Microarray Data Analysis

Clustering and Visualization

國立台灣大學資訊所

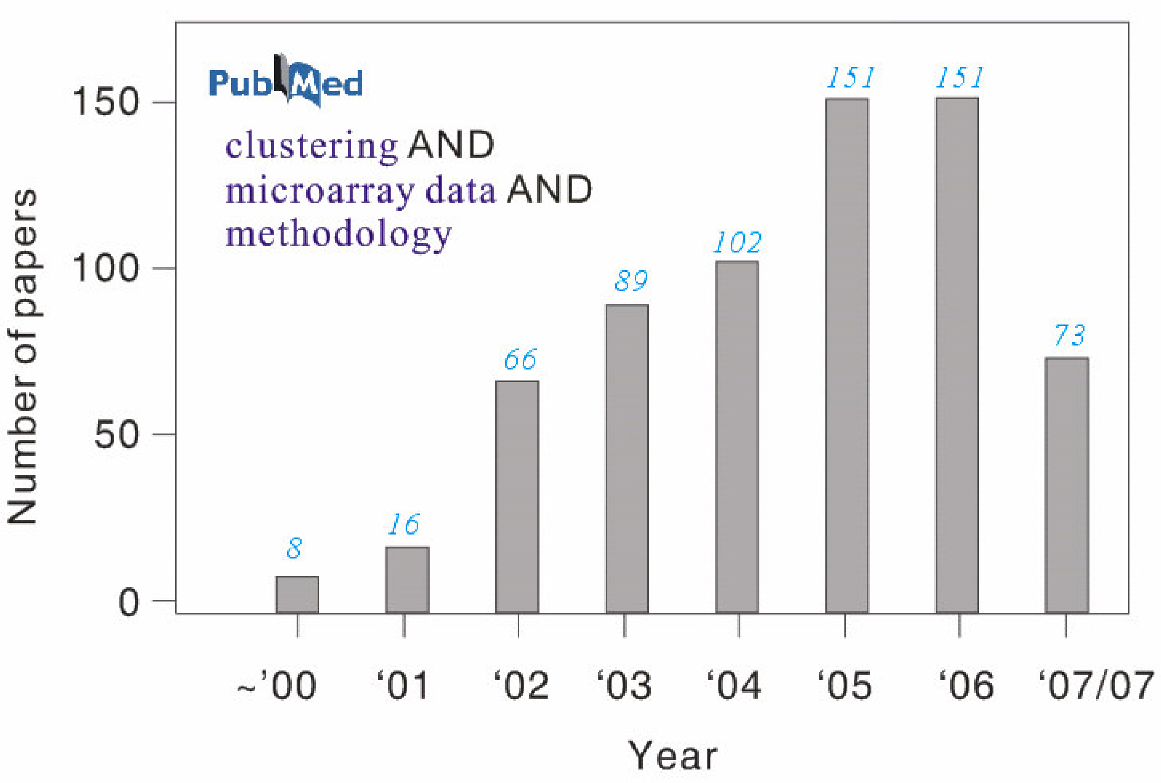
Course: 生物資訊與計算分子生物學

2007/11/06

吳漢銘
hmwu@stat.sinica.edu.tw
http://idv.sinica.edu.tw/hmwu

中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

Clustering of Microarray Gene Expression Data

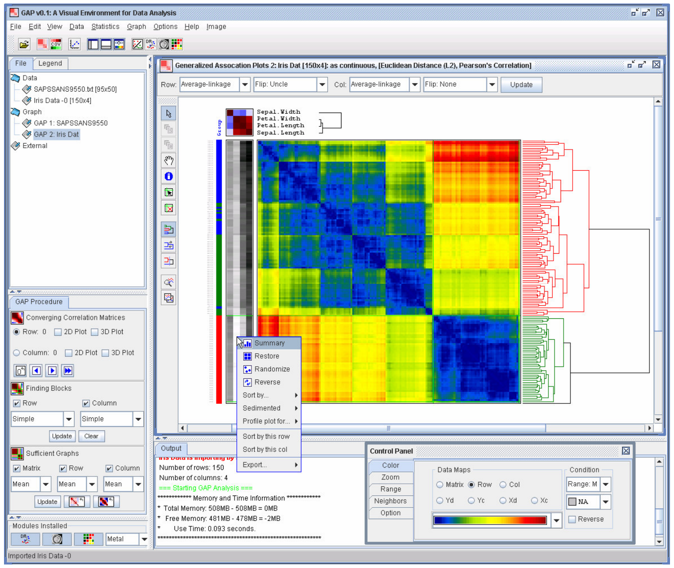


A continuing and active topic of research and application!

Outlines

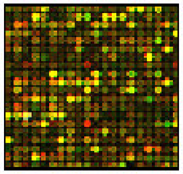
- Overview of Microarray Experiment
- Clustering Analysis and Visualization
- Distance and Similarity Measure
- K-Means Clustering
- Visualizing Clustering Results: Dimension Reduction Techniques
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
- Clustering Analysis and Visualization
  - Self-Organizing Maps (SOM)
  - Heat Map
  - Hierarchical Clustering
  - Tree Flip
- Cluster Validation
- Software

GAP

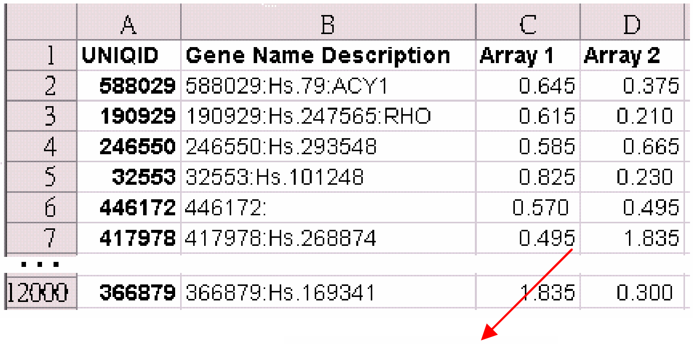


```
Iris Data is importing by ...
Number of rows: 150
Number of columns: 4
=== Starting GAP Analysis ===
************ Memory and Time Information ************
*  Total Memory: 508MB - 508MB = 0MB
*   Free Memory: 481MB - 478MB = -2MB
*       Use Time: 0.093 seconds.
*****************************************************
```

Overview of Microarray Experiment

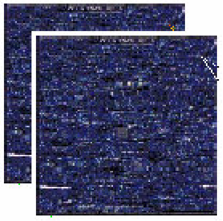


cDNA Microarray Data

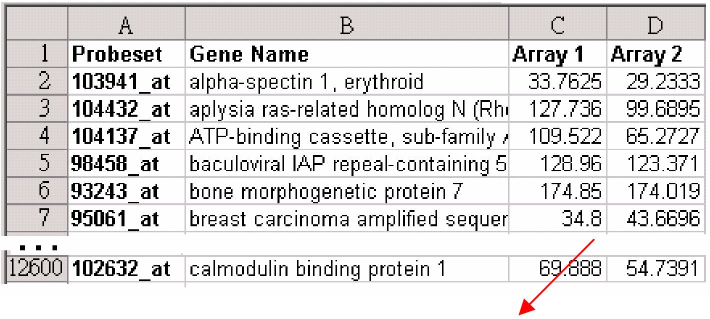


	A	B	C	D
1	**UNIQID**	**Gene Name Description**	**Array 1**	**Array 2**
2	**588029**	588029:Hs.79:ACY1	0.645	0.375
3	**190929**	190929:Hs.247565:RHO	0.615	0.210
4	**246550**	246550:Hs.293548	0.585	0.665
5	**32553**	32553:Hs.101248	0.825	0.230
6	**446172**	446172:	0.570	0.495
7	**417978**	417978:Hs.268874	0.495	1.835
…				
12000	**366879**	366879:Hs.169341	1.835	0.300

log2(Cy5/Cy3)



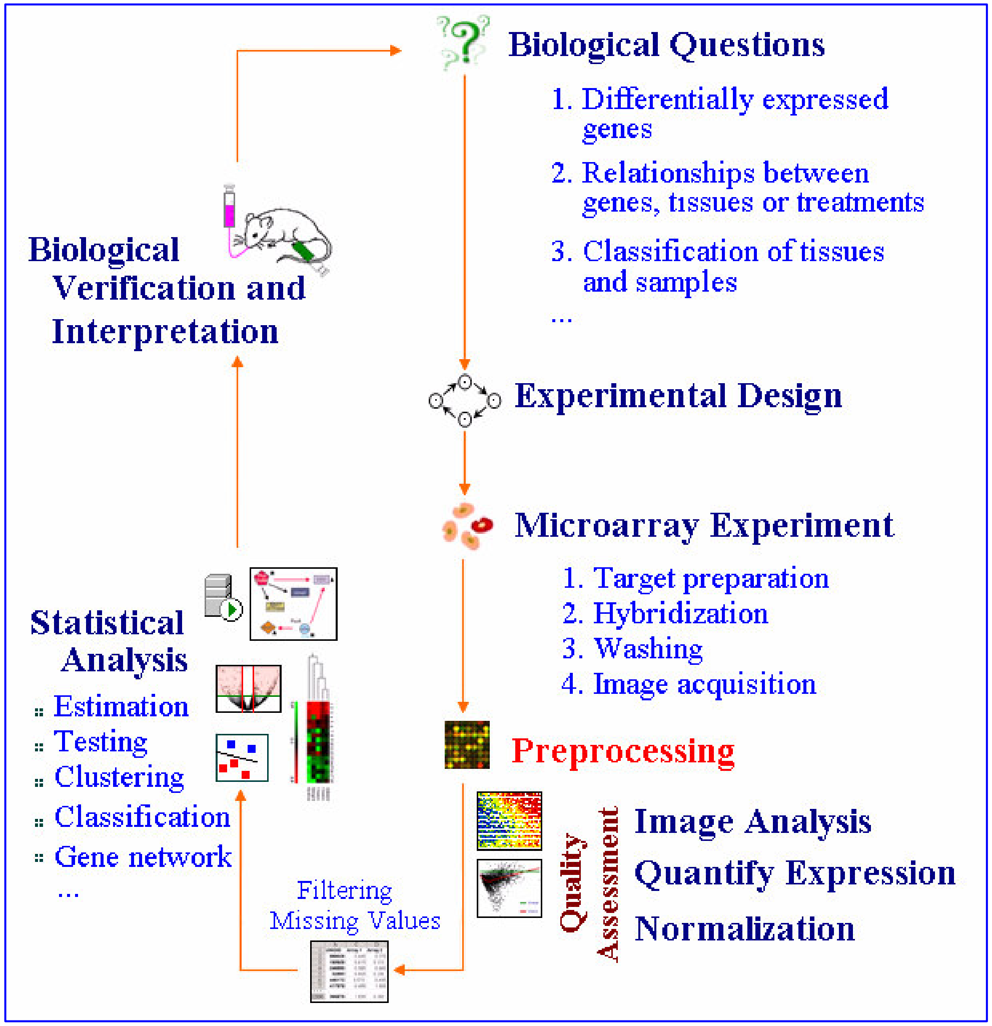
Oligonucleotide Array Data



	A	B	C	D
1	**Probeset**	**Gene Name**	**Array 1**	**Array 2**
2	**103941_at**	alpha-spectin 1, erythroid	33.7625	29.2333
3	**104432_at**	aplysia ras-related homolog N (Rh	127.736	99.6895
4	**104137_at**	ATP-binding cassette, sub-family /	109.522	65.2727
5	**98458_at**	baculoviral IAP repeal-containing 5	128.96	123.371
6	**93243_at**	bone morphogenetic protein 7	174.85	174.019
7	**95061_at**	breast carcinoma amplified sequer	34.8	43.6696
…				
12600	**102632_at**	calmodulin binding protein 1	69.888	54.7391

Expression index

Microarray Life Cycle



Biological Questions

1. Differentially expressed genes
2. Relationships between genes, tissues or treatments
3. Classification of tissues and samples

…

Experimental Design

Microarray Experiment

1. Target preparation
2. Hybridization
3. Washing
4. Image acquisition

Preprocessing

Quality Assessment

Image Analysis

Quantify Expression

Normalization

Filtering
Missing Values

Statistical Analysis

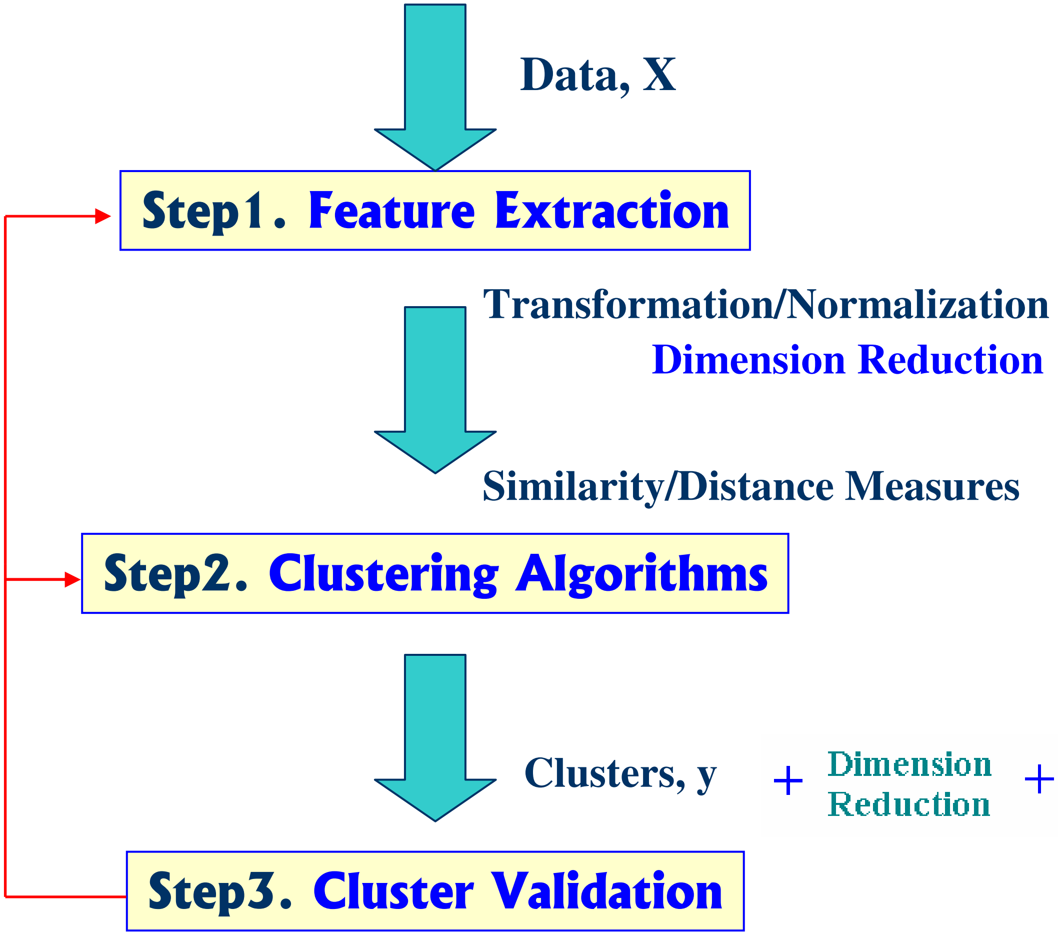
- Estimation
- Testing
- Clustering
- Classification
- Gene network

…

Biological Verification and Interpretation

Cluster Analysis (Unsupervised Learning)

Group a given collection of unlabeled patterns into meaningful clusters.



Data, X

Step1. Feature Extraction

Transformation/Normalization

Dimension Reduction

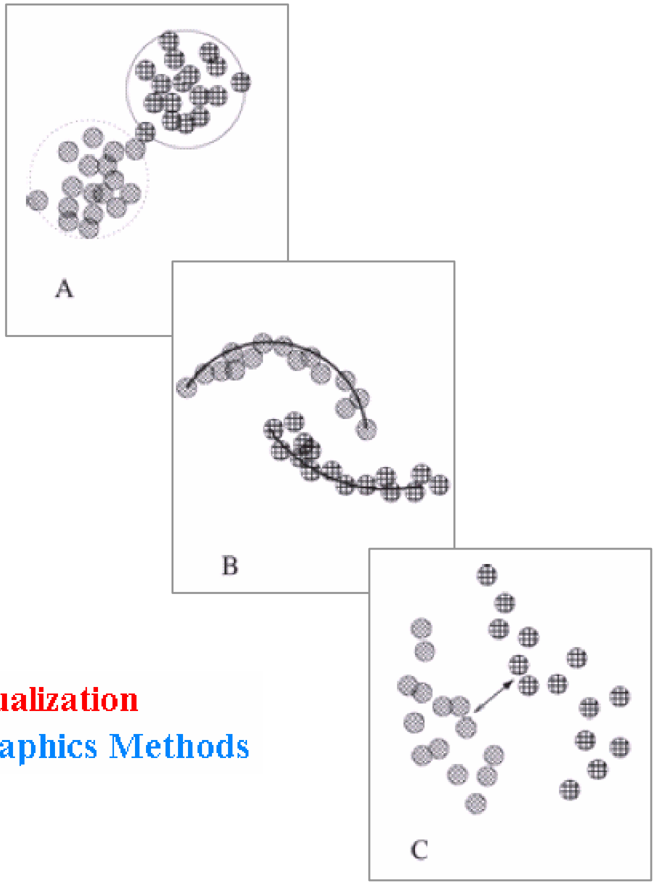
Similarity/Distance Measures

Step2. Clustering Algorithms

Clusters, y

Step3. Cluster Validation

+ Dimension Reduction + Visualization Graphics Methods



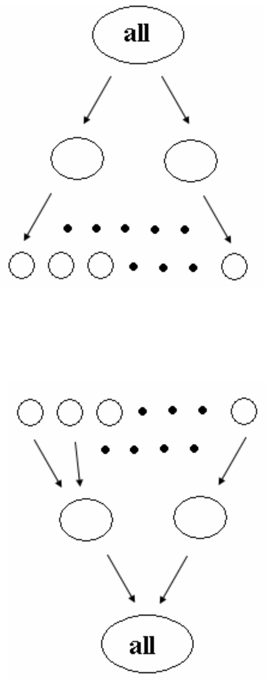
Daxin Jiang, Chun Tang and Aidong Zhang, (2004), **Cluster analysis for gene expression data: a survey**, IEEE Transactions on Knowledge and Data Engineering 16(11), 1370- 1386.

Clustering Analysis

Hierarchical clustering

The result is a tree that depicts the relationships between the objects.

- **Divisive clustering:**
  begin at step 1 with all the data in one cluster.
- **Agglomerative clustering:**
  all the objects start apart., there are n clusters at step 0.



Non-Hierarchical clustering

- k-means, The EM algorithm, K Nearest Neighbor,…

Two important properties of a clustering definition:

1. Most of data has been organized into **non-overlapping clusters**.
2. Each cluster has a within variance and one between variance for each of the other clusters. A good cluster should have a small within variance and large between variance.

Data/Information Visualization

What is Visualization?

- To visualize = to make visible, to transform into pictures.
- Making things/processes visible that are not directly accessible by the human eye.
- Transformation of an abstraction to a picture.
- Computer aided extraction and display of information from data.

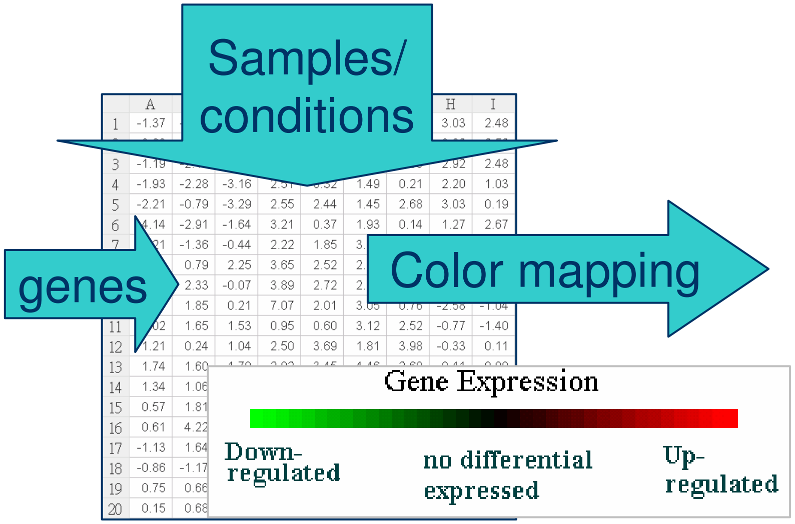
Data/Information Visualization

- Exploiting the human visual system to extract information from data.
- Provides an overview of complex data sets.
- Identifies structure, patterns, trends, anomalies, and relationships in data.
- Assists in identifying the areas of interest.

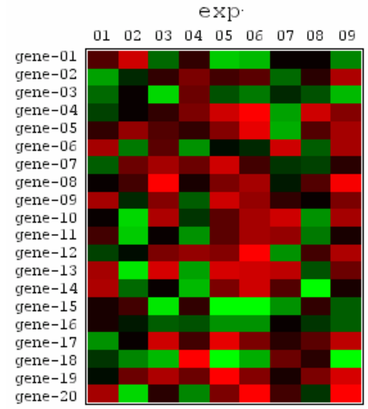
Visualization = Graphing for Data + Fitting + Graphing for Model

Tegarden, D. P. (1999). Business Information Visualization. Communications of AIS 1, 1-38.

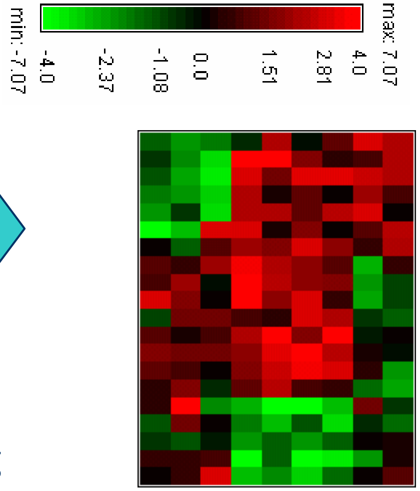
Visualizing Clustering Results: Heat Map



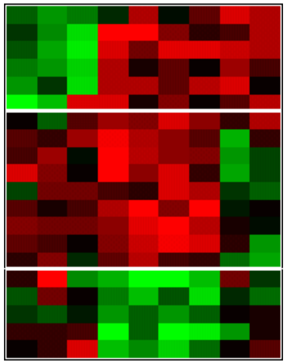
Without ordering



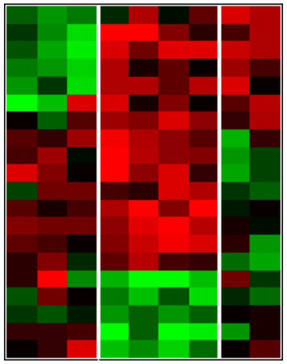
Ordering/ Seriation/ Clustering



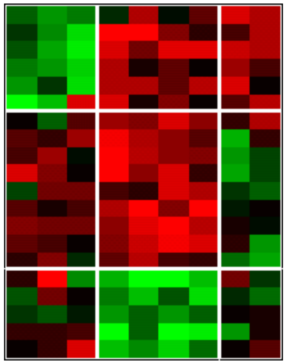
Gene-based clustering



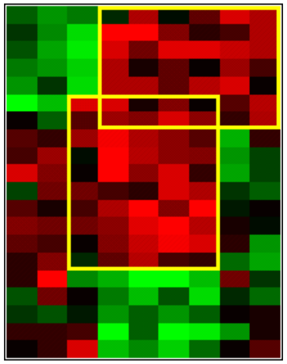
Sample-based clustering



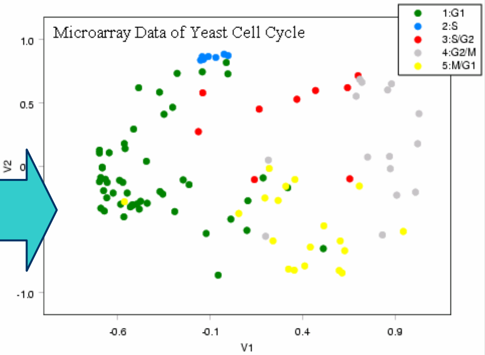
Twoway-based clustering



Subspace clustering



Dimension Reduction



e.g., K-means, SOM, Hierarchical Clustering, Model-based clustering,…

e.g., Bi-clustering

Clustering Analysis in Microarray Experiments

Goals

- Find natural classes in the data
- Identify new classes/gene correlations
- Refine existing taxonomies
- Support biological analysis/discovery

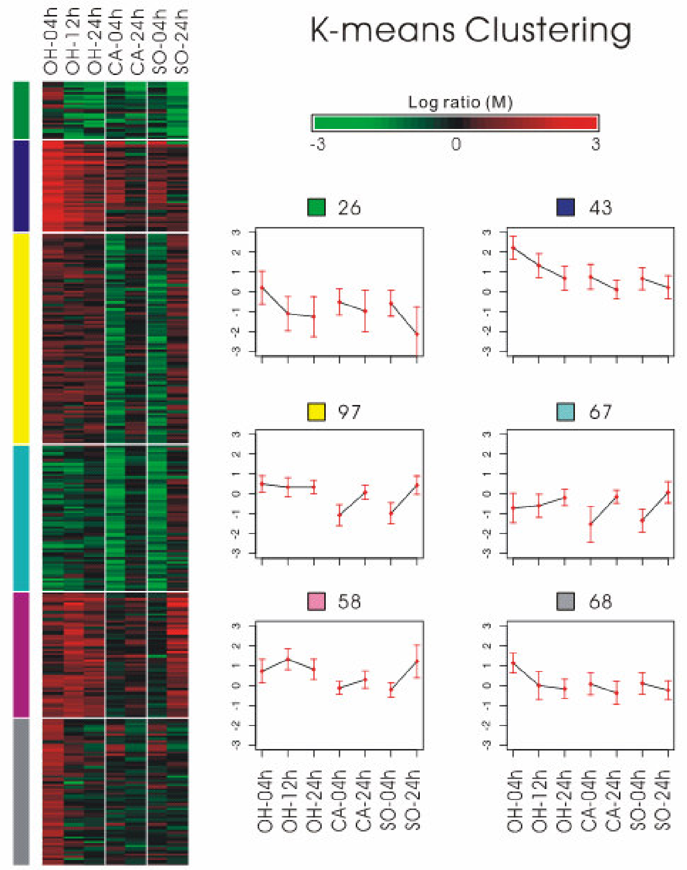
- cluster genes based on samples profiles
- cluster samples based on genes profiles

Hypothesis:

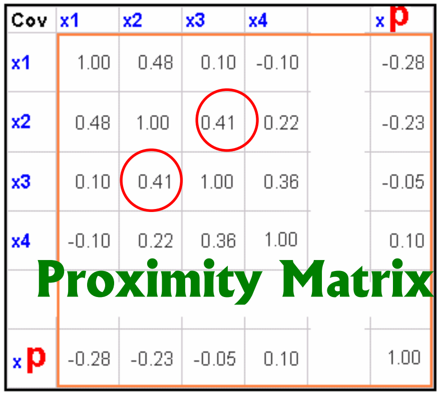
- genes with similar function have similar expression profiles.
- Clustering results in groups of co-expressed genes, groups of samples with a common phenotype, or blocks of genes and samples involved in specific biological processes.

