms – k-means

- **K-means** is a **partitioning algorithm** with a prefixed number **k** of clusters. It tries to minimize the sum of **within-cluster variances**.
- The a random algorithm chooses sample of **k** different objects as initial cluster midpoints. Then it alternates between two steps until convergence:
  1. Assign each object to its closest of the **k** midpoints with respect to **Euclidean distance**.
  2. Calculate **k new midpoints** as the averages of all points assigned to the old midpoints, respectively.
- **K-means** is a randomized algorithm, two runs usually produce different results. Thus, it has to be applied a few times to the same data set and the result with **minimal sum of within-cluster variances** should be chosen.

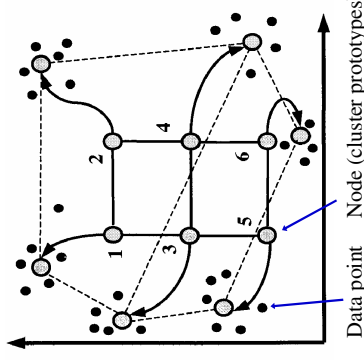


Adrian Alexa

Basel, August 31, 2007

## Cluster algorithms – Self-Organizing maps

- SOM's are similar to **K-means**, but with additional **constraints**.
- Mapping from input space onto one or two-dimensional array of **k** total nodes.
- Iteration steps (20000-50000):
  - Pick data point **P** at random
  - Move **all nodes** in direction of **P**: the closest node the most, the further a node is in network topology, the less
  - Decrease amount of **movement** with iteration steps



Tamayo et al. (1999): First use of SOM's for gene clustering from microarrays



Adrian Alexa

Basel, August 31, 2007

## Cluster algorithms - PAM

- **PAM** (Partitioning around medoids, Kaufman and Rousseeuw (1990)) is a partitioning algorithm, a generalization of **K-means**.
- For an arbitrary dissimilarity matrix **d** it tries to minimize the sum (over all objects) of distances to the closest of **k** prototypes.
- Objective function: 
$$\sum_{i=1}^n \min_{j=1, \dots, k} d(i, m_j)$$
 (**d**: Manhattan, Correlation, etc.)
- **Build phase**: Initial "medoids".
- **Swap phase**: Repeat until convergence:
  - Consider all pairs of objects (i,j), where i is a medoid and j not, and make the  $i \leftrightarrow j$  swap (if any) which decreases the objective function most.



Adrian Alexa

Basel, August 31, 2007

## Comparative study

- **Comparative study for tumor classification with microarrays**: Comparison of hierarchical clustering, K-means, PAM and SOM's
- **Data sets**:
  - Golub et al: Leukemia dataset, <http://www.genome.wi.mit.edu/MPR>, 3 cancer classes: 25 acute myeloid leukemia (AML) and 47 acute lymphoblastic leukemia (ALL) (9 T-cell and 38 B-cell), Affymetrix.
  - Ross et al.: NCI60 cancer dataset, <http://genome-www.stanford.edu/nci60>, 9 cancer classes: 9 breast, 6 central nervous system, 7 colon, 8 leukemia, 8 melanoma, 9 lung, 6 ovarian, 2 prostate, 8 renal, cDNA microarray
- Rahnenführer (2002): **Efficient clustering methods for tumor classification with gene expression arrays**, *Proc. of 26th Ann. Conf. of the Gesellschaft für Klassifikation*, Mannheim, July 2002.



Adrian Alexa

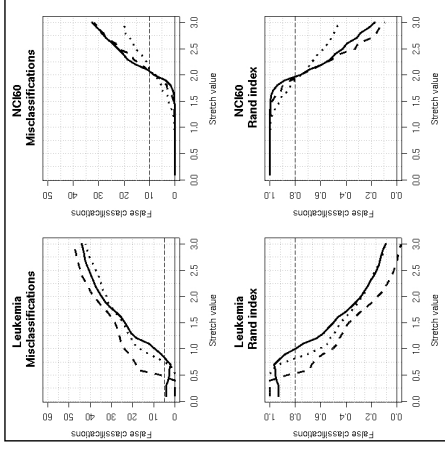
Basel, August 31, 2007



## Comparative study - winners

### Winners

- K-means
- - Hierarchical, Correlation
- ..... PAM, Manhattan



Adrian Alexa

Basel,

August 31, 2007



## Comparative study - results

- **Superiority of k-means with repeated runs**  
Similar for discriminant analysis: FLDA best (Dudoit et al., 2001)
- **Superiority of PAM with Manhattan distance** for noisy data
- Differences depend on the specific dataset
- **Preselection of genes**

Various approaches have been proposed for gene selection, especially in *supervised* learning.