Characteristic of Microarray Data:

- High-throughput, Noise, Outliers



Distance and Similarity Measure

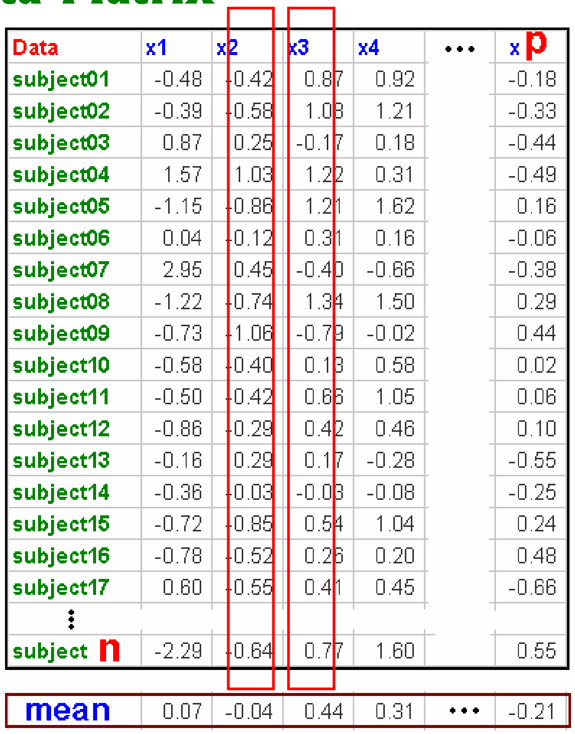


Cov	x1	x2	x3	x4		xp
x1	1.00	0.48	0.10	-0.10		-0.28
x2	0.48	1.00	0.41	0.22		-0.23
x3	0.10	0.41	1.00	0.36		-0.05
x4	-0.10	0.22	0.36	1.00		0.10
xp	-0.28	-0.23	-0.05	0.10		1.00

Proximity Matrix

Data Matrix

x y



Data	x1	x2	x3	x4	•••	xp
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-1.22	-0.74	1.34	1.50		0.29
subject09	-0.73	-1.06	-0.79	-0.02		0.44
subject10	-0.58	-0.40	0.13	0.58		0.02
subject11	-0.50	-0.42	0.66	1.05		0.06
subject12	-0.86	-0.29	0.42	0.46		0.10
subject13	-0.16	0.29	0.17	-0.28		-0.55
subject14	-0.36	-0.03	-0.03	-0.08		-0.25
subject15	-0.72	-0.85	0.54	1.04		0.24
subject16	-0.78	-0.52	0.26	0.20		0.48
subject17	0.60	-0.55	0.41	0.45		-0.66
⋮						
subject n	-2.29	-0.64	0.77	1.60		0.55
mean	0.07	-0.04	0.44	0.31	•••	-0.21

Pearson Correlation Coefficient

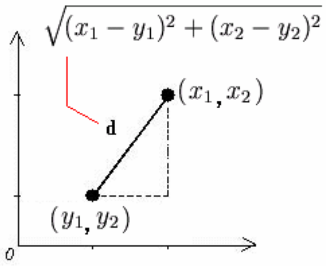
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$x = (x_1, x_2, \cdots, x_n)$$

$$y = (y_1, y_2, \cdots, y_n)$$

Euclidean Distance

$$d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

More Similarity Measures

Dissimilarity/Similarity Measure for Quantitative Data

Kendall’s tau

Two pairs of observation (x_i, y_i) and (x_j, y_j)

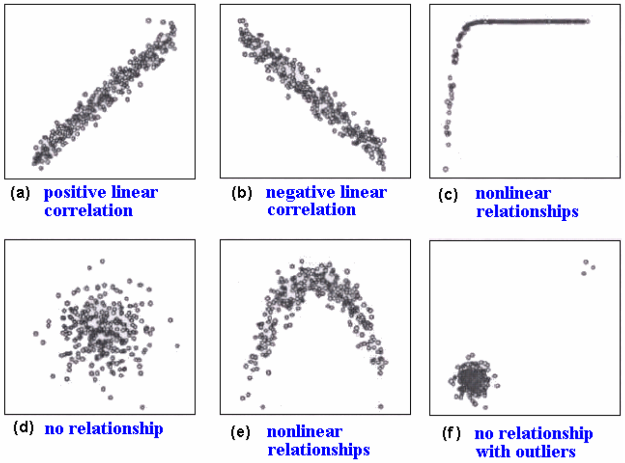
- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

E_y: extra y pair in x’s: $(x_j - x_i) = 0$

E_x : extra x pair in y’s: $(y_j - y_i) = 0$

$$\tau = \frac{C - D}{\sqrt{C + D - E_y}\,\sqrt{C + D - E_x}}$$

Similarity	Formula
Pearson correlation	$s(i,j) = \dfrac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i)\,\text{var}(x_j)}}$
Spearman correlation (r_i is ranked x_i)	$s(i,j) = \dfrac{\text{cov}(r_i, r_j)}{\sqrt{\text{var}(r_i)\,\text{var}(r_j)}}$
Kendall’s Tau	$s(i,j) = \dfrac{1}{\binom{p}{2}} \sum_{k \neq k'} sign\ [(x_{ik} - x_{ik'})(x_{jk} - x_{jk'})]$


(a) positive linear correlation
(b) negative linear correlation
(c) nonlinear relationships
(d) no relationship
(e) nonlinear relationships
(f) no relationship with outliers

Data	Pearson's rho	Spearman's rho	Kendall's tau
(a)	0.98	0.98	0.87
(b)	-0.98	-0.98	-0.87
(c)	0.50	0.99	0.98
(d)	-0.02	-0.03	-0.02
(e)	-0.06	-0.02	-0.02
(f)	0.68	0.00	0.00

- Pearson’s rho measures the strength of a linear relationship [(a), (b)].
- Spearman’s rho and Kendall’s tau measure any monotonic relationship between two variables [(a), (b) ,(c)].
- If the relationship between the two variables is non-monotonic, all three correlation coefficients fail to detect the existence of a relationship [(e)].
- Both Spearman’s rho and Kendall’s tau are rank-based non-parametric measures of association between variable X and Y.
- The rank-based correlation coefficients are more robust against outliers.

Algorithm they use different logic for computing the correlation coefficient, they seldom lead to markedly different conclusions (Siegel and Castellan, 1988).

K-Means Clustering

- K-means is a partition methods for clustering.
- Data are classified into k groups as specified by the user.
- Two different clusters cannot have any objects in common, and the k groups together constitute the full data set.

Optimization problem:

Minimize the sum of squared within-cluster distances

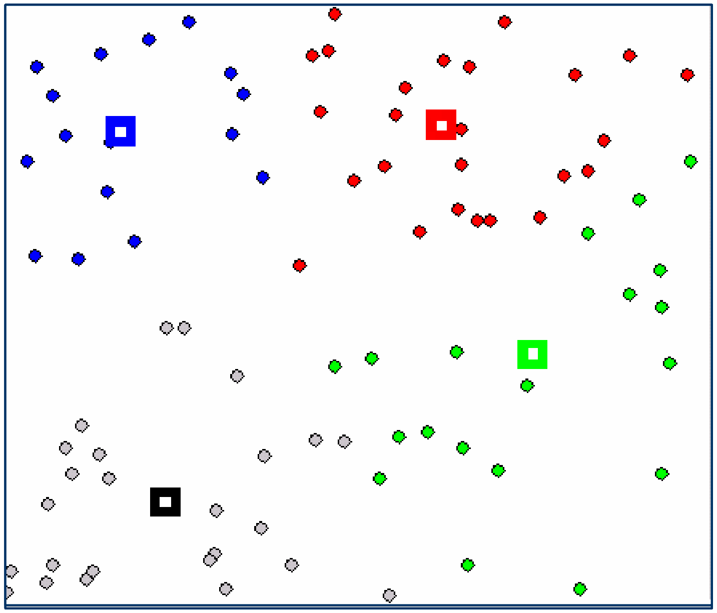
$$W(C) = \frac{1}{2}\sum_{k=1}^{K} \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

The K-Means Algorithm

1. The data points are randomly assigned to one of the K clusters.
2. The position of the K centroids are determined (initial group centroids).
3. For each data point:
   - Calculate the distance from the data point to each cluster.
   - Assign data point to the cluster that has the closest centroid.
4. Repeat the above step until the centroids no longer move.

The choice of initial partition can greatly affect the final clusters that result.

Converged



Visualizing Clustering Results:

Dimension Reduction Techniques

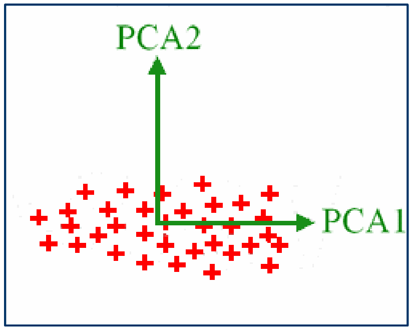
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)

Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.

Principal Component Analysis (PCA)

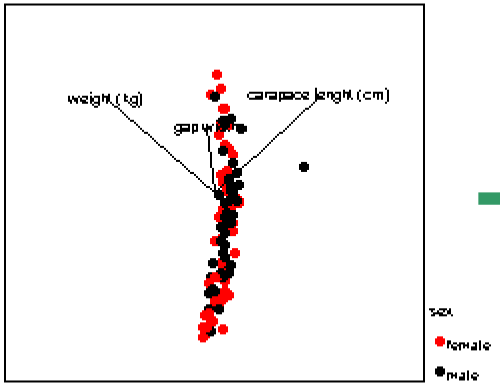
(Pearson 1901; Hotelling 1933; Jolliffe 2002)

PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



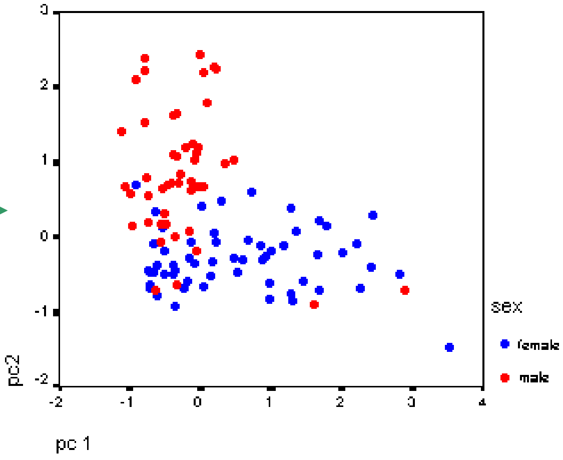
$$\mathrm{PCA}_1 = \boxed{a_1}\mathbf{X} + \boxed{b_1}\mathbf{Y}$$

$$\mathrm{PCA}_2 = \boxed{a_2}\mathbf{X} + \boxed{b_2}\mathbf{Y}$$



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y} + c_1\mathbf{Z}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$

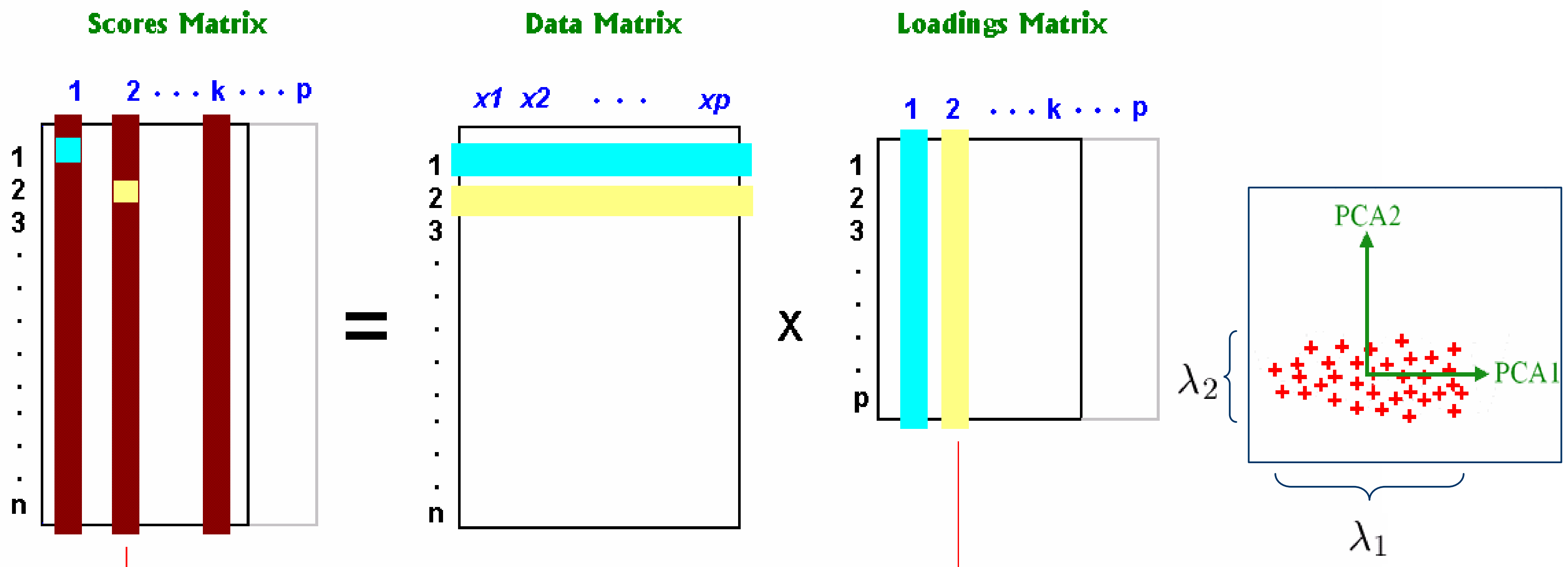


Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

$$\boxed{\begin{aligned}\mathrm{PCA}_1 &= a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p \\ \mathrm{PCA}_2 &= a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p\end{aligned}}$$

PCA: Loadings and Scores

$$\mathbf{Z} = \mathbf{X}\ \mathbf{W}$$



The ith principal component of $\mathbf{X}$ is $\mathbf{X}\mathbf{w}_i$, where $\mathbf{w}_i$ is the ith normalized eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to the ith largest eigenvaules.

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

$$\text{proportion} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

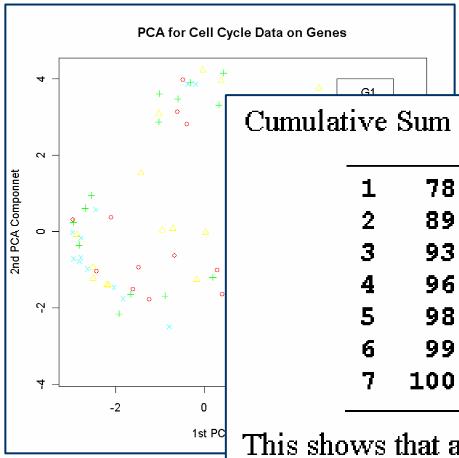
PCA (conti.)

Microarray Data Matrix

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

PCA on Conditions

MA Table	PCA-1	PCA-2	PCA-3
gene001	-0.18	-0.11	-0.03
gene002	0.51	-0.53	0.54
gene003	-0.35	-0.39	0.26
gene004	-0.18	-1.08	0.41
gene005	-0.62	-0.8	0.13
gene006	-0.09	-0.23	0.77
gene007	-0.38	-0.32	1.08
gene008	-0.88	-0.55	1.03
gene009	-1.26	0.45	0.41
gene010	0.12	-0.36	-0.16
gene011	-0.28	-0.44	2.13
gene012	-0.45	-0.23	0.82
gene013	-0.2	-0.43	0.44
gene014	0.03	-0.26	-0.68
gene015	-0.7	-0.76	0.5
gene016	-0.61	0.07	-0.04
gene017	-0.23	-0.71	0.01
gene018	0.1	0.1	0.11
gene019	-0.94	-0.97	0.24
gene020	-0.55	-0.53	0.86
gene021	-0.47	-0.87	-0.02
gene022	-0.34	-1.1	0.51
gene•••	-0.49	-0.2	0.91
gene n	-0.15	-1.04	-0.01



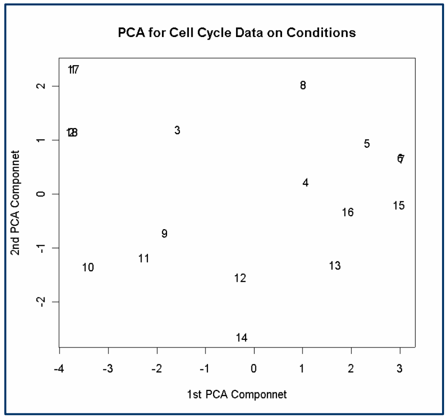
Cumulative Sum of the Variances:

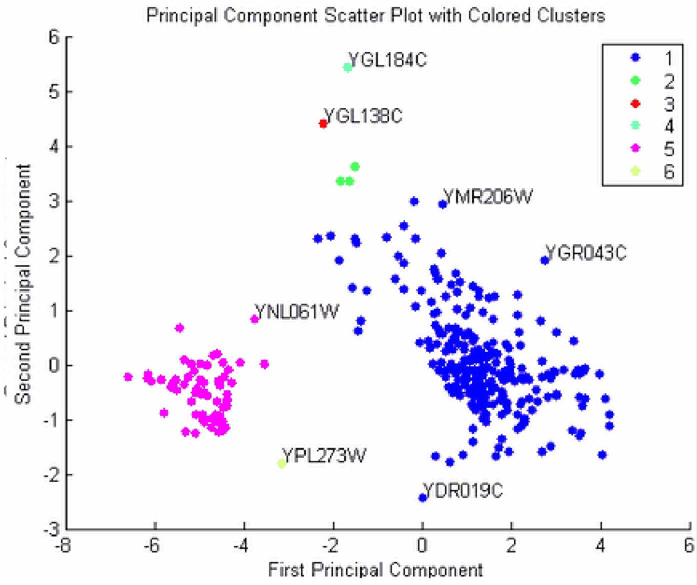
1	78.3719
2	89.2140
3	93.4357
4	96.0831
5	98.3283
6	99.3203
7	100.0000

This shows that almost 90% of the variance is accounted for by the first two principal components.

PCA on Genes

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
PCA-1	0.18	0.3	-0.12	-0.44	0.19	-0.39	-0.61
PCA-2	-0.16	-0.58	-0.43	-0.22	0.53	0.69	0.08
PCA-3	0.16	-0.44	-0.93	-1.23	-0.62	0.62	1.3

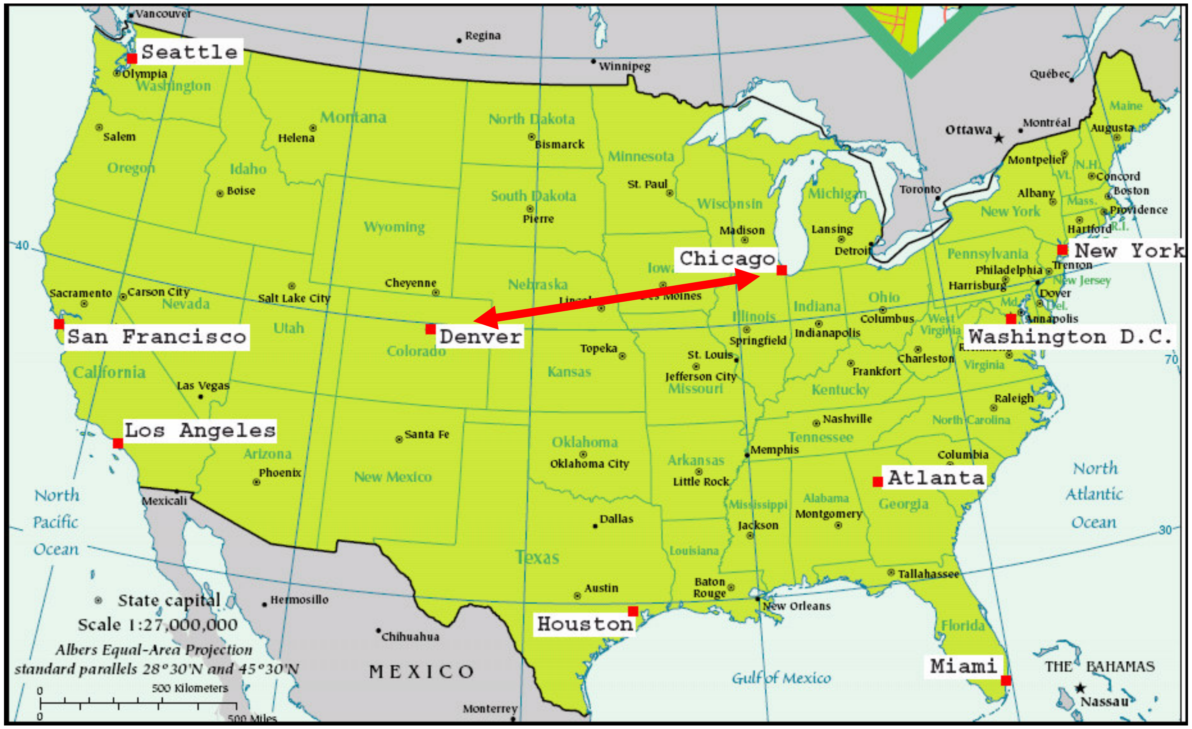




Yeast Microarray Data is from
DeRisi, JL, Iyer, VR, and Brown, PO.(1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale"; Science, Oct 24;278(5338):680-6.

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)



- Classical MDS takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.

- projection from some unknown dimensional space to 2-d dimension.

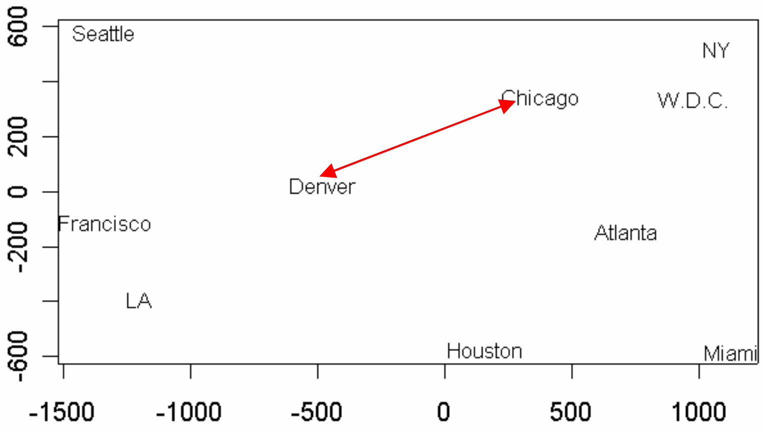
http://www.lib.utexas.edu/maps/united_states.html

?

Flying Mileages Between Ten U.S. Cities

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

MDS



MDS: Metric and Non-Metric Scaling

Question

Given a *dissimilarity matrix* D of certain objects, can we construct points in k-dimensional (often 2-dimensional) space such that

Goal of metric scaling

the Euclidean distances between these points approximate the entries in the dissimilarity matrix?

Goal of non-metric scaling

the order in distances coincides with the order in the entries of the dissimilarity matrix approximately?

$$S = \sum_{i,j} (\hat{d}_{ij} - d_{ij})^2$$

Mathematically: for given $\boldsymbol{k}$, compute points $\boldsymbol{x_1, \ldots, x_n}$ in $\boldsymbol{k}$-dimensional space such that the object function is minimized.

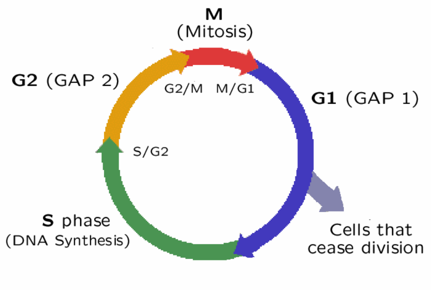
$$Stress = \sqrt{\frac{\sum_{i,j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

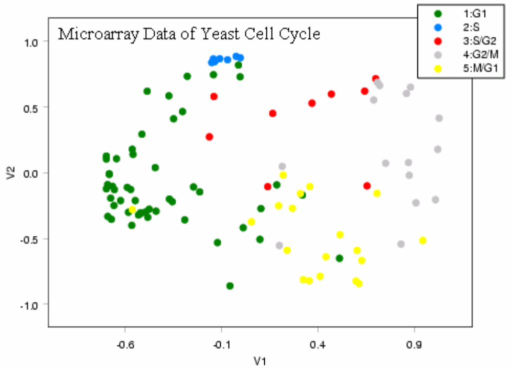
Microarray Data of Yeast Cell Cycle

- **Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)**

- **103 known genes: every 7 minutes and totally 18 time points.**

- **2D MDS Configuration Plot for 103 known genes.**



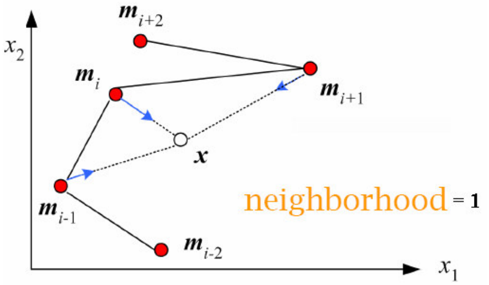


Clustering Analysis and Visualization

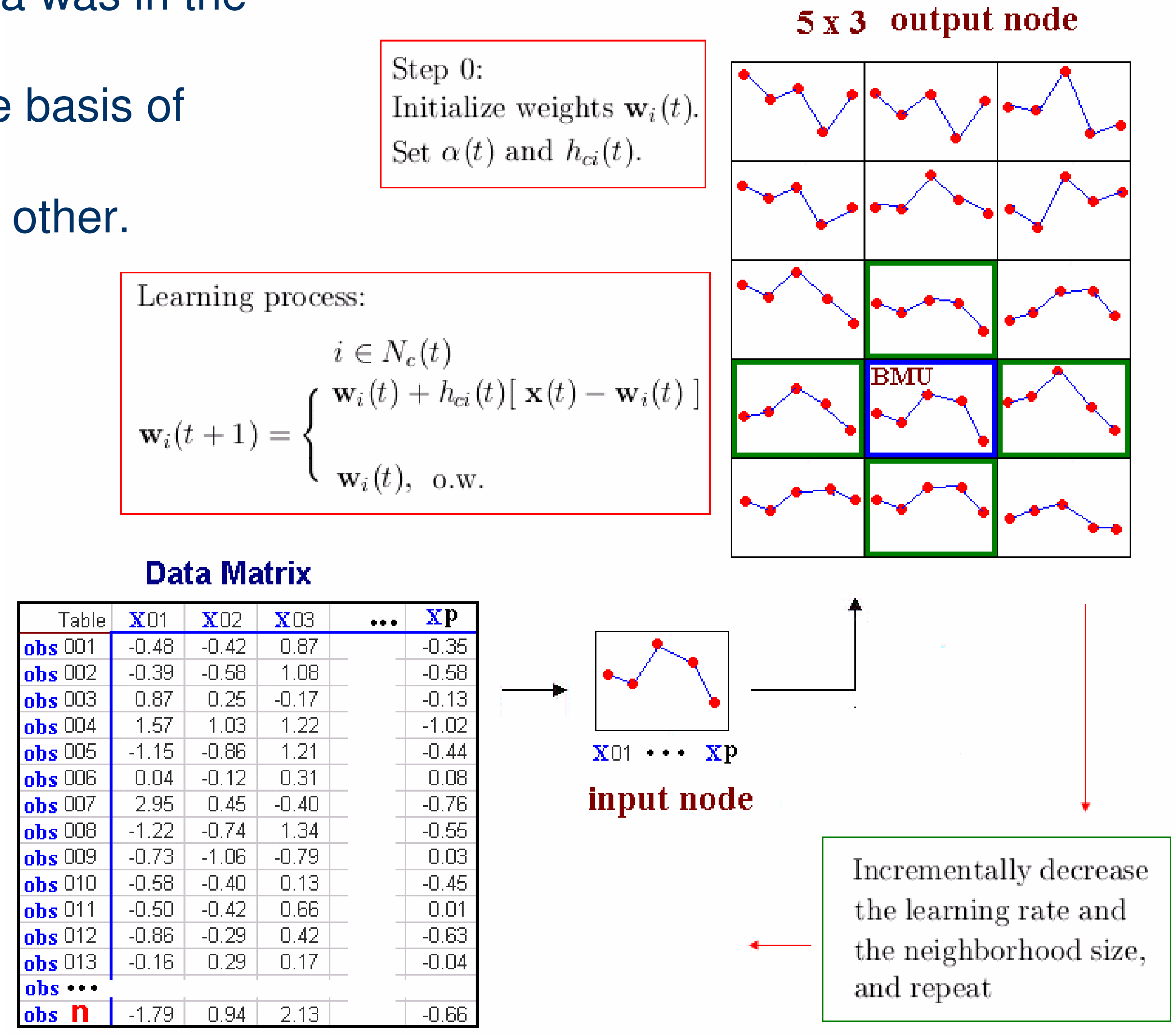
- Self-Organizing Maps (SOM)
- Heat Map
- Hierarchical Clustering

Self-Organizing Maps (SOM)

- ■ SOMs were developed by **Kohonen** in the early **1980's**, original area was in the area of speech recognition.
- ■ ***Idea:*** Organise data on the basis of **similarity** by putting entities **geometrically** close to each other.



- ■ SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display.



Step 0:
Initialize weights $\mathbf{w}_i(t)$.
Set $\alpha(t)$ and $h_{ci}(t)$.

Learning process:

$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,] & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

Data Matrix

Table	X01	X02	X03	•••	Xp
obs 001	-0.48	-0.42	0.87		-0.35
obs 002	-0.39	-0.58	1.08		-0.58
obs 003	0.87	0.25	-0.17		-0.13
obs 004	1.57	1.03	1.22		-1.02
obs 005	-1.15	-0.86	1.21		-0.44
obs 006	0.04	-0.12	0.31		0.08
obs 007	2.95	0.45	-0.40		-0.76
obs 008	-1.22	-0.74	1.34		-0.55
obs 009	-0.73	-1.06	-0.79		0.03
obs 010	-0.58	-0.40	0.13		-0.45
obs 011	-0.50	-0.42	0.66		0.01
obs 012	-0.86	-0.29	0.42		-0.63
obs 013	-0.16	0.29	0.17		-0.04
obs •••					
obs n	-1.79	0.94	2.13		-0.66

Incrementally decrease the learning rate and the neighborhood size, and repeat

Algorithm of SOM

Step 0: Initialize weights $\mathbf{w}_i(t)$.

Set topological neighborhood parameters $N_c(t)$.

Set learning rate parameters $\alpha(t)$ and $h_{ci}(t)$.

Step 1: For each input vector $\mathbf{x}(t)$, do

a. Finding a BMU: $\|\mathbf{x}(t) - \mathbf{w}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{w}_i(t)\|$

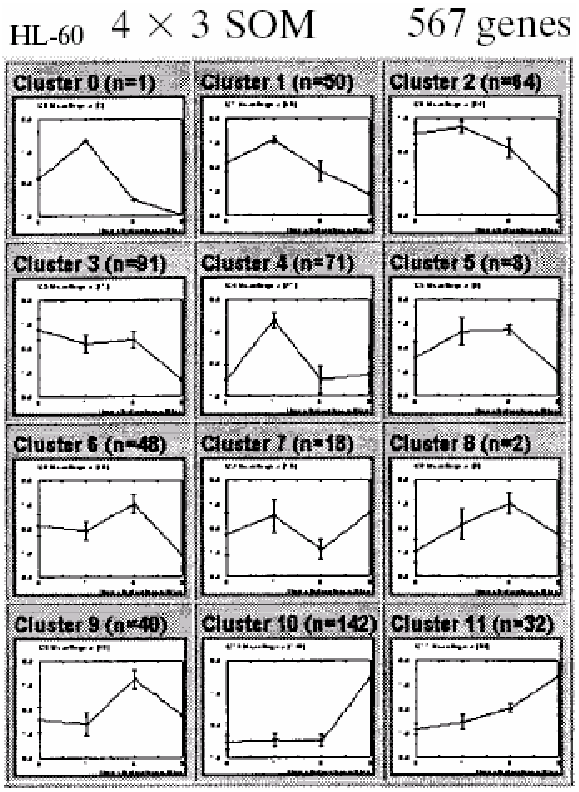
b. Learning process:

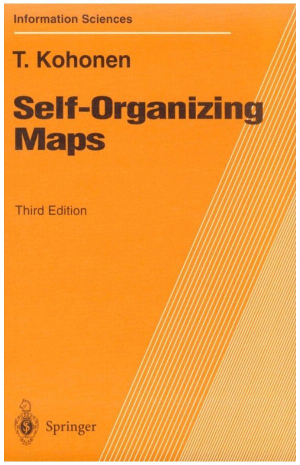
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

c. Go to the next unvisited input vector. If there are no unvisited input vector left then go back to the very first one and go to Step 2.

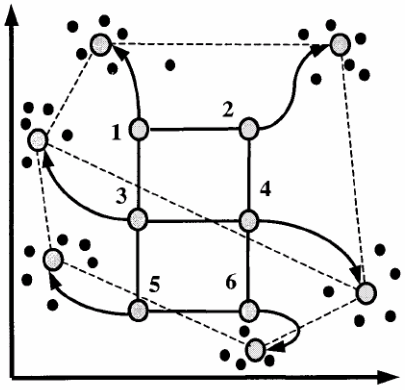
Step 2: Incrementally decrease the learning rate and the neighborhood size, and repeat Step 1.

Step 3: Keep doing Steps 1 and 2 for a sufficient number of iterations.





1995, 1997, 2001



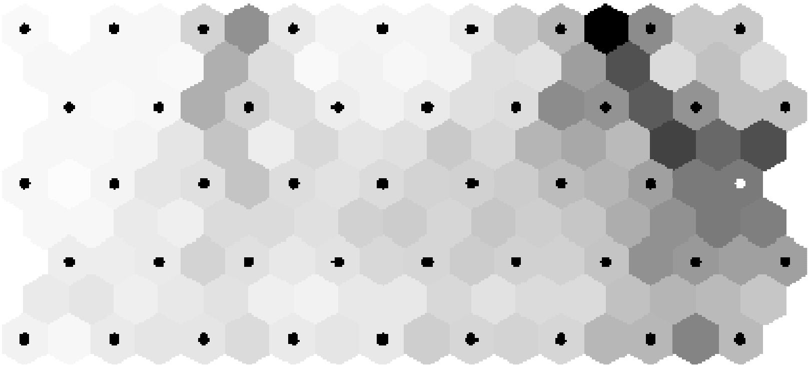
Macrophage Differentiation in HL-60 cells

Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96:2907-2912.

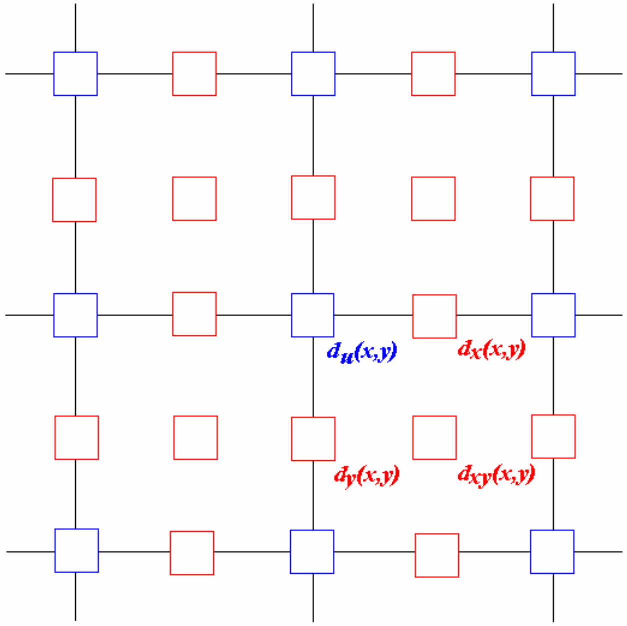
U-matrix: Unified Matrix Method

(Ultsch and Siemon 1989, Ultsch 1993)

U-matrix representation of SOM visualizes the distance between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes.



U-matrix representation of the SOM



$b(x,y)$: matrix of neurons, of size $n_x \times n_y$.

$w_i(x,y)$: matrix of weights.

$u(x,y)$: U-matrix of size $(2n_x - 1) \times (2n_y - 1)$.

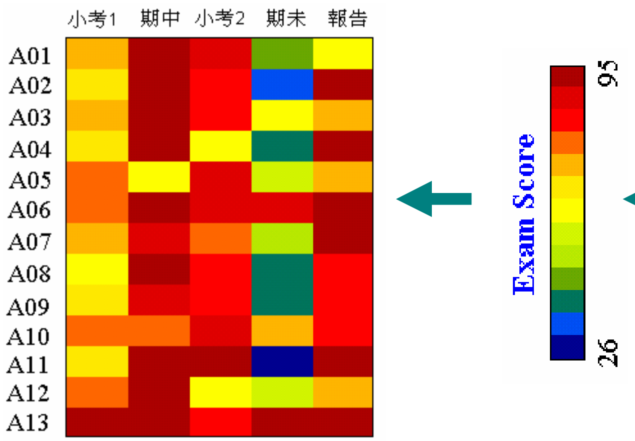
$d_x(x,y)$: $\|b(x,y) - b(x+1,y)\| = \sqrt{\sum_i [w_i(x,y) - w_i(x+1,y)]^2}$

$d_y(x,y)$: $\|b(x,y) - b(x,y+1)\| = \sqrt{\sum_i [w_i(x,y) - w_i(x,y+1)]^2}$

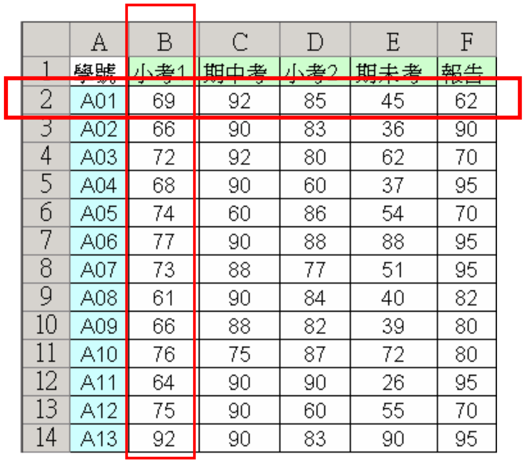
$d_{xy}(x,y)$: $\frac{1}{2}[\frac{\|b(x,y)-b(x+1,y+1)\|}{\sqrt{2}} + \frac{\|b(x,y+1)-b(x+1,y)\|}{\sqrt{2}}]$

$d_u(x,y)$: the median of the surrounding elements.

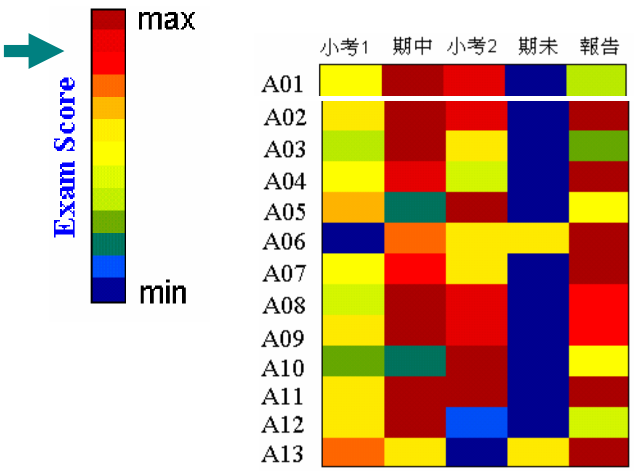
Heat Map: Data Image, Matrix Visualization



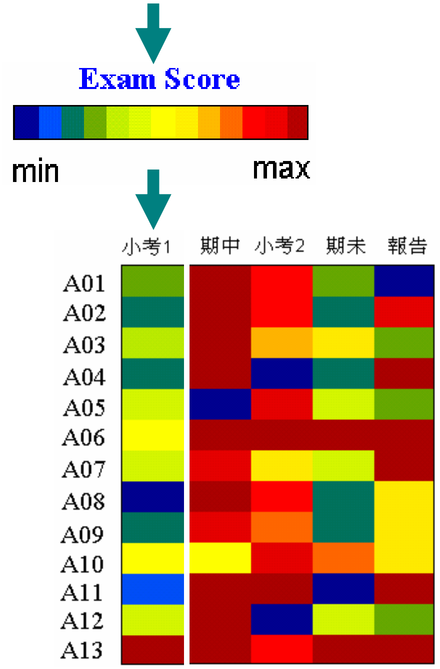
Range Matrix Condition



	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92	90	83	90	95

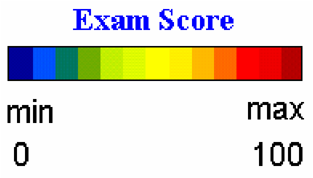


Range Raw Condition

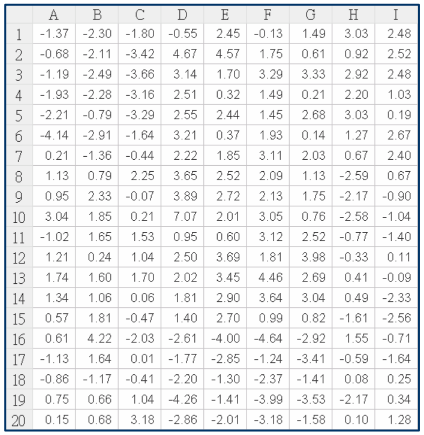


Range Column Condition

What about this one?

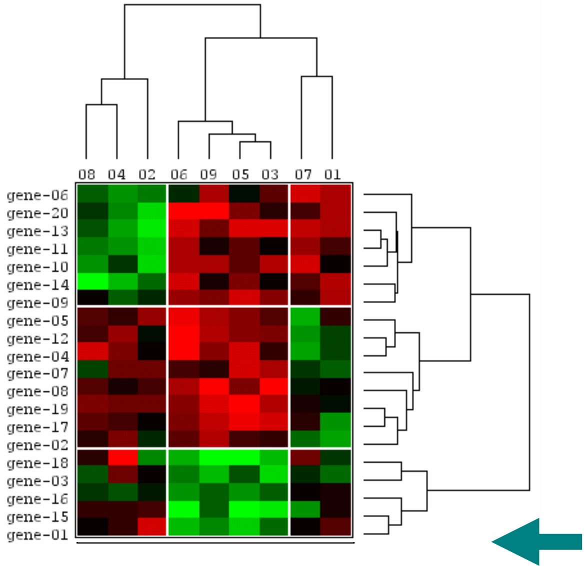


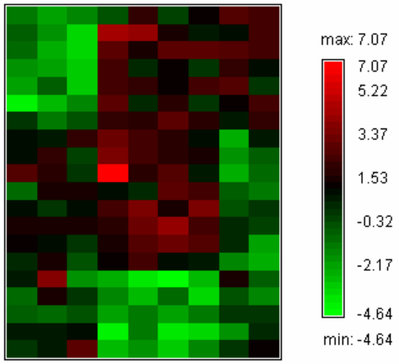
Heat Map: Display Conditions

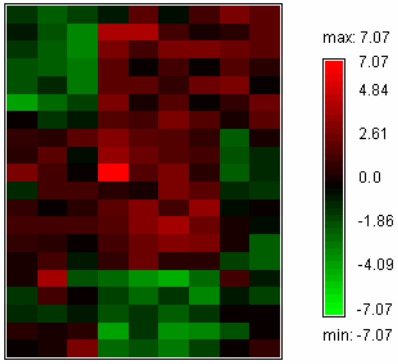


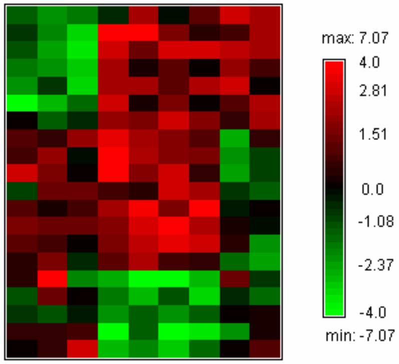
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28

Center Matrix Condition

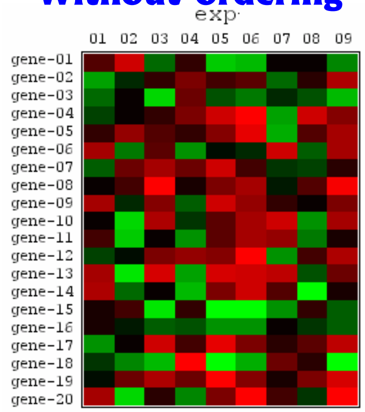








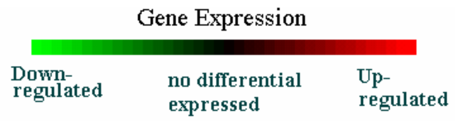
Without ordering

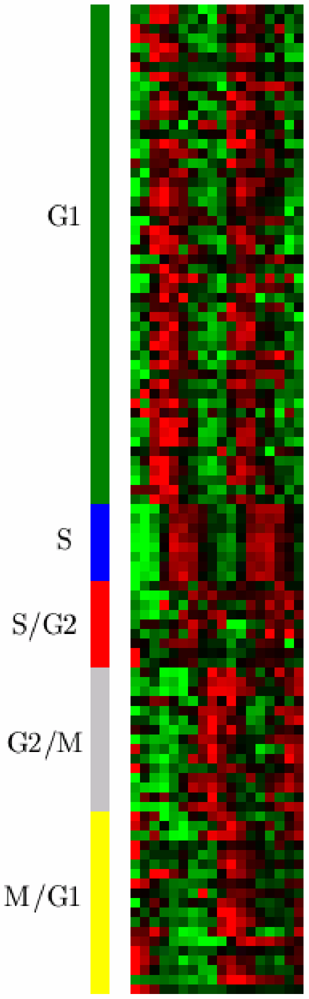


Microarray Data of Yeast Cell Cycle

- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)

- 103 known genes: every 7 minutes and totally 18 time points.