For clustering samples, a practical proceeding is to choose the **top 100-200 genes with respect to variance (across samples)**. This decreases noise and computation time.

Adrian Alexa

Basel,

August 31, 2007



## Classification

### MESSAGE 3

Simple cluster algorithms work better  
in case of little model knowledge!

**(But: More sophisticated methods might be more appropriate with more a priori knowledge)**

**Combinatorial cluster algorithms are approximation algorithms, they converge to local optima!**

Adrian Alexa

Basel,

August 31, 2007



## Recommendations


- **Interest in specific genes:**  
If you search for genes that are co-regulated with a specific gene of your choice, **DO SO!** Don't use clustering, but generate a list of genes close to your gene with respect to some distance.
- **Clustering after feature selection?**  
**NO!** Do not first select genes based on the outcome of some covariable (e.g. tumor type) for your clustering. You will **ALWAYS** find differences w.r.t. this covariable, since this is how you selected the genes!
- **Number of clusters**  
No general rule how to select the '**correct**' number of clusters. Adhoc approach is to **try different numbers** and choose cutoff, for which performance of the clustering algorithm breaks down.  
**The quality of a clustering result depends on the concept of a cluster!**

Adrian Alexa

Basel,

August 31, 2007





  
mpi

---

## R commands and libraries

- **library(mva)**
  - Hierarchical clustering: *hclust()*
  - Kmeans: *kmeans()*
  - Principal components: *princomp()*
- **library(cluster)**
  - PAM: *pam()*
  - Silhouette information: *silhouette()*
- **ISIS package:** <http://www.molgen.mpg.de/~heydebre>

Adrian Alexa      Basel,      August 31, 2007


  
mpi

---

## SUMMARY

MESSAGE 1:

**Discriminant analysis: CLASSES KNOWN**  
**Cluster analysis: CLASSES NOT KNOWN**


MESSAGE 2:

**Appropriate choice of distance measure depends on your intention!**

MESSAGE 3:

**Simple cluster algorithms work better in case of little model knowledge!**

Adrian Alexa      Basel,      August 31, 2007



  
mpi

---

## Literature

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531-37.
2. Alizadeh AA, Eisen MB, Davis RE and 28 others. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; 403: 503-11.
3. Jain A, Dubes RC. Algorithms for Clustering Data. Englewood Cliffs, New Jersey: Prentice Hall; 1988.
4. Azaaje F. Clustering-based approaches to discovering and visualising microarray data patterns. *Brief. Bioinformatics* 2003; 4: 31-42.
5. Eisen MB, Spellman PT, Brown PO, Bostein D. Cluster analysis and display of genome-wide expression patterns. *PNAS* 1998; 95: 14863-68.
6. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitarcewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS* 1999; 96: 2907-12.
7. Kaufman L, Rousseeuw P. Finding Groups in Data. New York: John Wiley and Sons; 1990.
8. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J. Comput Biol.* 1999; 6: 281-97.

Adrian Alexa      Basel,      August 31, 2007


  
mpi

---

## Literature

9. Cheng Y, Church GM. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol.* 2000; 8:93-103.
10. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 2002; Suppl 1: 136-44.
11. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Bostein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000; 1(2): RESEARCH0003.
12. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics* 2001; 17: 309-18.
13. Rahnenfuhrer J. Efficient clustering methods for tumor classification with microarrays. In: *Between Data Science and Applied Data Analysis* (Eds: M. Schader, W. Gaul, M. Vichi), Springer, Proc. 26th Ann. Conf. GfKI 2002; 670-679.
14. Dudoit S, Fridlyand J. A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology* 2002; 3:RESEARCH0036.
15. Smolkin M, Ghosh, D. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* 2003; 4:36.

Adrian Alexa      Basel,      August 31, 2007