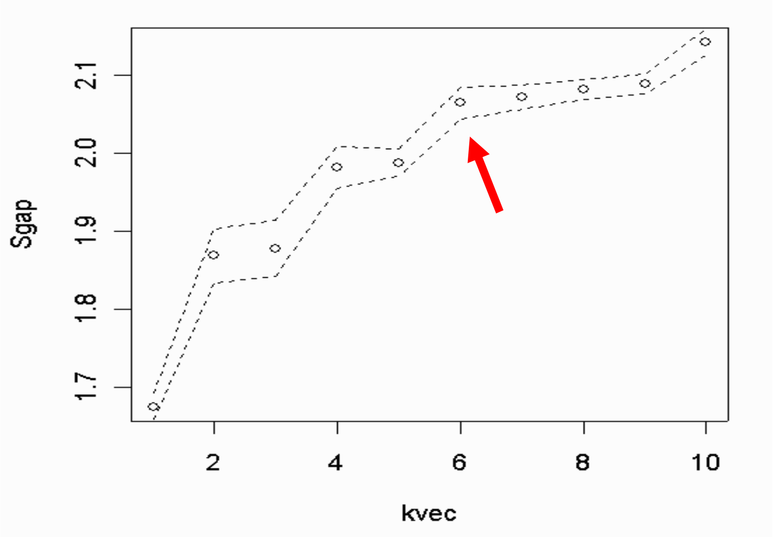


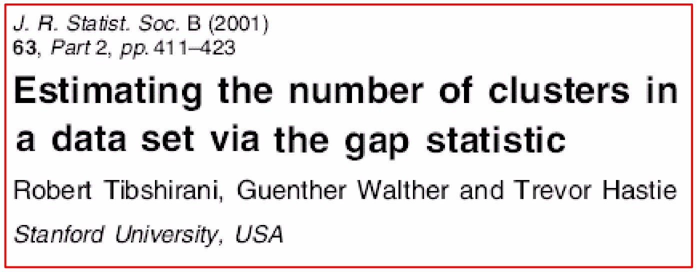
K-Means Clustering

- Data

  Baseline: Culture Medium (CM-00h)
  OH-04h, OH-12h, OH-24h
  CA-04h, CA-24h
  SO-04h, SO-24h

- A set of 359 genes was selected for

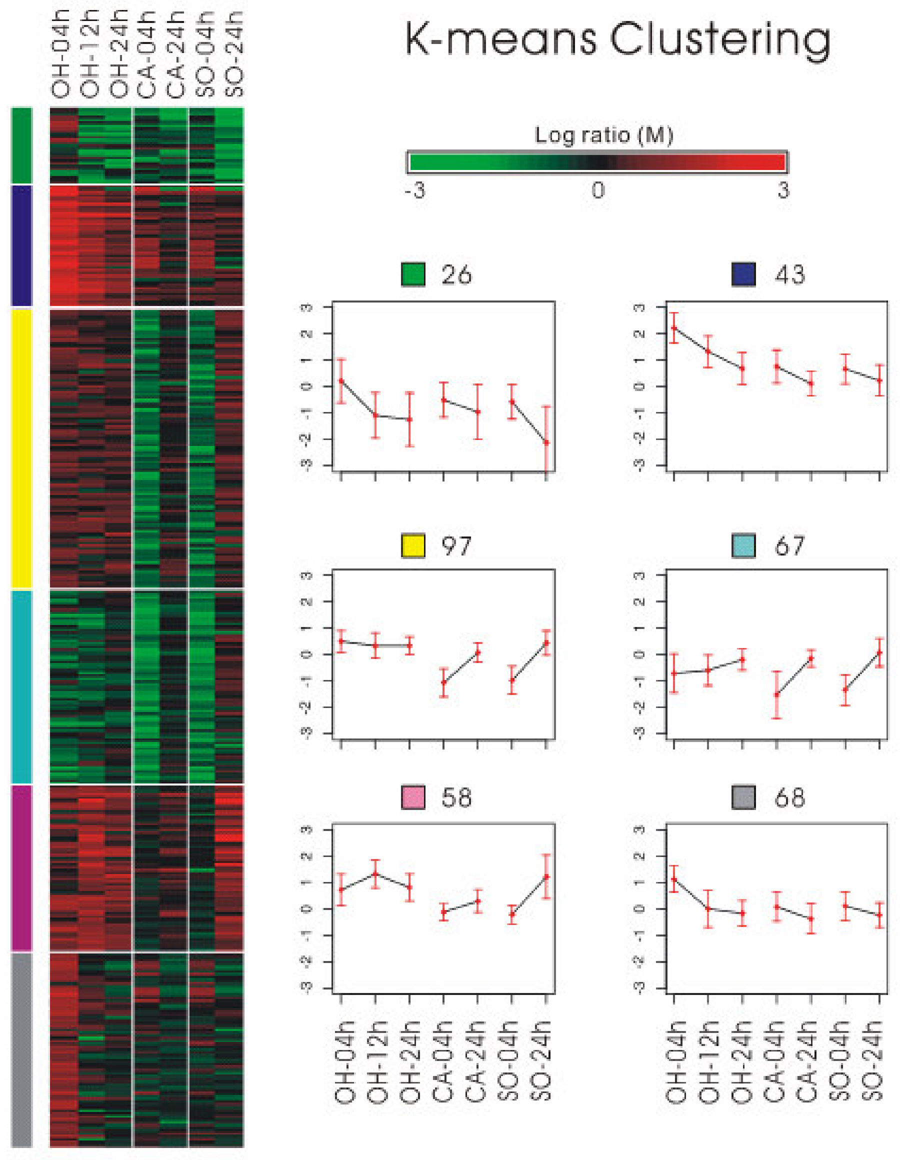



J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

Estimating the number of clusters in a data set via the gap statistic

Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA



Hierarchical Clustering and Dendrogram

(Kaufman and Rousseeuw, 1990)

Example: Average-Linkage

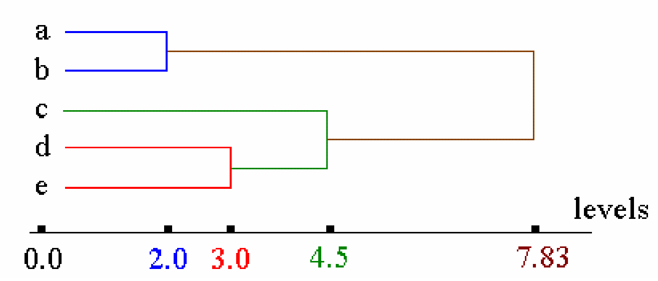
distance matrix

	a	b	c	d	e
a	0	2	6	10	9
b		0	5	9	8
c			0	4	5
d				0	3
e					0

	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

	{a, b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

	{a, b}	{c, d, e}
{a, b}	0	7.83
{c, d, e}		0



$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)]$$

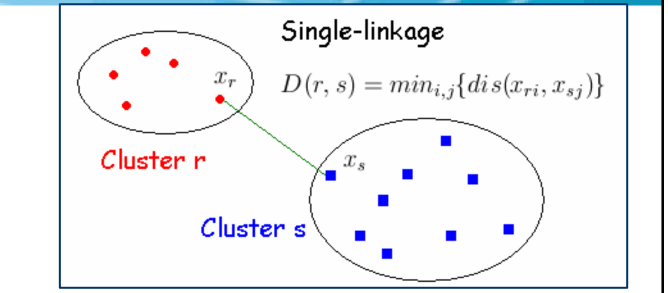
$$= \frac{1}{2}(6+5) = 5.5$$

$$D(\{a,b\},\{d,e\})$$

$$= \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$

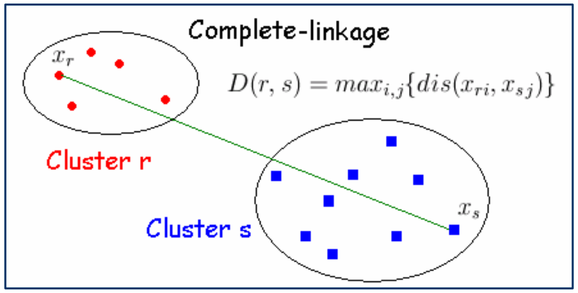
$$= \frac{1}{4}(10+9+9+8) = 9$$

UPGMA (Unweighted Pair-Groups Method Average)



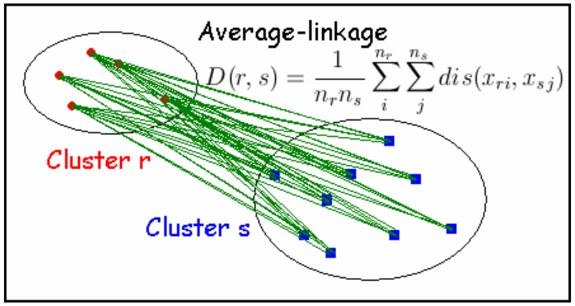
Single-linkage

$D(r,s) = min_{i,j}\{dis(x_{ri}, x_{sj})\}$



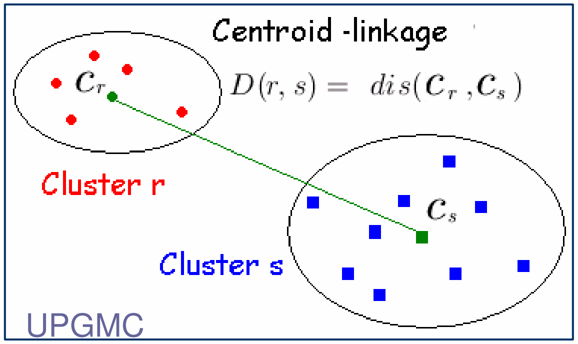
Complete-linkage

$D(r,s) = max_{i,j}\{dis(x_{ri}, x_{sj})\}$



Average-linkage

$D(r,s) = \frac{1}{n_r n_s}\sum_i^{n_r}\sum_j^{n_s} dis(x_{ri}, x_{sj})$

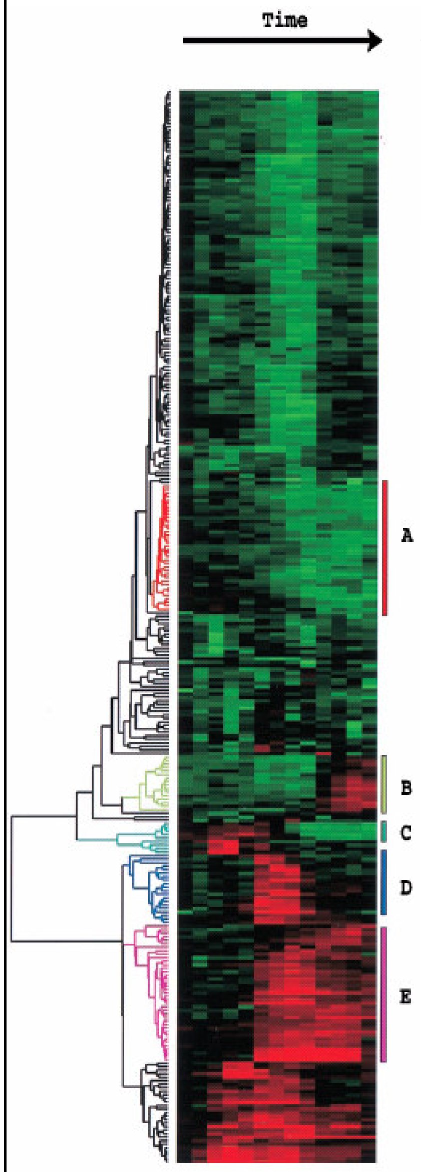


Centroid -linkage

$D(r,s) = dis(C_r, C_s)$

UPGMC

Display of Genome-Wide Expression Patterns



Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

Software:

Cluster and TreeView

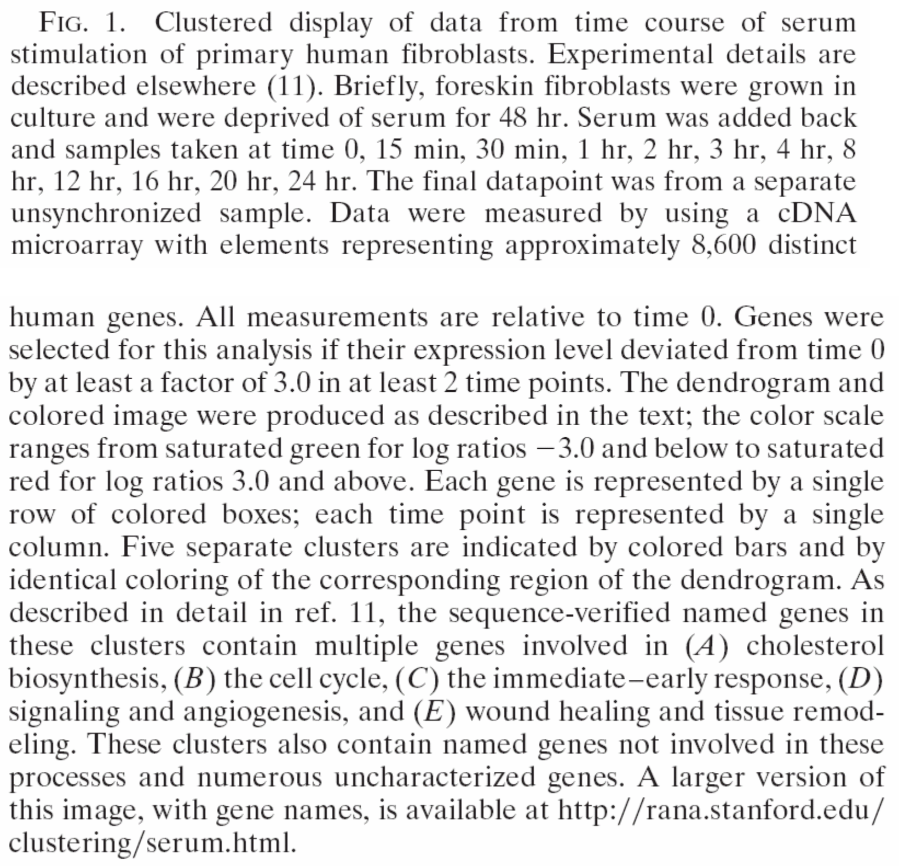


FIG. 1. Clustered display of data from time course of serum stimulation of primary human fibroblasts. Experimental details are described elsewhere (11). Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8,600 distinct human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated green for log ratios −3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. As described in detail in ref. 11, the sequence-verified named genes in these clusters contain multiple genes involved in (*A*) cholesterol biosynthesis, (*B*) the cell cycle, (*C*) the immediate–early response, (*D*) signaling and angiogenesis, and (*E*) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes. A larger version of this image, with gene names, is available at http://rana.stanford.edu/clustering/serum.html.

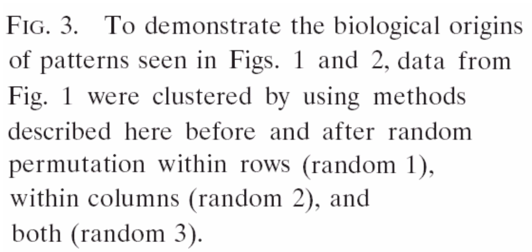
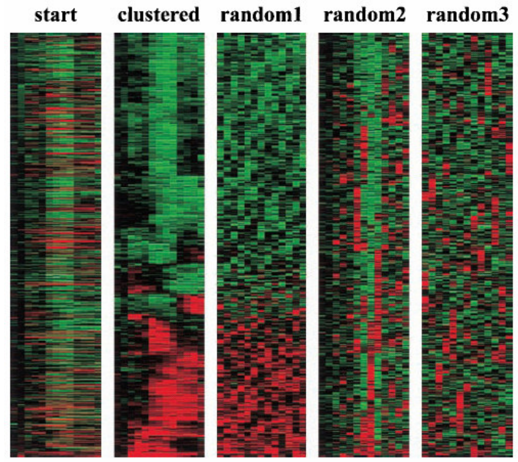
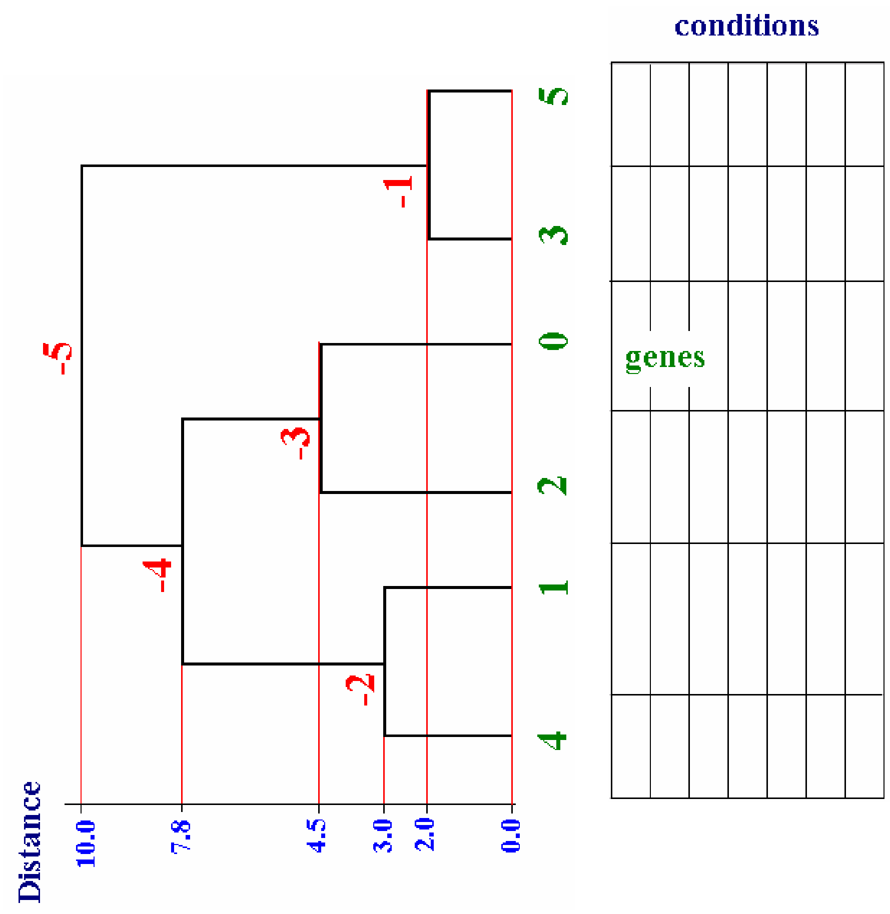
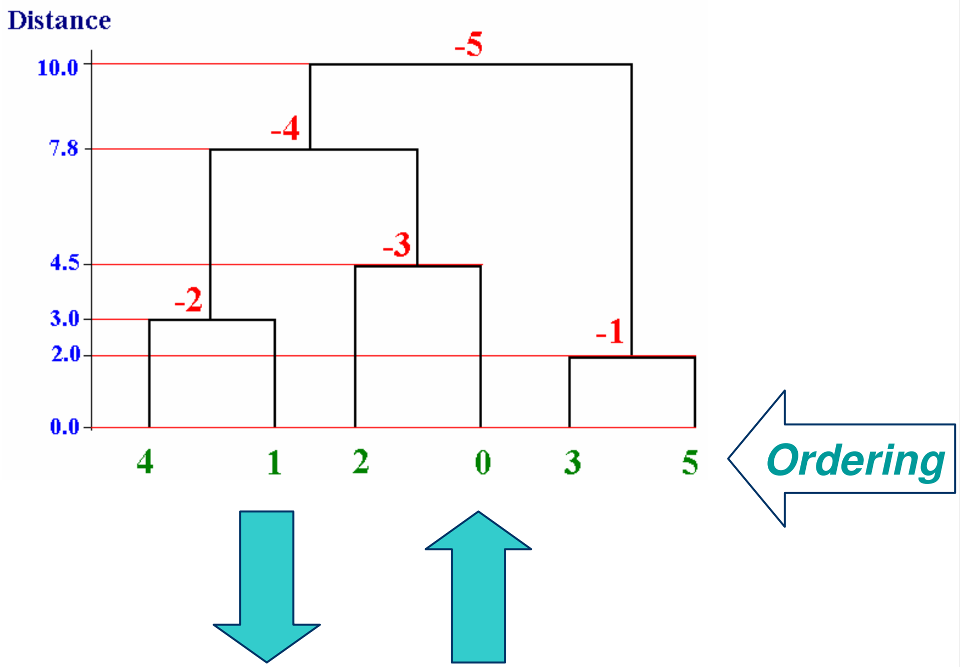


FIG. 3. To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).



Dendrogram and Tree Storage





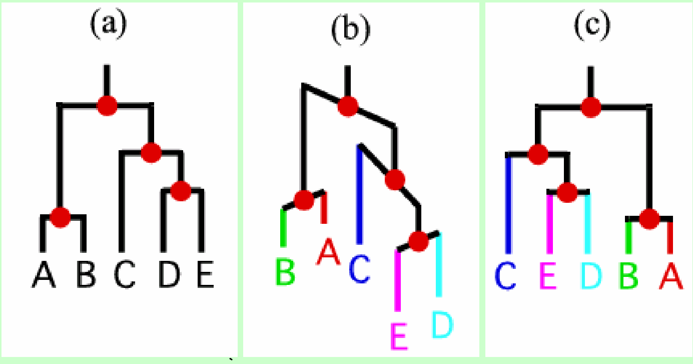
no	NodeID	Left	Right	Distance
0	-1	3	5	2
1	-2	4	1	3
2	-3	2	0	4.5
3	-4	-2	-3	7.8
4	-5	-4	-1	10

For example:

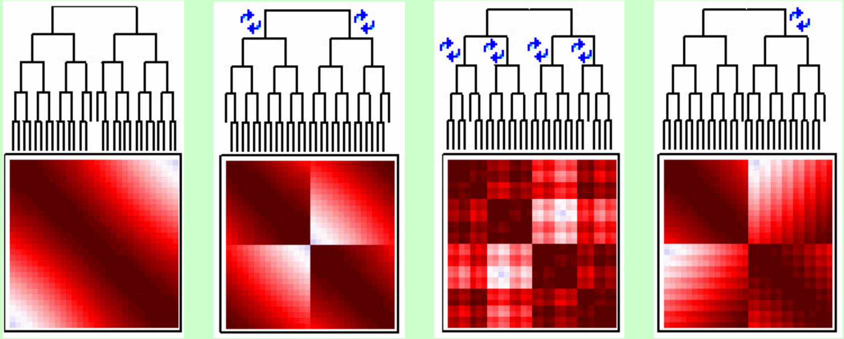
Cluster and TreeView, R

Seriation Problem for Hierarchical Clustering

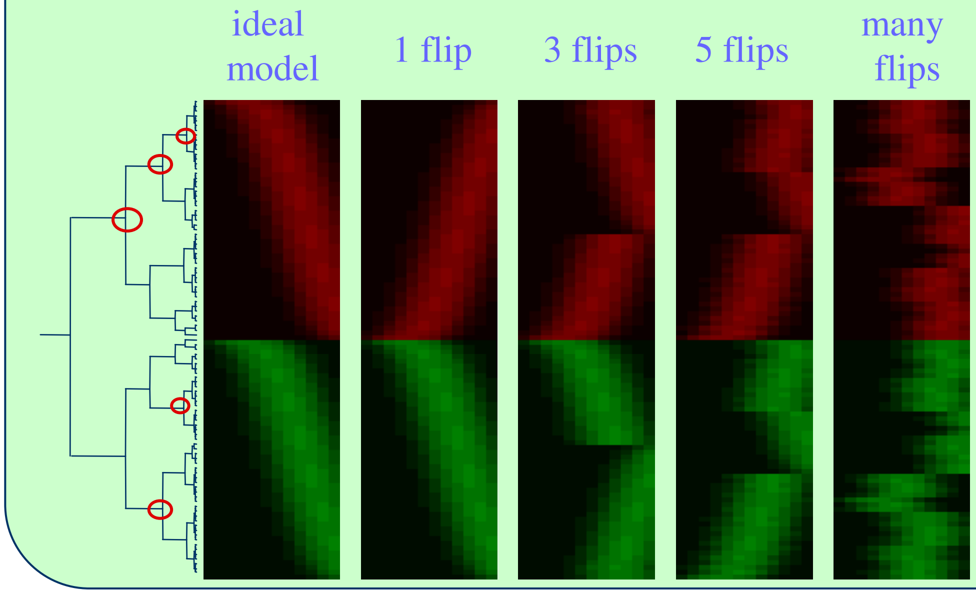
Tree seriation



Tree seriation for proximity matrices



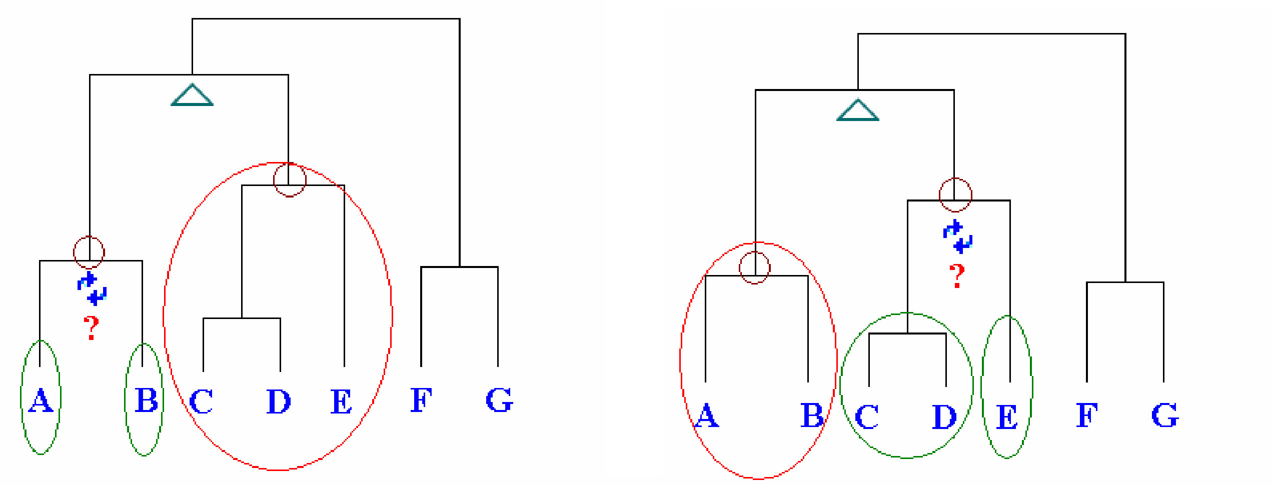
Tree seriation for raw data matrices



Different Seriations Generated from Identical Tree Structure

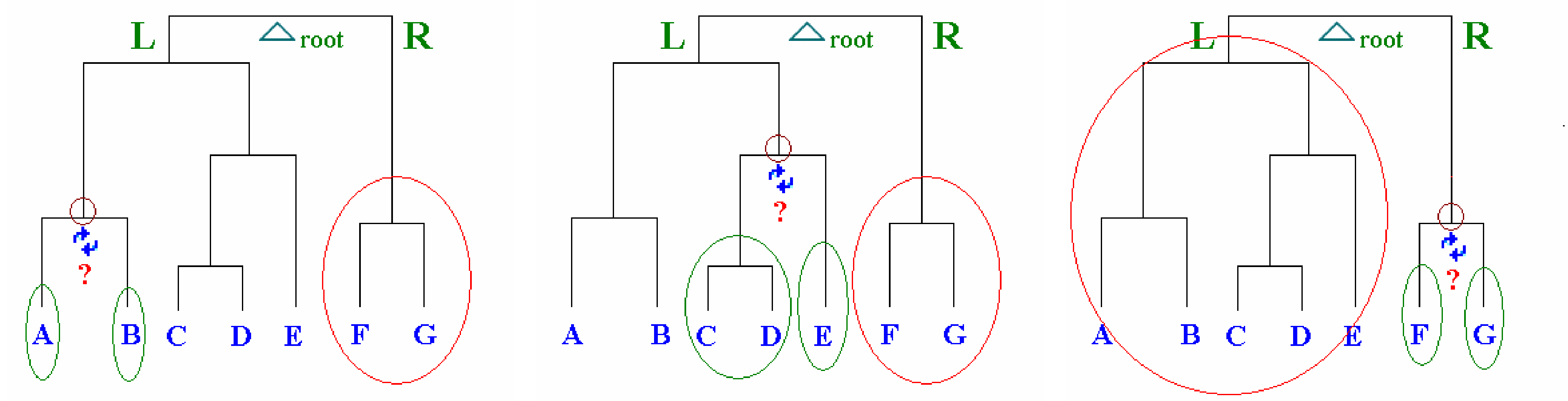
Internal Tree Flips

Uncle Approach



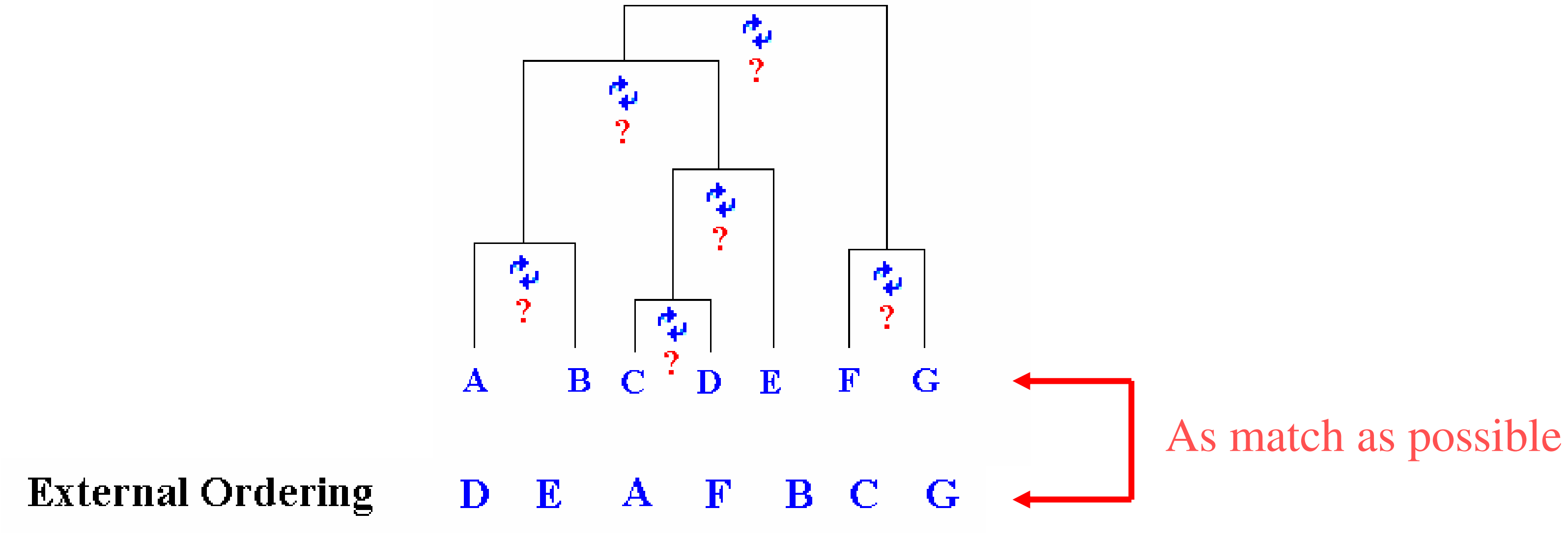
if d(A, {C,D, E}) < d(B, {C,D, E}) then flip

GrandPa Approach



Further reading: Ziv Bar-Joseph, David K. Gifford, and Tommi S. Jaakkola, (2001), **Fast Optimal Leaf Ordering** for Hierarchical Clustering. Bioinformatics 17(Suppl. 1):S22–S29.

External Tree Flips



External Ordering D E A F B C G

As match as possible

How to build an external ordering?

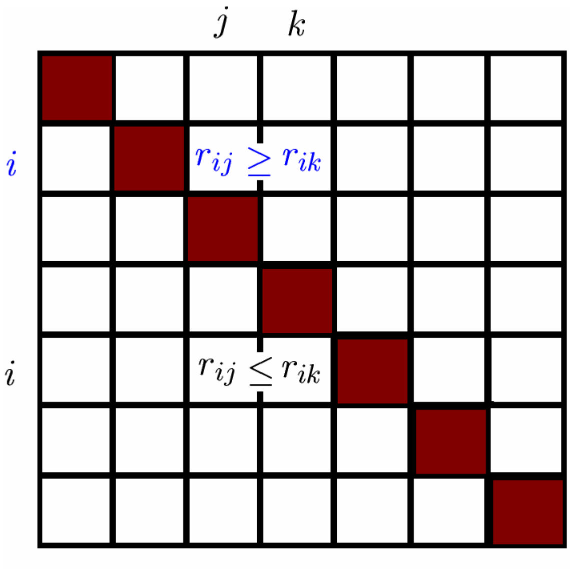
(1) Based on average expression level (Cluster Software, Eisen et al 1998)
(2) Using the results of a one-dimensional SOM
(3) …

Further reading: Tien, Y. J., Lee, Y. S, Wu, H. M. and Chen, C. H. (2006) Integration of clustering and visualization methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles.

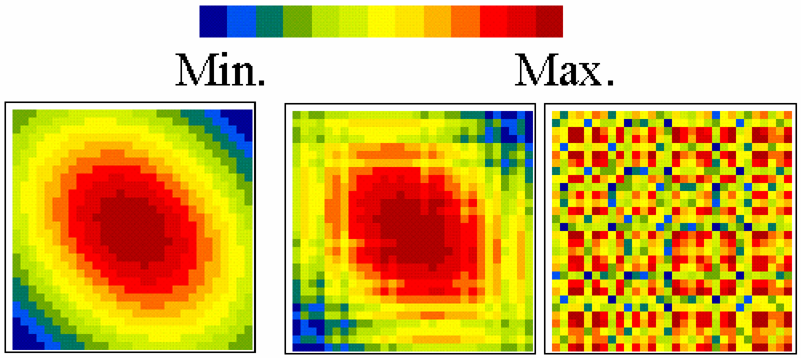
Criteria for a “good” Permutation

When T is symmetric, we usually want T’ to approximate a Robinson form (Robinson (1951)).

Robinson Form



$r_{ij} \leq r_{ik}$ if $j < k < i$, $r_{ij} \geq r_{ik}$ if $i < j < k$



Robinson **pre-Robinson**

Global/local Criterion:

Anti-Robinson Measurements

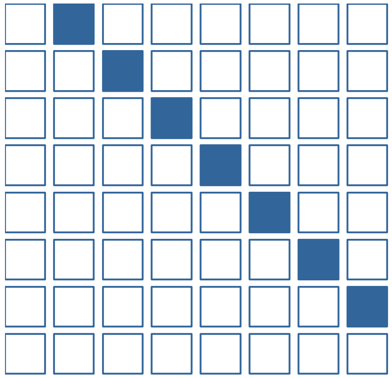
permuted matrix, $D = [d_{ij}]$

$$AR(i) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k} I(d_{ij} > d_{ik}) \Big],$$

$$AR(s) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}| \Big],$$

$$AR(w) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) |j-k| |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) |j-k| |d_{ij} - d_{ik}| \Big].$$

Local criterion: Minimal Span Loss Function



$$MS = \sum_{i=1}^{n-1} d_{i,i+1}$$

Further Reading

Michael Friendly , Ernest Kwan, (2003) Effect ordering for data displays, Computational Statistics & Data Analysis, v.43 n.4, p.509-539.

Global vs Local Seriation

GAP Elliptical Seriation
An algorithm for identifying global clustering patterns and smoothing temporal expression profiles

GAP Elliptical Seriation

Michael Eisen Tree Seriation

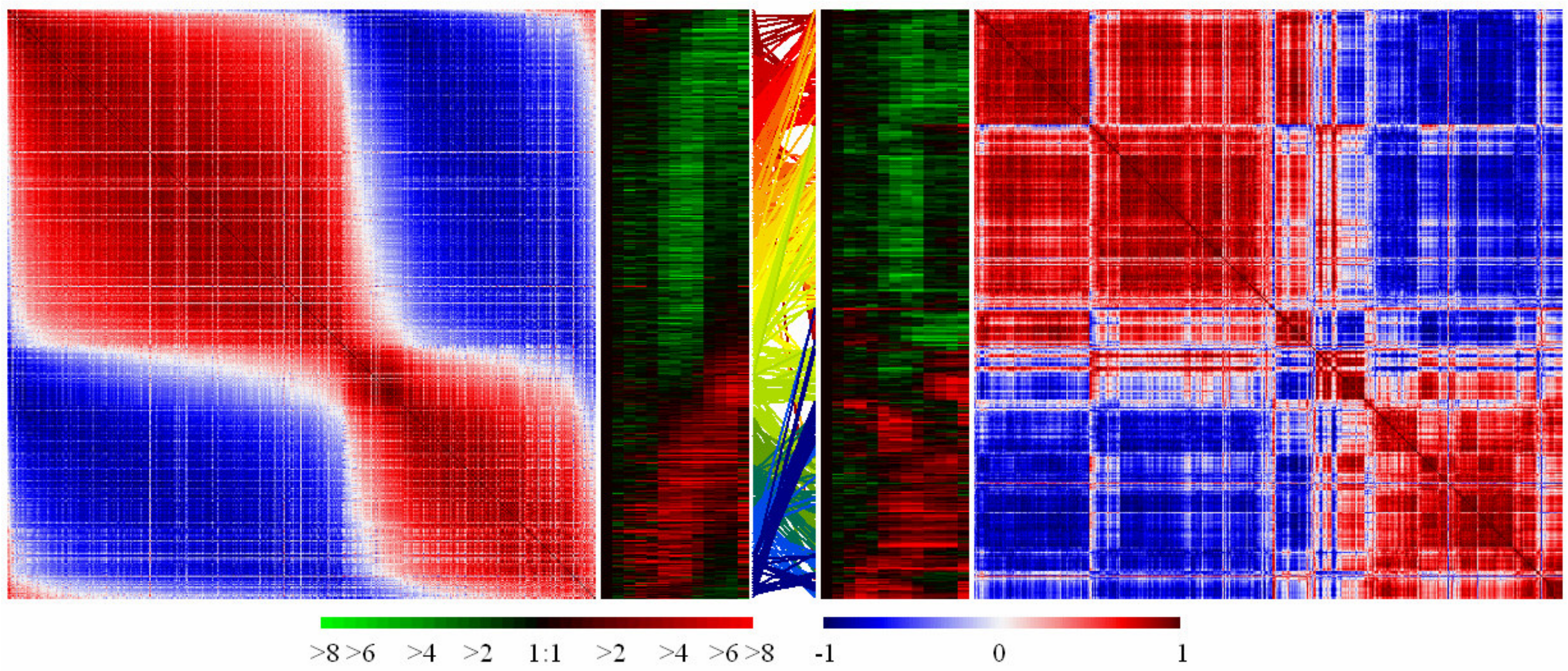
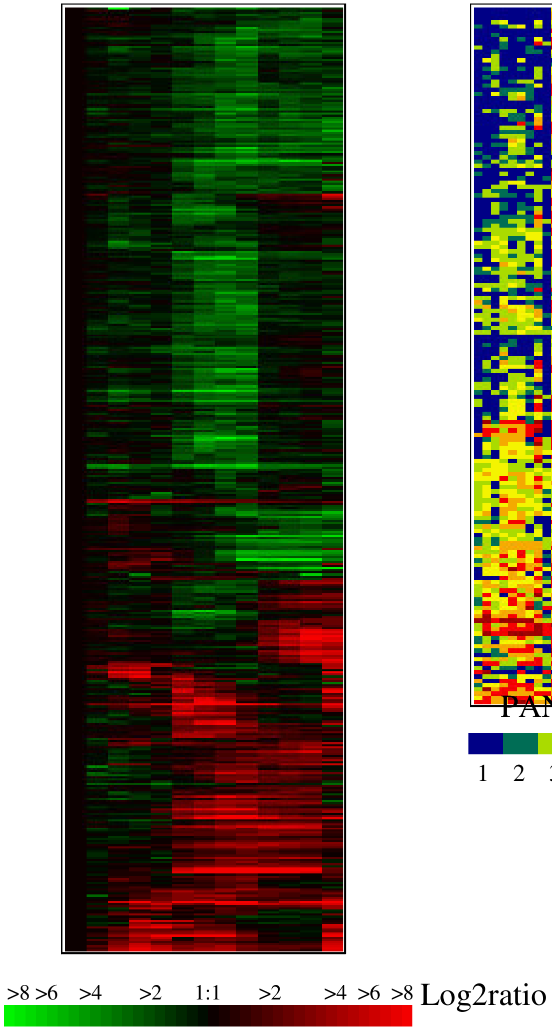


Image source: Dr. Chen Chun-houh’s slide

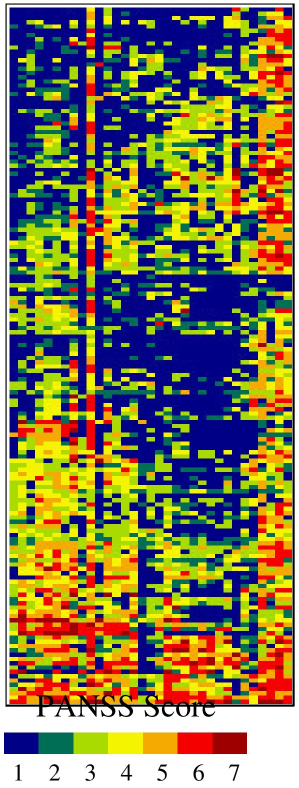
Visualization of Data Matrices

Simple ⟵ Information Visualization of Data Matrices ⟶ Difficult

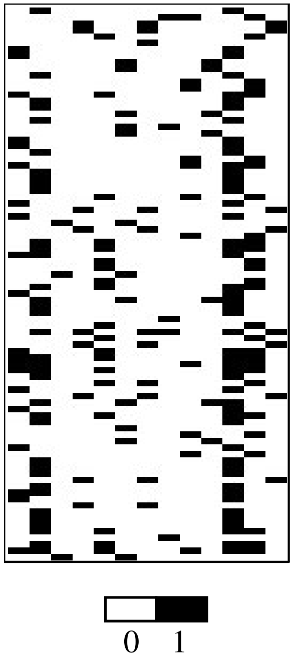
Continuous
(Gene/Time)



Ordinal
(Patient/Symptom)



Binary
(Mouse/Tumor)



Categorical
(Subject/SNP)

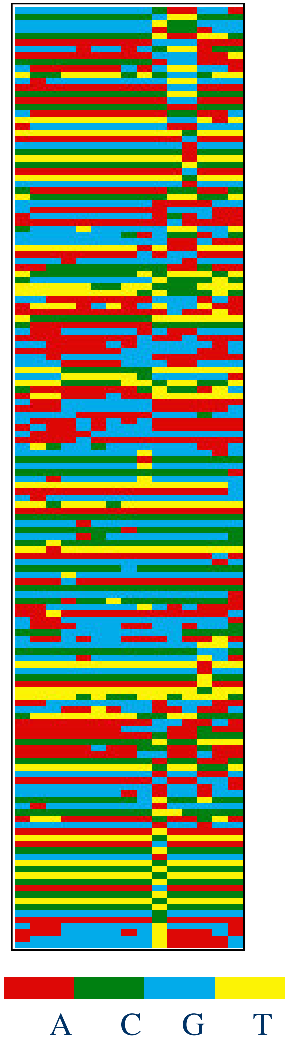


Image source: Chen Chun-houh's slide

Cluster Validation

Assess the quality and reliability of the cluster sets.

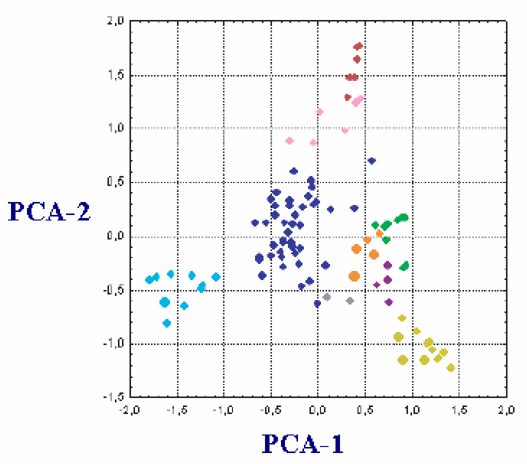
- **Quality:** **clusters can be measured in terms of homogeneity and separation.**
- **Reliability:** **cluster structure is not formed by chance.**
- **Ground Truth**: **from domain knowledge.**

NOTE:
Help to decide the number of clusters **in the data.**

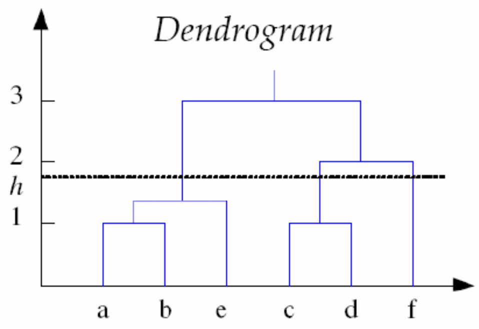
Choosing the Number of Clusters

(1) K is defined by the application.

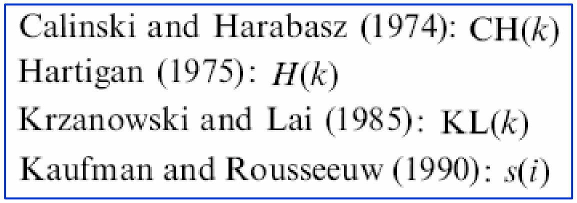
(2) Plot the data in two PAC dimensions.



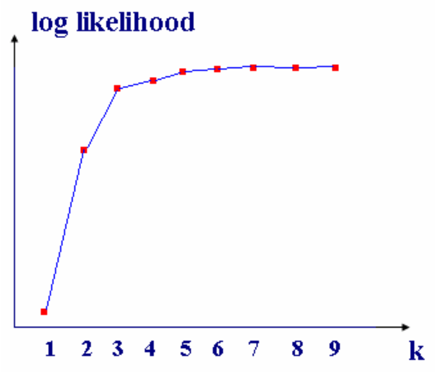
(4) Hierarchical clustering: look at the difference between levels in the tree.

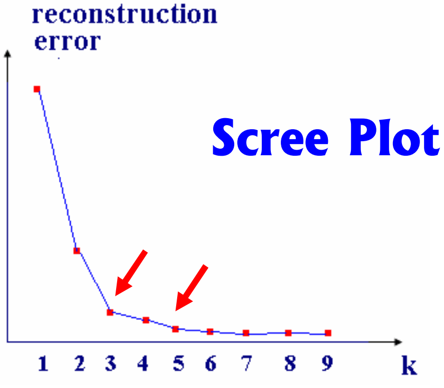


(e.g., k-means: within-cluster sum of squares)

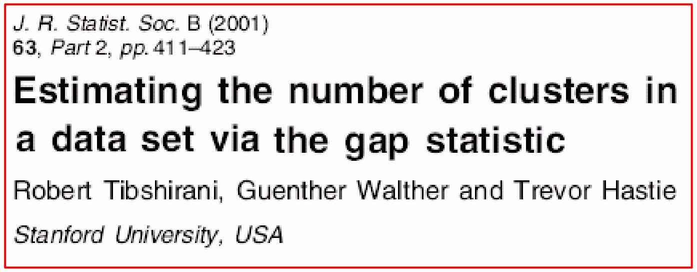

Calinski and Harabasz (1974): $\mathrm{CH}(k)$
Hartigan (1975): $H(k)$
Krzanowski and Lai (1985): $\mathrm{KL}(k)$
Kaufman and Rousseeuw (1990): $s(i)$

(3) Plot the **reconstruction error** or log likelihood as a function of k, and look for the elbow.





Scree Plot


J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

Estimating the number of clusters in a data set via the gap statistic

Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA

Literatures on Cluster Validation

2007

- Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh and Edward R. Dougherty, (2007), Model-based evaluation of clustering validation measures, Pattern Recognition 40(3), 807-824.
- Francisco R. Pinto, João A. Carriço, Mário Ramirez and Jonas S Almeida, (2007), Ranked Adjusted Rand: integrating distance and partition information in a measure of clustering agreement, BMC Bioinformatics, 8:44.

2006

- Susmita Datta and Somnath Datta, (2006), Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinformatics 2006, 7:397. [web]
- Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng and George C. Tseng, (2006), Evaluation and comparison of gene clustering methods in microarray analysis, Bioinformatics 22(19), 2405-2412.
- Giorgio Valentini , (2006), Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data, Bioinformatics, 22(3), 369-370.
- Susmita Datta and Somnath Datta, (2006), Evaluation of clustering algorithms for gene expression data, BMC Bioinformatics 2006, 7(Suppl 4):S17. [web]

2005

- Tibshirani, Robert; Walther, Guenther (2005), Cluster Validation by Prediction Strength, Journal of Computational & Graphical Statistics 14(3), pp. 511-528(18)
- Julia Handl, Joshua Knowles and Douglas B. Kell, (2005), Computational cluster validation in post-genomic data analysis, Bioinformatics 21(15), 3201-3212. [web] [supp]
- Nadia B,Francisco A,Padraig C. (2005), An integrated tool for microarray data clustering and cluster validity assessment, Bioinformatics 21:451. [Web]
- Julia Handl and Joshua Knowles, (2005) Exploiting the trade-off -- the benefits of multiple objectives in data clustering, Proceedings of the Third International Conference on E
- Nikhil R Garge, C ither? BMC Bioinformatics 2

More than 30 papers for Microarray!

2004

- Daxin Jiang, Chun Tang and Aidong Zhang, (2004), Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16(11), 1370- 1386. [web]
- Kimberly D. Siegmund, Peter W. Laird and Ite A. Laird-Offringa, (2004), A comparison of cluster analysis methods using DNA methylation data, Bioinformatics 20(12), 1896-1904.
- Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann, Stability-Based Validation of Clustering Solutions, Neural Comp. 2004 16: 1299-1323.

2003

- Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003 Mar 1;19(4):459-66.
- N. Bolshakova and F. Azuaje, (2003), Cluster validation techniques for genome expression data, Signal Processing 83(4), 825-833.

2001

- K. Y. Yeung, D. R. Haynor and W. L. Ruzzo, (2001), Validating clustering for gene expression data, Bioinformatics 17(4), 309-318. [web]
- Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis,(2001), On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17(2), 107 - 145.
- Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 2001 Jul 31;98(16):8961-5.
- Levine E, Domany E. Resampling method for unsupervised estimation of cluster validity. Neural Comput. 2001 Nov;13(11):2573-93.
- Maria Halkidi, Michalis Vazirgiannis, Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set, icdm, p. 187, First IEEE International Conference on Data Mining (ICDM'01), 2001

~2000

- Zhang K, Zhao H. Assessing reliability of gene clusters from gene expression data. Funct Integr Genomics. 2000 Nov;1(3):156-73.
- Xie, X.L. Beni, G. (1991), A validity measure for fuzzy clustering, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 13(8), 841-847.
- Peter Rousseeuw, (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20(1), 53-65.
- Lawrence Hubert and Phipps Arabie (1985), Comparing partitions, Journal of Classification 2(1), 193-218.
- Wallace, D. L. 1983. A method for comparing two hierarchical clusterings: comment. Journal of the American Statistical Association 78:569-576.
- E. B. Fowlkes; C. L. Mallows, (1983), A Method for Comparing Two Hierarchical Clusterings, Journal of the American Statistical Association, 78(383), 553-569.
- William M. Rand, (1971), Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association 66(336), 846-850.

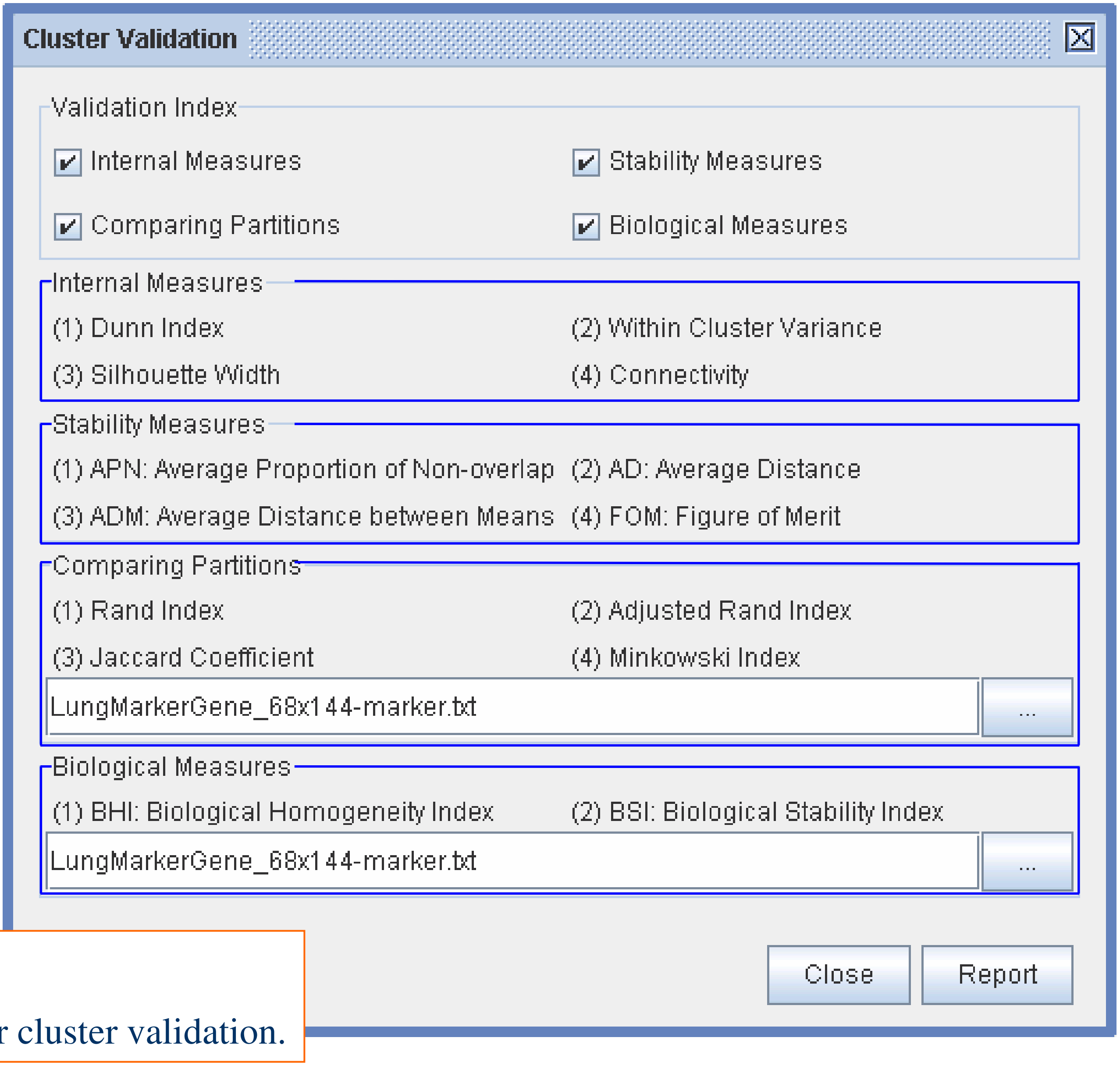
Cluster Validation Index

Internal Measures

Stability Measures

Comparing Partitions

Biological Measures



Validation Index

☑ Internal Measures ☑ Stability Measures

☑ Comparing Partitions ☑ Biological Measures

Internal Measures

(1) Dunn Index (2) Within Cluster Variance

(3) Silhouette Width (4) Connectivity

Stability Measures

(1) APN: Average Proportion of Non-overlap (2) AD: Average Distance

(3) ADM: Average Distance between Means (4) FOM: Figure of Merit

Comparing Partitions

(1) Rand Index (2) Adjusted Rand Index

(3) Jaccard Coefficient (4) Minkowski Index

LungMarkerGene_68x144-marker.txt

Biological Measures

(1) BHI: Biological Homogeneity Index (2) BSI: Biological Stability Index

LungMarkerGene_68x144-marker.txt

See also

clValid: an R package for cluster validation.

Statistical Evaluation: Internal Measures

Compactness | Homogeneity | Separation

Connectivity

$$Conn(\mathcal{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} d_{i,nn_{i(j)}}$$

$$d_{i,nn_{i(j)}} = \begin{cases} 0, & \text{for } i \text{ and } j \text{ are in the same cluster,} \\ 1/j, & \text{otherwise.} \end{cases}$$

Dunn index (Dunn, 1974)

$$D(\mathcal{C}) = \frac{\min\limits_{C_k, C_l \in \mathcal{C},\, C_k \neq C_l} \left(\min\limits_{i \in C_k,\, j \in C_l} \operatorname{dist}(i,j) \right)}{\max\limits_{C_m \in \mathcal{C}} \operatorname{diam}(C_m)}$$

Within-cluster Variance

$$V(\mathcal{C}) = \sqrt{\frac{1}{N} \sum_{C_k \in \mathcal{C}} \sum_{i \in C_k} \operatorname{dist}(i, \mu_k)}$$

Silhouette Width (Rousseeuw, 1987)

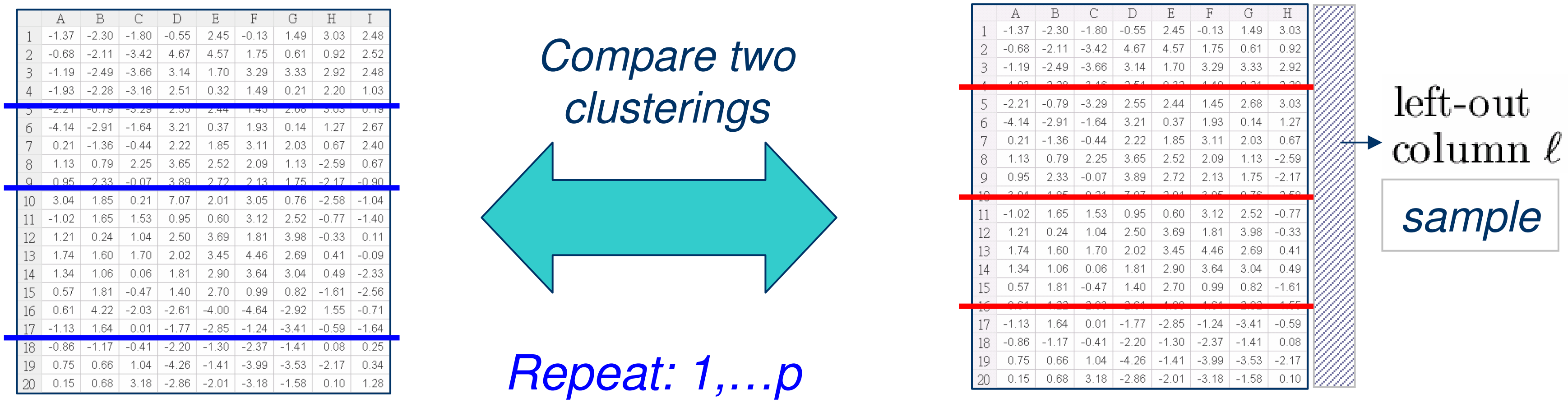
$$S(\mathcal{C}) = \sum_{i=1}^{N} \frac{S(i)}{N}, \quad S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

the average distance between i and the observations in the closet other cluster

the average distance between i and all other observations in the same cluster.

Statistical Evaluation: Stability

- Average Proportion of Non-overlap (APN)
- Average Distance (AD)
- Average Distance between Means (ADM)
- Prediction Strength: Figure of Merit (FOM)



	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28

Full data (nxp)

Compare two clusterings

Repeat: 1,…,p

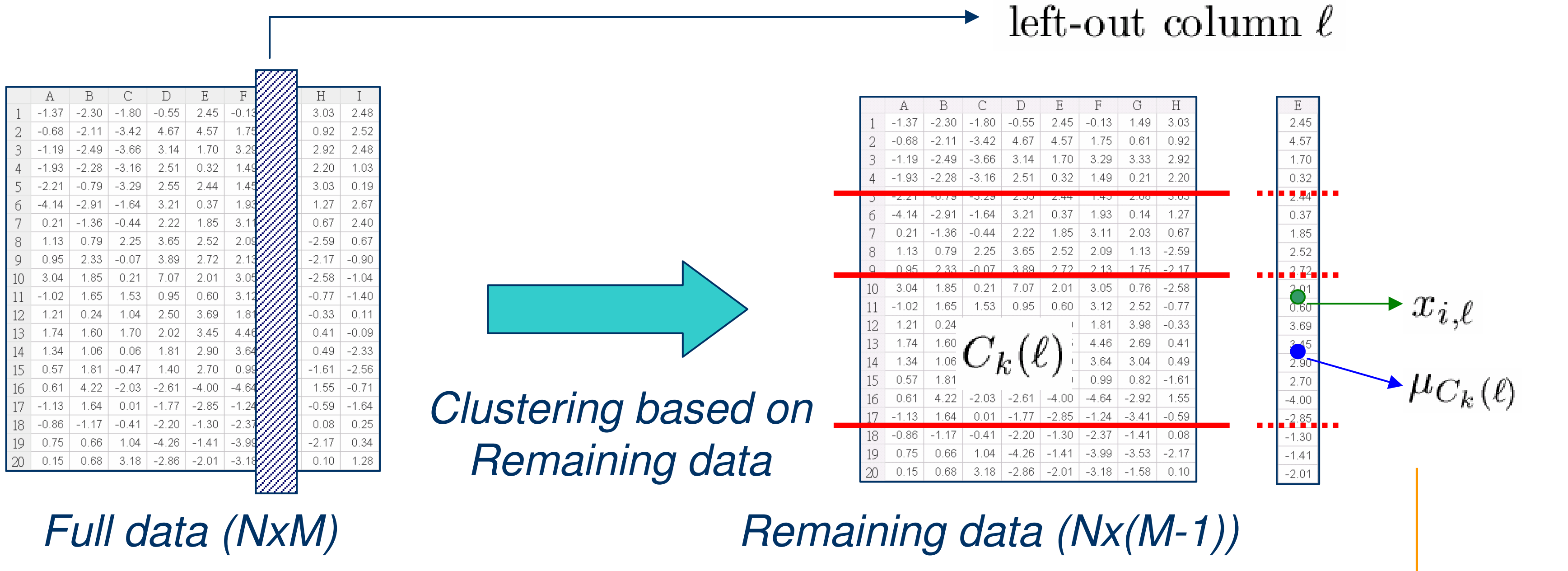
	A	B	C	D	E	F	G	H
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10

left-out column ℓ

sample

Remaining data (nx(p-1))

Figure of Merit (FOM)



Full data (NxM)

	A	B	C	D	E	F		H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13		3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75		0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29		2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49		2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45		3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93		1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11		0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09		-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13		-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05		-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12		-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81		-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46		0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64		0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99		-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64		1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24		-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37		0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99		-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18		0.10	1.28

Clustering based on Remaining data

Remaining data (Nx(M-1))

	A	B	C	D	E	F	G	H
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.08	3.03
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77
12	1.21	0.24				1.81	3.98	-0.33
13	1.74	1.60				4.46	2.69	0.41
14	1.34	1.06				3.64	3.04	0.49
15	0.57	1.81				0.99	0.82	-1.61
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10

E
2.45
4.57
1.70
0.32
2.44
0.37
1.85
2.52
2.72
2.01
0.60
3.69
3.45
2.90
2.70
-4.00
-2.85
-1.30
-1.41
-2.01

$$FOM(\ell,\mathcal{C}) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(\ell)} \operatorname{dist}(x_{i,\ell},\mu_{C_k(\ell)})} \times \sqrt{\frac{N}{N-K}}$$

$$FOM(\mathcal{C}) = \frac{1}{M}\sum_{\ell=1}^{M} FOM(\ell,\mathcal{C})$$

$\operatorname{dist}(x_{i,\ell},\mu_{C_k(\ell)})$

K. Y. Yeung, D. R. Haynor and W. L. Ruzzo, (2001), Validating clustering for gene expression data, Bioinformatics 17(4), 309-318.

Agreement with Reference Partition

- Rand index
- Jaccard coefficient
- Minikowski Measure
- Adjusted Rand index

$$\mathcal{U} = \{U_1, \cdots, U_K\}$$

$$\mathcal{V} = \{V_1, \cdots, V_S\}$$

$$A = \left[\begin{array}{c|cccc} & O_1 & O_1 & \cdots & O_n \\ \hline O_1 & & & & \\ O_2 & & & & \\ \cdots & & & & \\ O_n & & & & \end{array}\right] \qquad B = \left[\begin{array}{c|cccc} & O_1 & O_1 & \cdots & O_n \\ \hline O_1 & & & & \\ O_2 & & & & \\ \cdots & & & & \\ O_n & & & & \end{array}\right]$$

$$A_{ij} = \mathrm{I}(O_i \in U_k, O_j \in U_k) \qquad B_{ij} = \mathrm{I}(O_i \in V_s, O_j \in V_s)$$

$$n_{11} = \#\{(O_i, O_j); \mathrm{I}(A_{ij} = 1, B_{ij} = 1)\}$$
$$n_{10} = \#\{(O_i, O_j); \mathrm{I}(A_{ij} = 1, B_{ij} = 0)\}$$
$$n_{01} = \#\{(O_i, O_j); \mathrm{I}(A_{ij} = 0, B_{ij} = 1)\}$$
$$n_{00} = \#\{(O_i, O_j); \mathrm{I}(A_{ij} = 0, B_{ij} = 0)\}$$

- Rand index, [0, 1], maximum:

$$R(\mathcal{U}, \mathcal{V}) = \frac{n_{11} + n_{00}}{n_{00} + n_{01} + n_{10} + n_{11}}$$

- Jaccard coefficient, [0, 1], maximum:

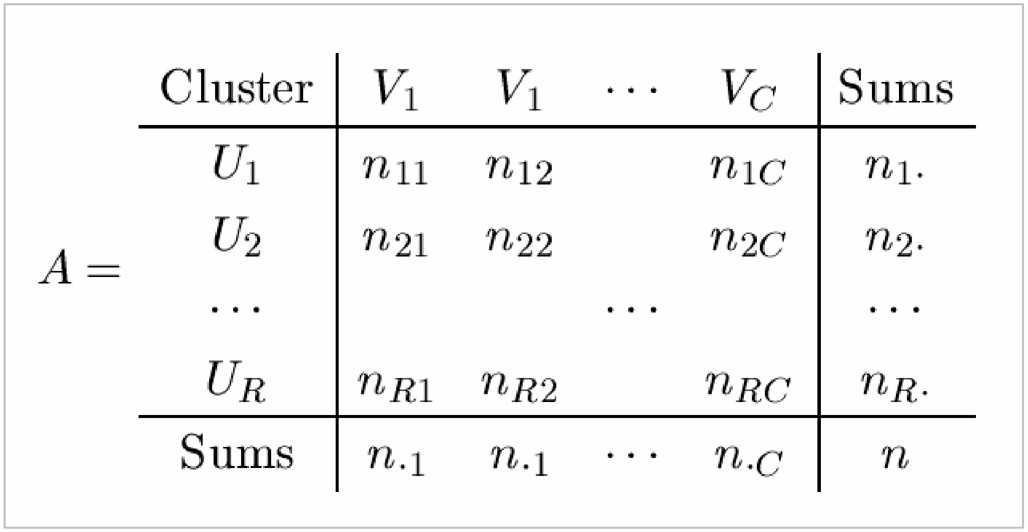
$$J(\mathcal{U}, \mathcal{V}) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

- Minikowski measure:

$$M(\mathcal{U}, \mathcal{V}) = \sqrt{\frac{n_{10} + n_{01}}{n_{11} + n_{01}}}$$

Adjusted Rand index

May be the most widely used Cluster Validation Index!



$A =$

Cluster	V_1	V_1	$\cdots$	V_C	Sums
U_1	n_{11}	n_{12}		n_{1C}	$n_{1\cdot}$
U_2	n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
$\cdots$			$\cdots$		$\cdots$
U_R	n_{R1}	n_{R2}		n_{RC}	$n_{R\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot C}$	n

$$R(\mathcal{U},\mathcal{V}) = 1 + \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right)}{\binom{n}{2}}$$

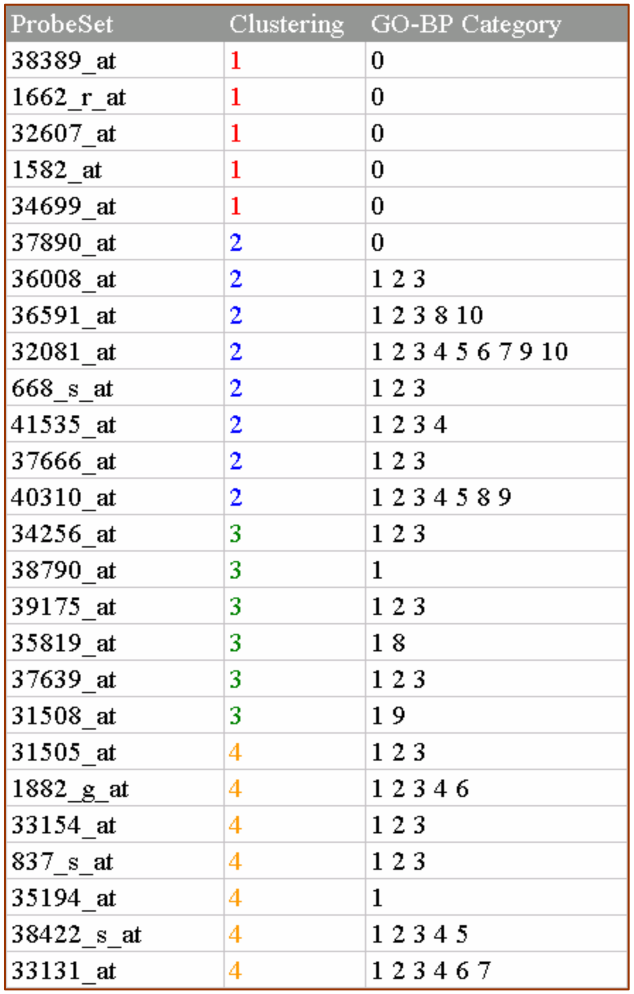
$$R(\mathcal{U},\mathcal{V})_{adj} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{C}\binom{n_{\cdot j}}{2}\right] - \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}/\binom{n}{2}}$$

Lawrence Hubert and Phipps Arabie (1985), Comparing partitions, Journal of Classification 2(1), 193-218.

Biological Evaluation

- Biological Homogeneity Index (BHI)
- Biological Stability Index (BSI)

Example:
GO (Gene Ontology)
Multiple Functional Categories



ProbeSet	Clustering	GO-BP Category
38389_at	1	0
1662_r_at	1	0
32607_at	1	0
1582_at	1	0
34699_at	1	0
37890_at	2	0
36008_at	2	1 2 3
36591_at	2	1 2 3 8 10
32081_at	2	1 2 3 4 5 6 7 9 10
668_s_at	2	1 2 3
41535_at	2	1 2 3 4
37666_at	2	1 2 3
40310_at	2	1 2 3 4 5 8 9
34256_at	3	1 2 3
38790_at	3	1
39175_at	3	1 2 3
35819_at	3	1 8
37639_at	3	1 2 3
31508_at	3	1 9
31505_at	4	1 2 3
1882_g_at	4	1 2 3 4 6
33154_at	4	1 2 3
837_s_at	4	1 2 3
35194_at	4	1
38422_s_at	4	1 2 3 4 5
33131_at	4	1 2 3 4 6 7

Susmita Datta and Somnath Datta, (2006), Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinformatics 7:397.

Biological Evaluation: Homogeneity

Biological Homogeneity Index (BHI)

- $\mathcal{B} = \{B_1, \ldots, B_F\}$: a set of F functional classes, not necessarily disjoint,
- B^i: the functional class containing gene i (with possibly more than one functional class containing i).
- B^j: the function class containing gene j,
- $\mathrm{I}(B^i = B^j) = \begin{cases} 1, & \text{if } B^i \text{ and } B^j \text{ match }, \\ 0, & \text{otherwise.} \end{cases}$
- Given statistical clustering partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ and set of biological classes $\mathcal{B} = \{B_1, \ldots, B_F\}$, the BHI is defined as

$$BHI(\mathcal{C}, \mathcal{B}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k(n_k - 1)} \sum_{i \neq j; i,j \in C_k} \mathrm{I}\left(B^i = B^j\right) .$$

- $n_k = n(C_k \cap \mathcal{B})$: the number of annotated genes in statistical cluster C_k.
- Range: $[0, 1]$, maximum.

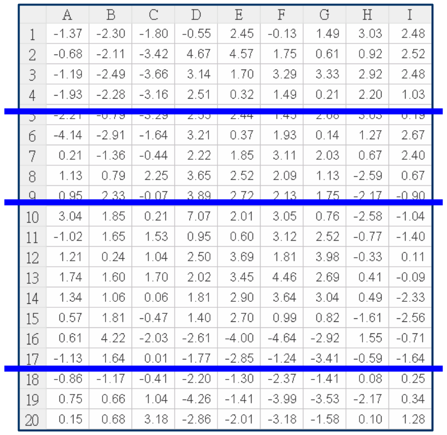
Biological Evaluation: Stability

Biological Stability Index (BSI)

- The BSI is defined as

$$BSI(\mathcal{C},\mathcal{B}) = \frac{1}{F}\sum_{k=1}^{F}\frac{1}{n(B_k)(n(B_k)-1)}\frac{1}{M}\sum_{\ell=1}^{M}\sum_{i\neq j;i,j\in B_k}\frac{n(C^{i,0}\cap C^{j,\ell})}{n(C^{i,0})},$$

- $C^{i,0}$: the statistical cluster containing observation i based on all the data.
- $C^{j,\ell}$: the statistical cluster containing observation j when column ℓ is removed.
- Range $[0,1]$: maximum.

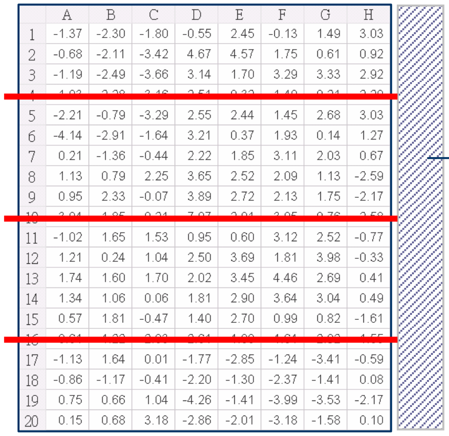


	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28

Full data (nxp)

Compare two clusterings

Repeat: 1,…p



	A	B	C	D	E	F	G	H
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10

left-out column ℓ

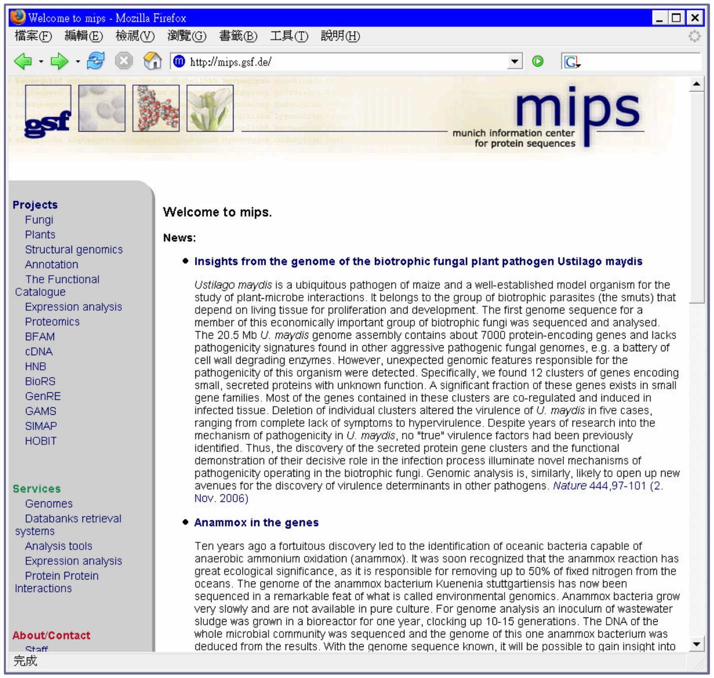
sample

Remaining data (nx(p-1))

Obtain Functional Categories (Annotation)

MIPS: the Munich Information Center for Protein Sequences

- http://mips.gsf.de/
- MIPS: a database for protein sequences and complete genomes, Nucleic Acids Research, 27:44-48, 1999


Welcome to mips - Mozilla Firefox
檔案(F) 編輯(E) 檢視(V) 瀏覽(G) 書籤(B) 工具(T) 說明(H)
http://mips.gsf.de/
gsf
mips
munich information center
for protein sequences
完成

Projects
Fungi
Plants
Structural genomics
Annotation
The Functional Catalogue
Expression analysis
Proteomics
BFAM
cDNA
HNB
BioRS
GenRE
GAMS
SIMAP
HOBIT

Services
Genomes
Databanks retrieval systems
Analysis tools
Expression analysis
Protein Protein Interactions

About/Contact
Staff

Welcome to mips.

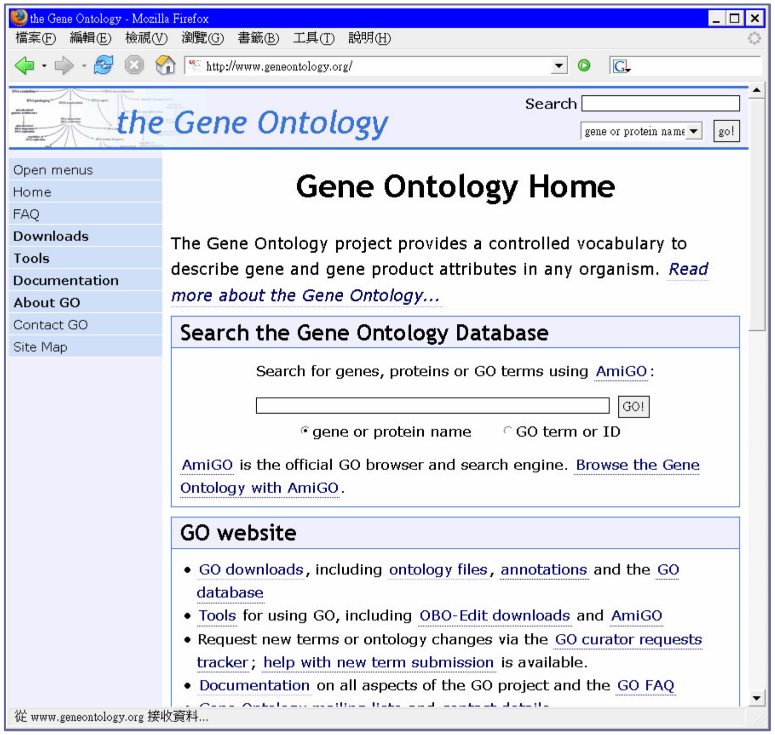
News:

- **Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis**

  Ustilago maydis is a ubiquitous pathogen of maize and a well-established model organism for the study of plant-microbe interactions. It belongs to the group of biotrophic parasites (the smuts) that depend on living tissue for proliferation and development. The first genome sequence for a member of this economically important group of biotrophic fungi was sequenced and analysed. The 20.5 Mb *U. maydis* genome assembly contains about 7000 protein-encoding genes and lacks pathogenicity signatures found in other aggressive pathogenic fungal genomes, e.g. a battery of cell wall degrading enzymes. However, unexpected genomic features responsible for the pathogenicity of this organism were detected. Specifically, we found 12 clusters of genes encoding small, secreted proteins with unknown function. A significant fraction of these genes exists in small gene families. Most of the genes contained in these clusters are co-regulated and induced in infected tissue. Deletion of individual clusters altered the virulence of *U. maydis* in five cases, ranging from complete lack of symptoms to hypervirulence. Despite years of research into the mechanism of pathogenicity in *U. maydis*, no "true" virulence factors had been previously identified. Thus, the discovery of the secreted protein gene clusters and the functional demonstration of their decisive role in the infection process illuminate novel mechanisms of pathogenicity operating in the biotrophic fungi. Genomic analysis is, similarly, likely to open up new avenues for the discovery of virulence determinants in other pathogens. *Nature* 444,97-101 (2. Nov. 2006)

- **Anammox in the genes**

  Ten years ago a fortuitous discovery led to the identification of oceanic bacteria capable of anaerobic ammonium oxidation (anammox). It was soon recognized that the anammox reaction has great ecological significance, as it is responsible for removing up to 50% of fixed nitrogen from the oceans. The genome of the anammox bacterium Kuenenia stuttgartiensis has now been sequenced in a remarkable feat of what is called environmental genomics. Anammox bacteria grow very slowly and are not available in pure culture. For genome analysis an inoculum of wastewater sludge was grown in a bioreactor for one year, clocking up 10-15 generations. The DNA of the whole microbial community was sequenced and the genome of this one anammox bacterium was deduced from the results. With the genome sequence known, it will be possible to gain insight into


the Gene Ontology - Mozilla Firefox
檔案(F) 編輯(E) 檢視(V) 瀏覽(G) 書籤(B) 工具(T) 說明(H)
http://www.geneontology.org/
the Gene Ontology
Search
gene or protein name
go!
GO!
從 www.geneontology.org 接收資料...

Open menus
Home
FAQ
Downloads
Tools
Documentation
About GO
Contact GO
Site Map

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. *Read more about the Gene Ontology...*

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

● gene or protein name ○ GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

GO website

- GO downloads, including ontology files, annotations and the GO database
- Tools for using GO, including OBO-Edit downloads and AmiGO
- Request new terms or ontology changes via the GO curator requests tracker; help with new term submission is available.
- Documentation on all aspects of the GO project and the GO FAQ
- Gene Ontology mailing lists and contact details

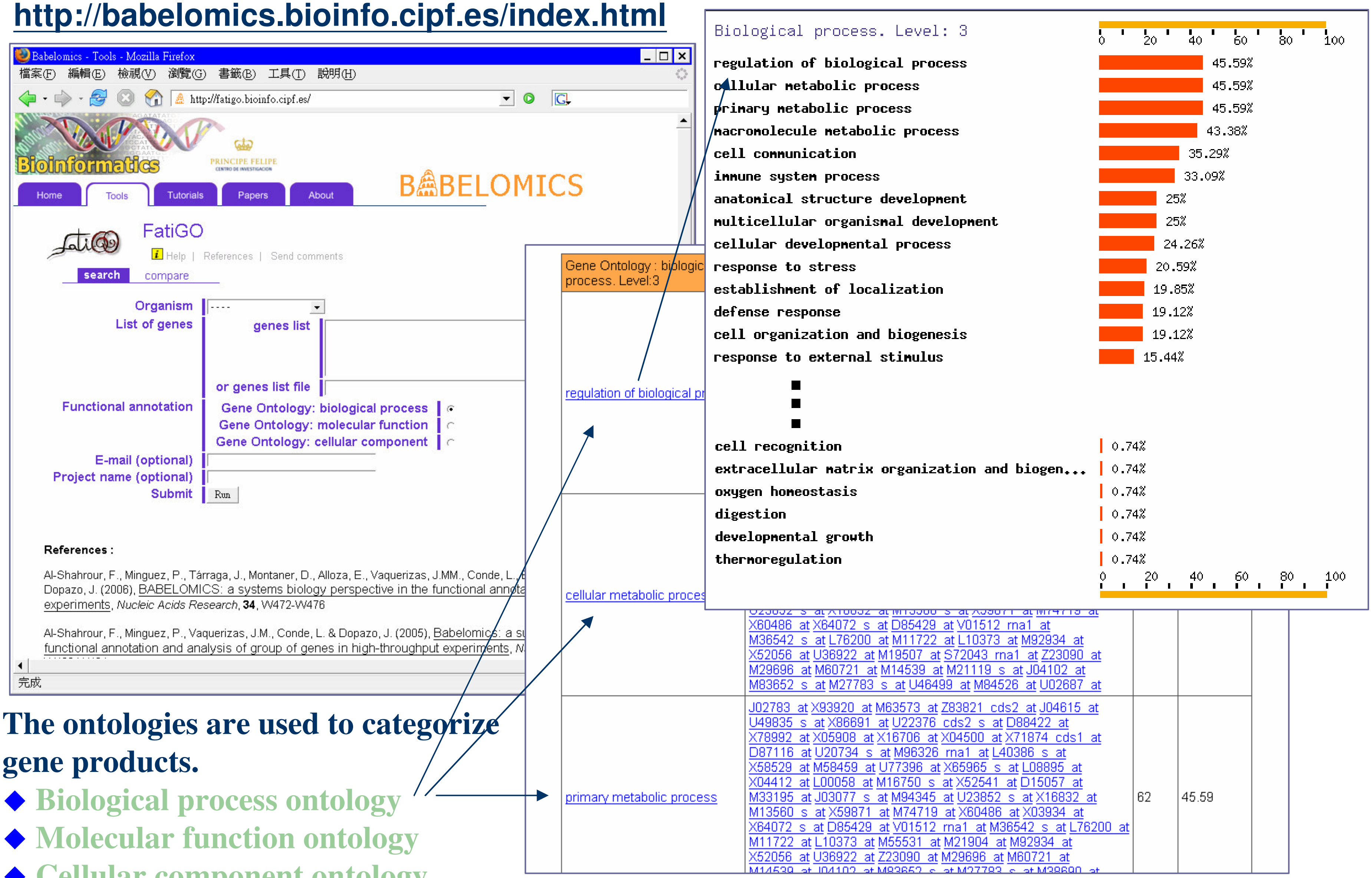
GO: Gene Ontology

- A GO annotation is a Gene Ontology term associated with a gene product.
- http://www.geneontology.org/
- The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. (2000) 25: 25-29.

- FatiGO (Al-Shahrour et al., 2004)
- FunCat (Ruepp et al., 2004)

FatiGO

http://babelomics.bioinfo.cipf.es/index.html



Babelomics - Tools - Mozilla Firefox

http://fatigo.bioinfo.cipf.es/

Organism: ----

List of genes: genes list / or genes list file

Functional annotation:
- Gene Ontology: biological process ●
- Gene Ontology: molecular function ○
- Gene Ontology: cellular component ○

E-mail (optional)

Project name (optional)

Submit: Run

References :

Al-Shahrour, F., Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J.MM., Conde, L., Dopazo, J. (2006), BABELOMICS: a systems biology perspective in the functional annota experiments, *Nucleic Acids Research*, **34**, W472-W476

Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. (2005), Babelomics: a su functional annotation and analysis of group of genes in high-throughput experiments, N

Gene Ontology : biologic process. Level:3			
regulation of biological pr			
cellular metabolic proces	U23852 s at X16832 at M13560 s at X59871 at M74719 at X60486 at X64072 s at D85429 at V01512 rna1 at M36542 s at L76200 at M11722 at L10373 at M92934 at X52056 at U36922 at M19507 at S72043 rna1 at Z23090 at M29696 at M60721 at M14539 at M21119 s at J04102 at M83652 s at M27783 s at U46499 at M84526 at U02687 at		
primary metabolic process	J02783 at X93920 at M63573 at Z83821 cds2 at J04615 at U49835 s at X86691 at U22376 cds2 s at D88422 at X78992 at X05908 at X16706 at X04500 at X71874 cds1 at D87116 at U20734 s at M96326 rna1 at L40386 s at X58529 at M58459 at U77396 at X65965 s at L08895 at X04412 at L00058 at M16750 s at X52541 at D15057 at M33195 at J03077 s at M94345 at U23852 s at X16832 at M13560 s at X59871 at M74719 at X60486 at X03934 at X64072 s at D85429 at V01512 rna1 at M36542 s at L76200 at M11722 at L10373 at M55531 at M21904 at M92934 at X52056 at U36922 at Z23090 at M29696 at M60721 at M14539 at J04102 at M83652 s at M27783 s at M38690 at	62	45.59

Biological process. Level: 3

regulation of biological process	45.59%
cellular metabolic process	45.59%
primary metabolic process	45.59%
macromolecule metabolic process	43.38%
cell communication	35.29%
immune system process	33.09%
anatomical structure development	25%
multicellular organismal development	25%
cellular developmental process	24.26%
response to stress	20.59%
establishment of localization	19.85%
defense response	19.12%
cell organization and biogenesis	19.12%
response to external stimulus	15.44%
⋮	
cell recognition	0.74%
extracellular matrix organization and biogen...	0.74%
oxygen homeostasis	0.74%
digestion	0.74%
developmental growth	0.74%
thermoregulation	0.74%

The ontologies are used to categorize gene products.

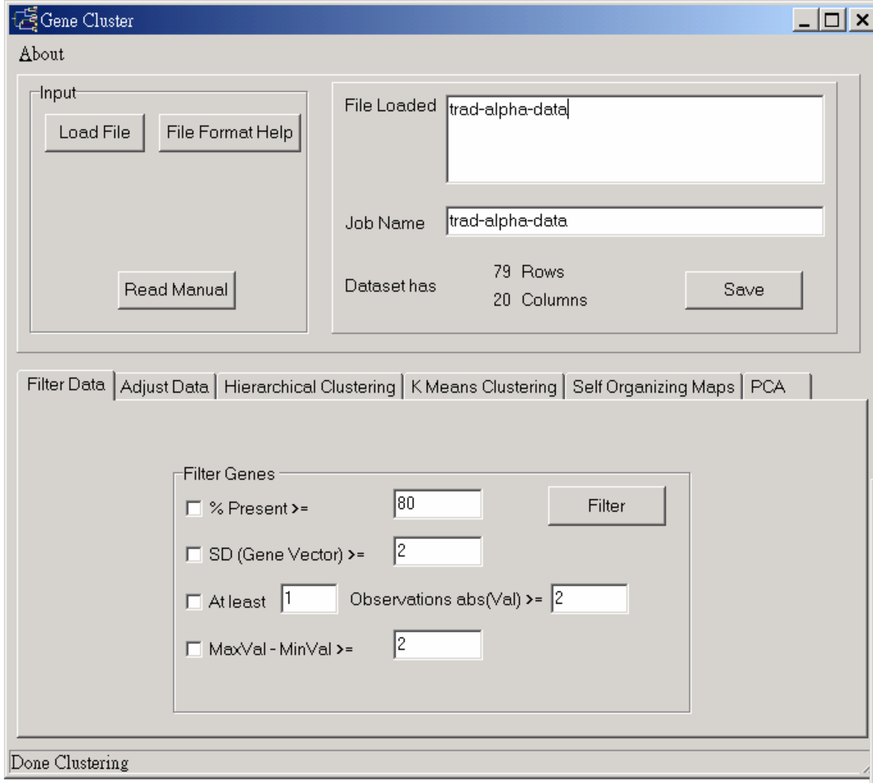
- ◆ Biological process ontology
- ◆ Molecular function ontology
- ◆ Cellular component ontology

Software

- Cluster and TreeView
- Bioconductor
- PermutMatrix
- GAP (Generalized Association Plots)

- GeneSpring GX v7.3

Cluster and TreeView



Gene Cluster

About

Input

Load File | File Format Help

Read Manual

File Loaded: trad-alpha-data

Job Name: trad-alpha-data

Dataset has 79 Rows, 20 Columns

Save

Filter Data | Adjust Data | Hierarchical Clustering | K Means Clustering | Self Organizing Maps | PCA

Filter Genes

- ☐ % Present >= 80
- ☐ SD (Gene Vector) >= 2
- ☐ At least 1 Observations abs(Val) >= 2
- ☐ MaxVal - MinVal >= 2

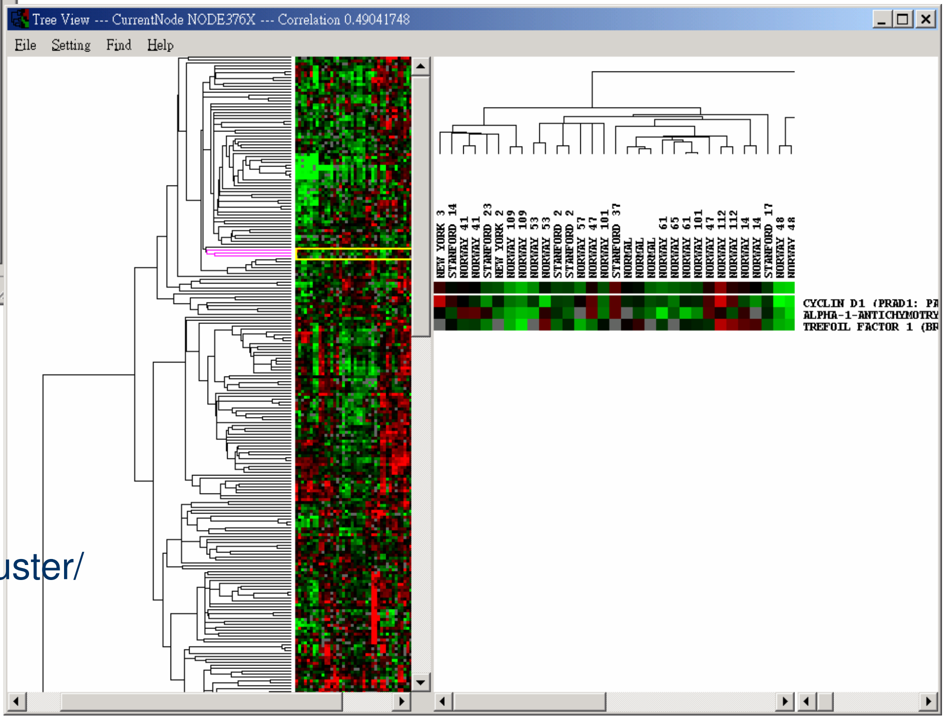
Filter

Done Clustering

Eisen Lab

http://rana.lbl.gov/EisenSoftware.htm

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci.* 95(25):14863-8.

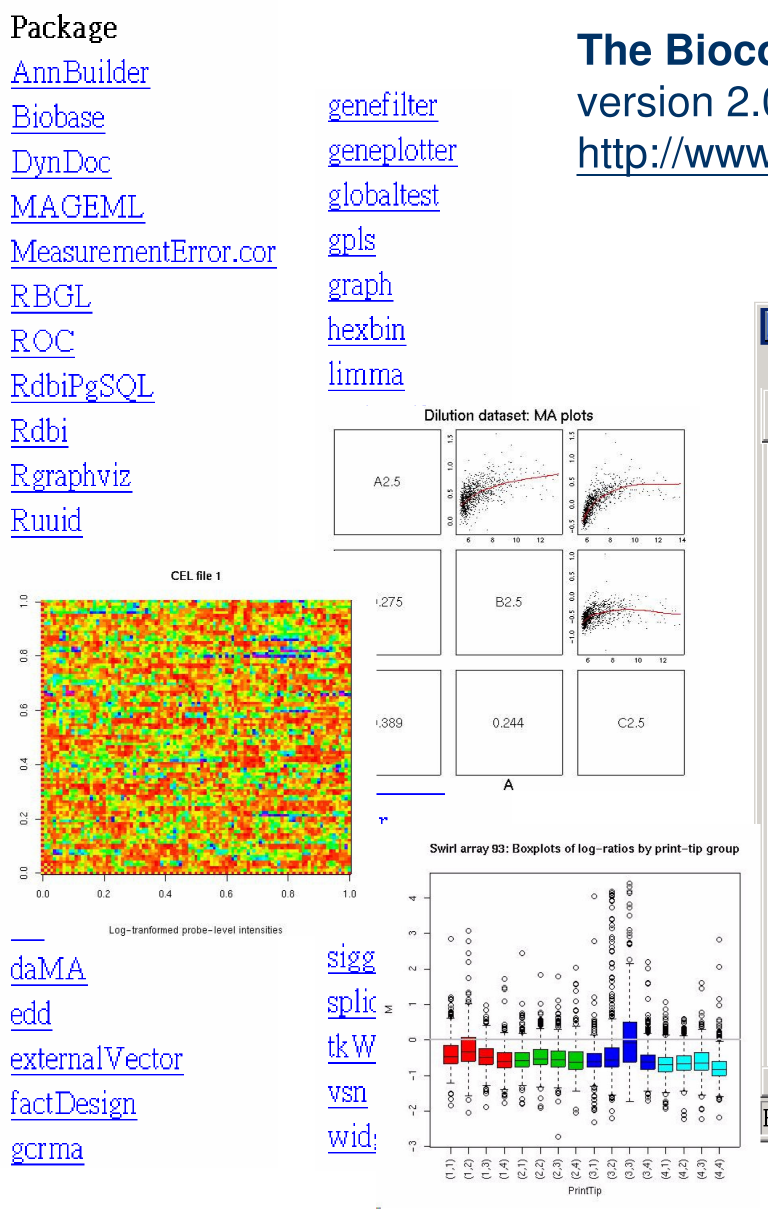


Tree View --- CurrentNode NODE376X --- Correlation 0.49041748

De Hoon, M.J.L.; Imoto, S.; Nolan, J.; Miyano, S.; "**Open source clustering software**". Bioinformatics, 20 (9): 1453--1454 (2004)

http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

Bioconductor



Package

AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Rgraphviz
Ruuid

genefilter
geneplotter
globaltest
gpls
graph
hexbin
limma

daMA
edd
externalVector
factDesign
gcrma

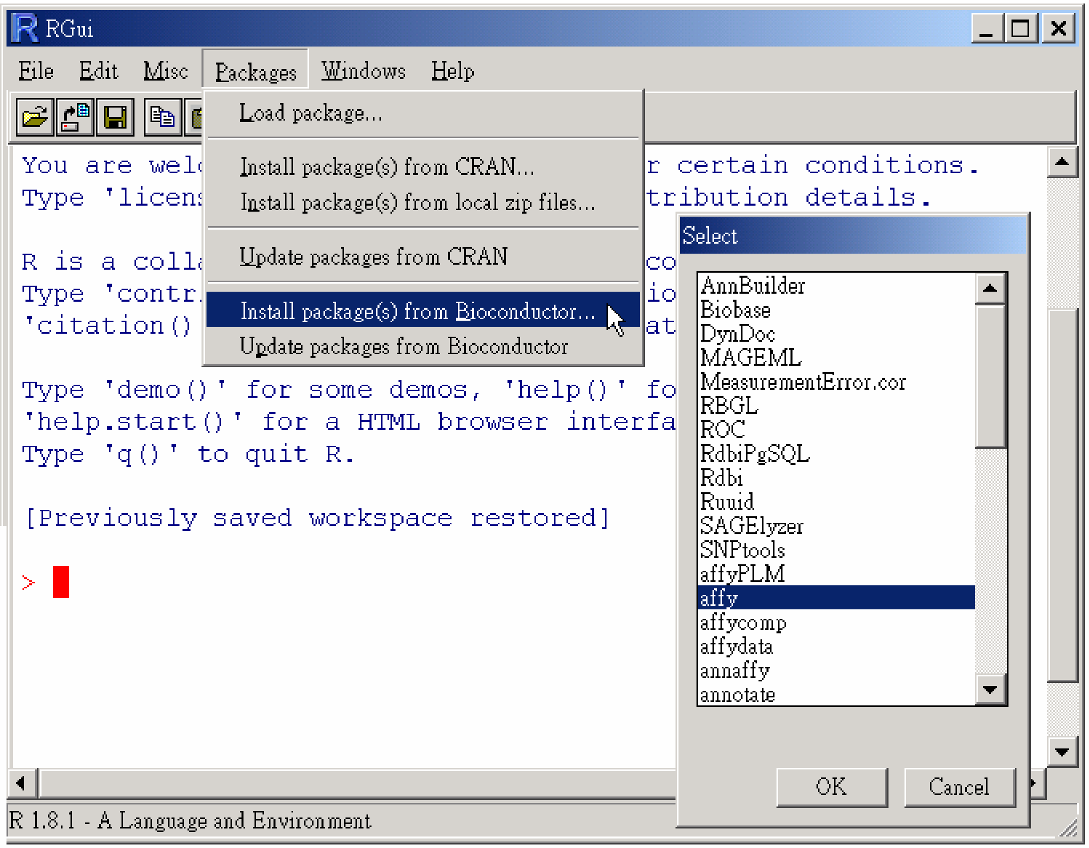
sigg
splic
tkW
vsn
wid

Dilution dataset: MA plots
A2.5
B2.5
C2.5
.275
.389
0.244
A
CEL file 1
Log-tranformed probe-level intensities
Swirl array 93: Boxplots of log-ratios by print-tip group
M
PrintTip

The Bioconductor
version 2.0
http://www.bioconductor.org

The R Project for Statistical Computing

R version 2.5.1 (**2007-06-28**)
http://www.r-project.org



RGui
File
Edit
Misc
Packages
Windows
Help
Load package...
Install package(s) from CRAN...
Install package(s) from local zip files...
Update packages from CRAN
Install package(s) from Bioconductor...
Update packages from Bioconductor
Select
OK
Cancel

```
You are welc                                  r certain conditions.
Type 'licens                                  tribution details.

R is a colla                                  co
Type 'contr                                   io
'citation()                                   at

Type 'demo()' for some demos, 'help()' fo
'help.start()' for a HTML browser interfa
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Ruuid
SAGElyzer
SNPtools
affyPLM
affy
affycomp
affydata
annaffy
annotate

R 1.8.1 - A Language and Environment

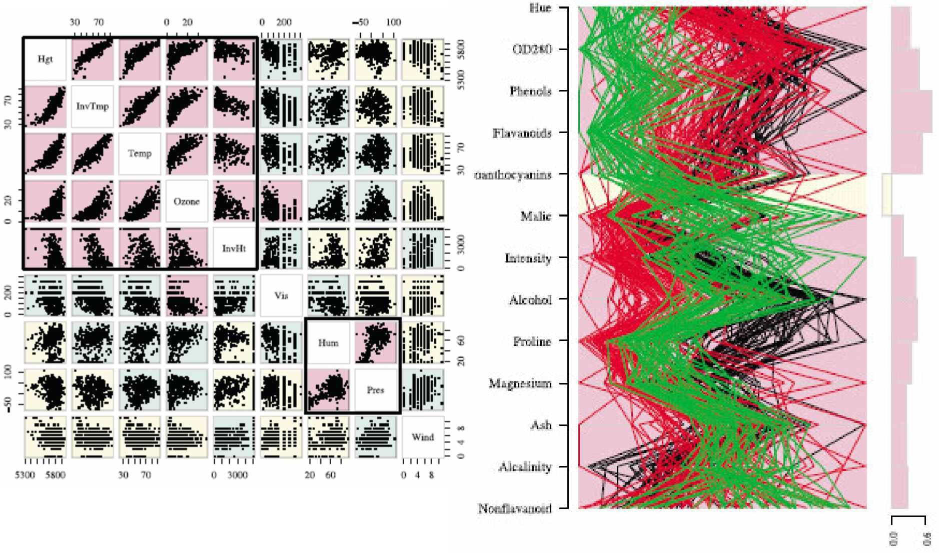
Gclus, PermutMatrix

- **gclus: Clustering Graphics**

  (R package)

  http://cran.r-project.org/src/contrib/Descriptions/gclus.html

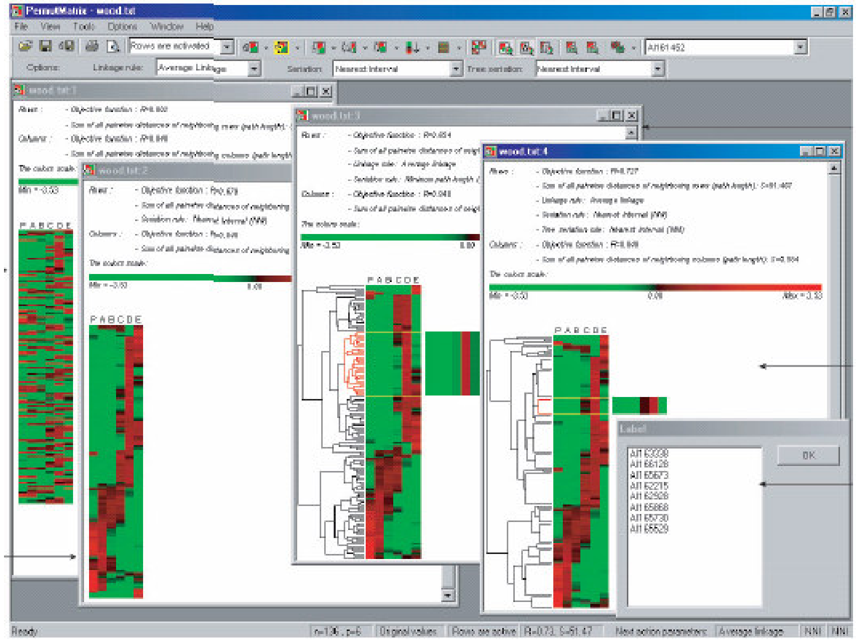
  Catherine B. Hurley, (2004), Clustering Visualizations of Multidimensional Data, Journal of Computational & Graphical Statistics, Vol. 13, No. 4, pp.788-806



- **PermutMatrix**

  http://www.lirmm.fr/~caraux/PermutMatrix

  Caraux, G., and Pinloche, S. (2005), "Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order," Bioinformatics, 21, 1280-1281.



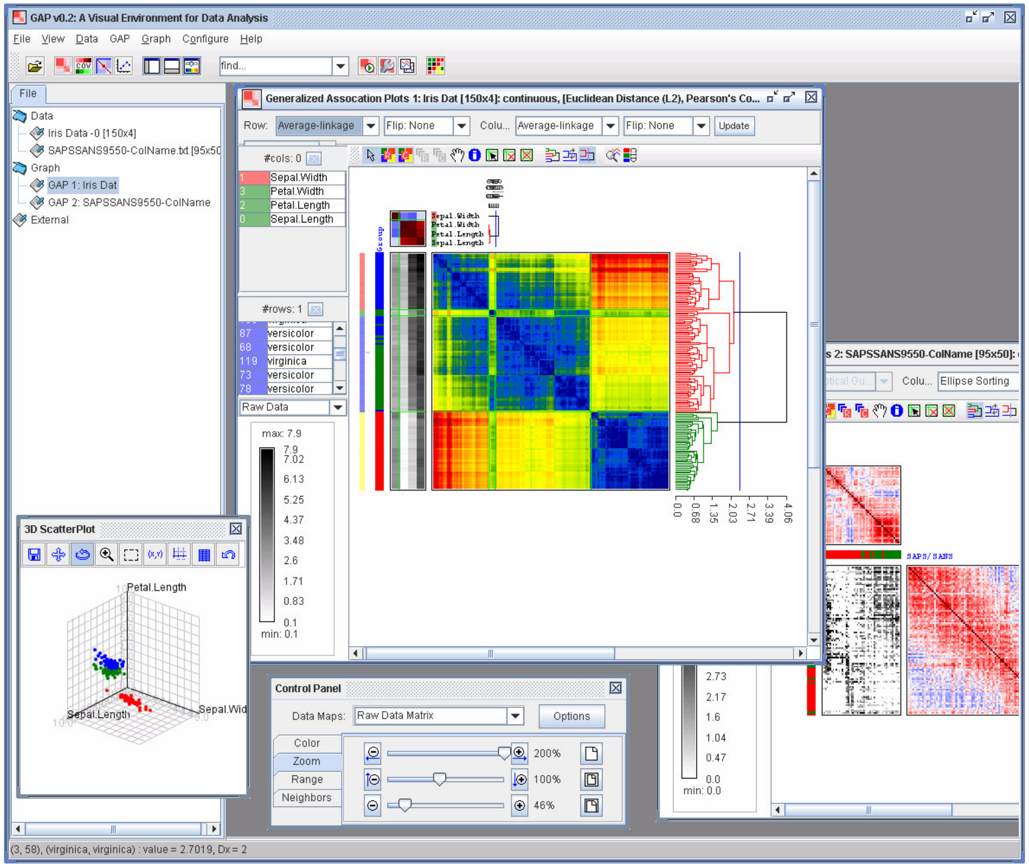
GAP Software verison 0.2

Generalized Association Plots

- Input Data Type: continuous or binary.
- Various seriation algorithms and **clustering analysis**.
- Various display conditions.
- Modules:
  GAP with Covaraite Adjusted,
  Nonlinear Association Analysis,
  Missing Value Imputation.

Statistical Plots

- 2D Scatterplot, 3D Scatterplot (Rotatable)



Chen, C. H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. Statistica Sinica 12, 7-29.
Wu, H. M., Tien, Y. J. and Chen, C. H. (2006). GAP: a Graphical Environment for Matrix Visualization and Information Mining.

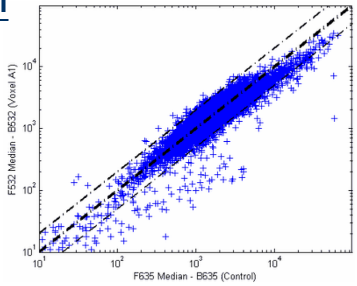
http://gap.stat.sinica.edu.tw/Software/GAP

Matlab: Bioinformatics ToolBox

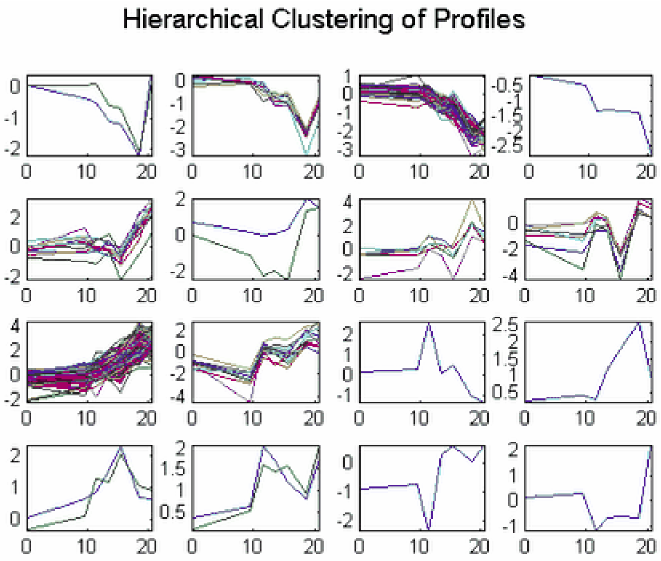
The MathWorks

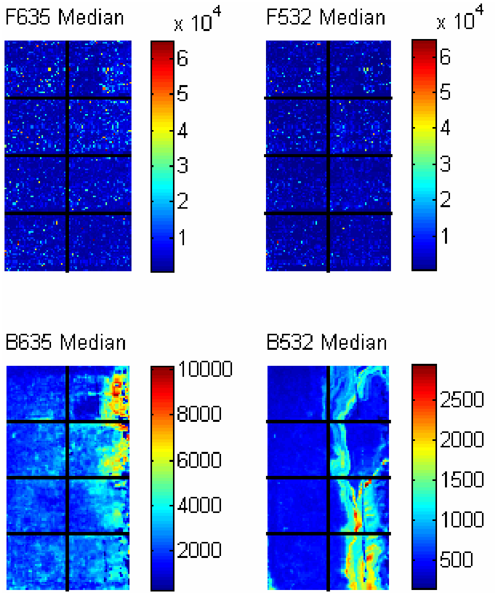
Bioinformatics Toolbox

http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html



- Data Formats and Databases — Access online databases, read and write to files with standard genome and proteome formats such as FASTA and PDB.
- Sequence Alignments — Compare nucleotide or amino acid sequences using pairwise and multiple sequence alignment functions.
- Sequence Utilities and Statistics — Manipulate sequences and determine physical, chemical, and biological characteristics.
- Microarray Analysis — Read, filter, normalize, and visualize microarray data.
- Protein Structure Analysis — Determine protein characteristics and simulate enzyme cleavage reactions.

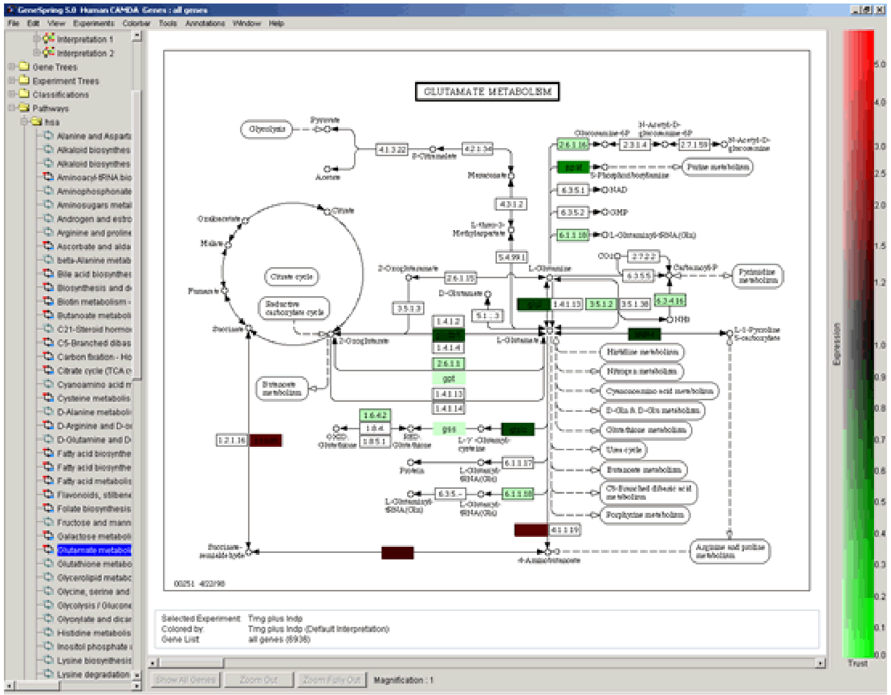




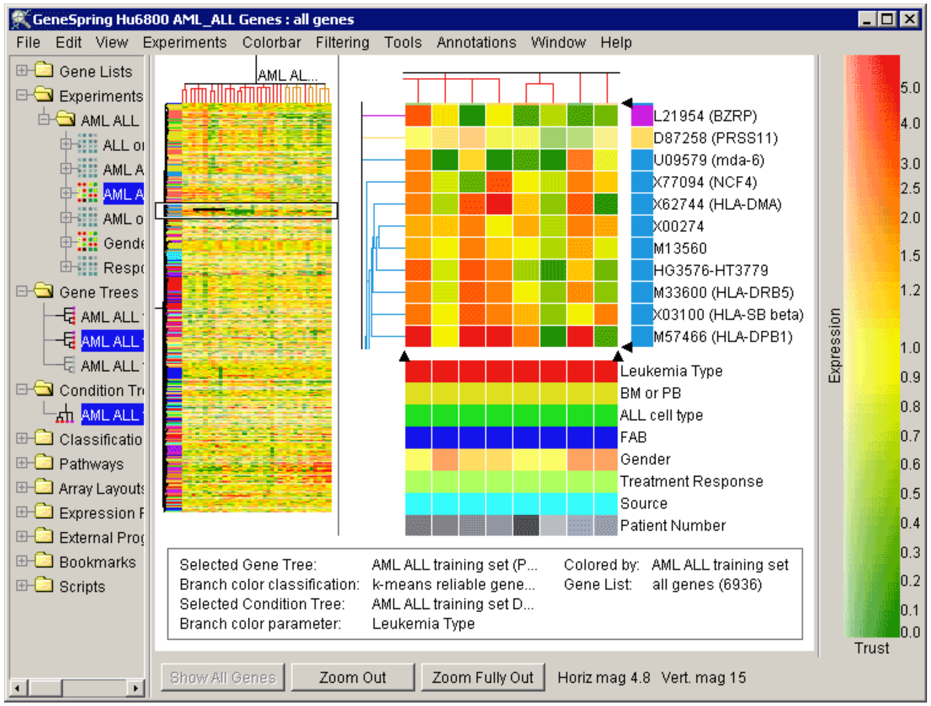
- Prototype and Development Environment — Create new algorithms, try new ideas, and compare alternatives.
- Share Algorithms and Deploy Applications — Create GUIs and stand-alone applications.

GeneSpring GX v7.3.1

- RMA or GC-RMA probe level analysis
- Advanced Statistical Tools
- Data Clustering
- Visual Filtering
- 3D Data Visualization
- Data Normalization (Sixteen)
- Pathway Views
- Search for Similar Samples
- Support for MIAME Compliance
- Scripting
- MAGE-ML Export



Selected Experiment: Tmg plus Indp
Colored by: Tmg plus Indp (Default Interpretation)
Gene List: all genes (6936)



Selected Gene Tree: AML ALL training set (P...
Branch color classification: k-means reliable gene...
Selected Condition Tree: AML ALL training set D...
Branch color parameter: Leukemia Type
Colored by: AML ALL training set
Gene List: all genes (6936)

Horiz mag 4.8 Vert. mag 15

Images from
http://www.silicongenetics.com

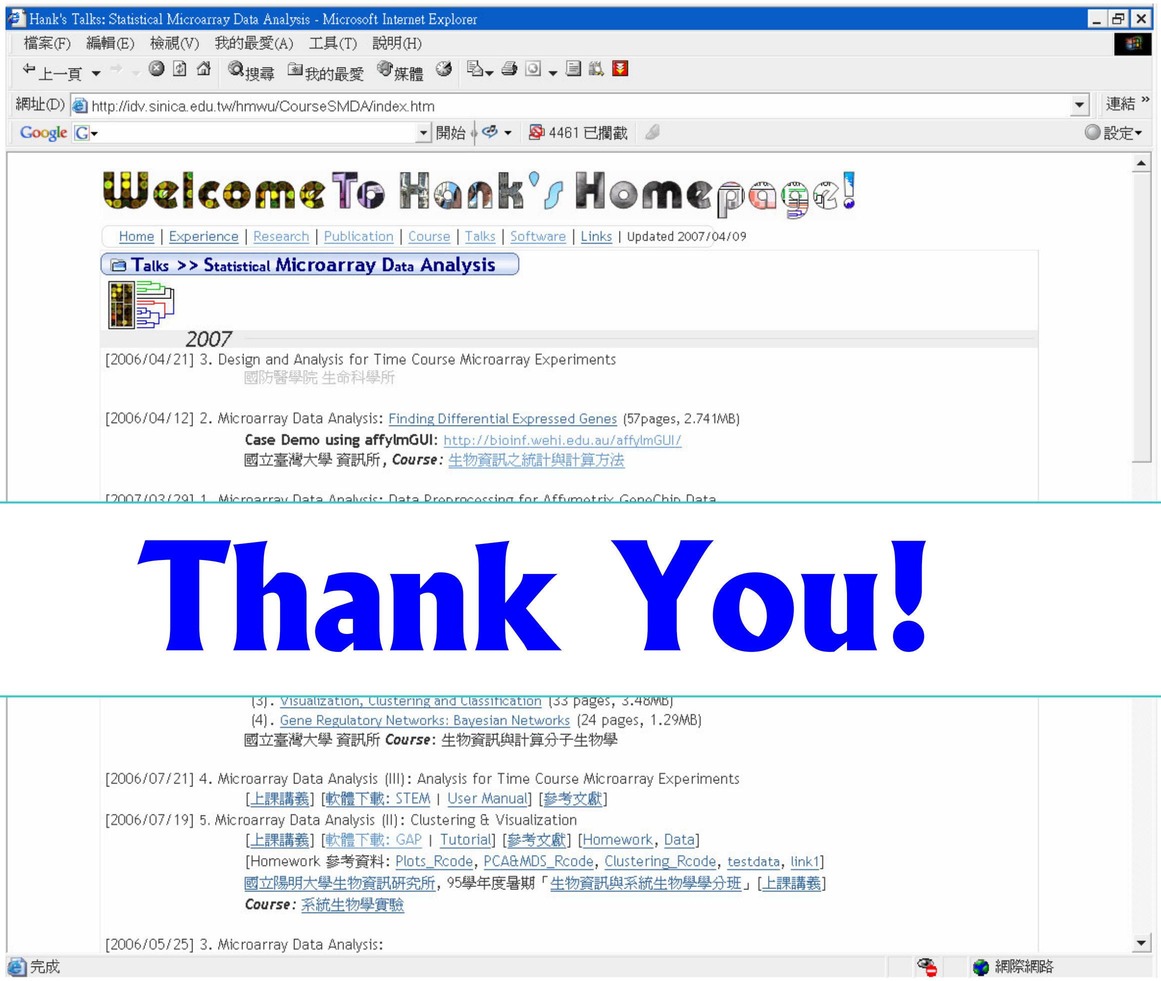
Agilent Technologies

2004 Articles Citing GeneSpring®

2004 : 2003 : 2002 : 2001 : pre-2001 : Reviews

More than 700 papers

Questions?


Hank's Talks: Statistical Microarray Data Analysis - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
http://idv.sinica.edu.tw/hmwu/CourseSMDA/index.htm
連結
Google
開始
4461 已攔截
設定
Welcome To Hank's Homepage!
完成
網際網路

Home | Experience | Research | Publication | Course | Talks | Software | Links | Updated 2007/04/09

Talks >> Statistical Microarray Data Analysis

2007

[2006/04/21] 3. Design and Analysis for Time Course Microarray Experiments
國防醫學院 生命科學所

[2006/04/12] 2. Microarray Data Analysis: Finding Differential Expressed Genes (57pages, 2.741MB)
Case Demo using affylmGUI: http://bioinf.wehi.edu.au/affylmGUI/
國立臺灣大學 資訊所, ***Course***: 生物資訊之統計與計算方法

[2007/03/29] 1. Microarray Data Analysis: Data Preprocessing for Affymetrix GeneChip Data

(3). Visualization, Clustering and Classification (33 pages, 3.48MB)
(4). Gene Regulatory Networks: Bayesian Networks (24 pages, 1.29MB)
國立臺灣大學 資訊所 ***Course***: 生物資訊與計算分子生物學

[2006/07/21] 4. Microarray Data Analysis (III): Analysis for Time Course Microarray Experiments
[上課講義] [軟體下載: STEM | User Manual] [參考文獻]
[2006/07/19] 5. Microarray Data Analysis (II): Clustering & Visualization
[上課講義] [軟體下載: GAP | Tutorial] [參考文獻] [Homework, Data]
[Homework 參考資料: Plots_Rcode, PCA&MDS_Rcode, Clustering_Rcode, testdata, link1]
國立陽明大學生物資訊研究所, 95學年度暑期「生物資訊與系統生物學學分班」[上課講義]
Course: 系統生物學實驗

[2006/05/25] 3. Microarray Data Analysis:

Thank You!

Reference: http://idv.sinica.edu.tw/hmwu/SMDA/Clustering/index.htm

吳漢銘
hmwu@stat.sinica.edu.tw
http://idv.sinica.edu.tw/hmwu

中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica