

Statistical
Microarray Data Analysis

Clustering and Visualization

96 陽明大學生物資訊與系統生物學學分班
Course: 系統生物學實驗

2007年8月15日

吳漢銘

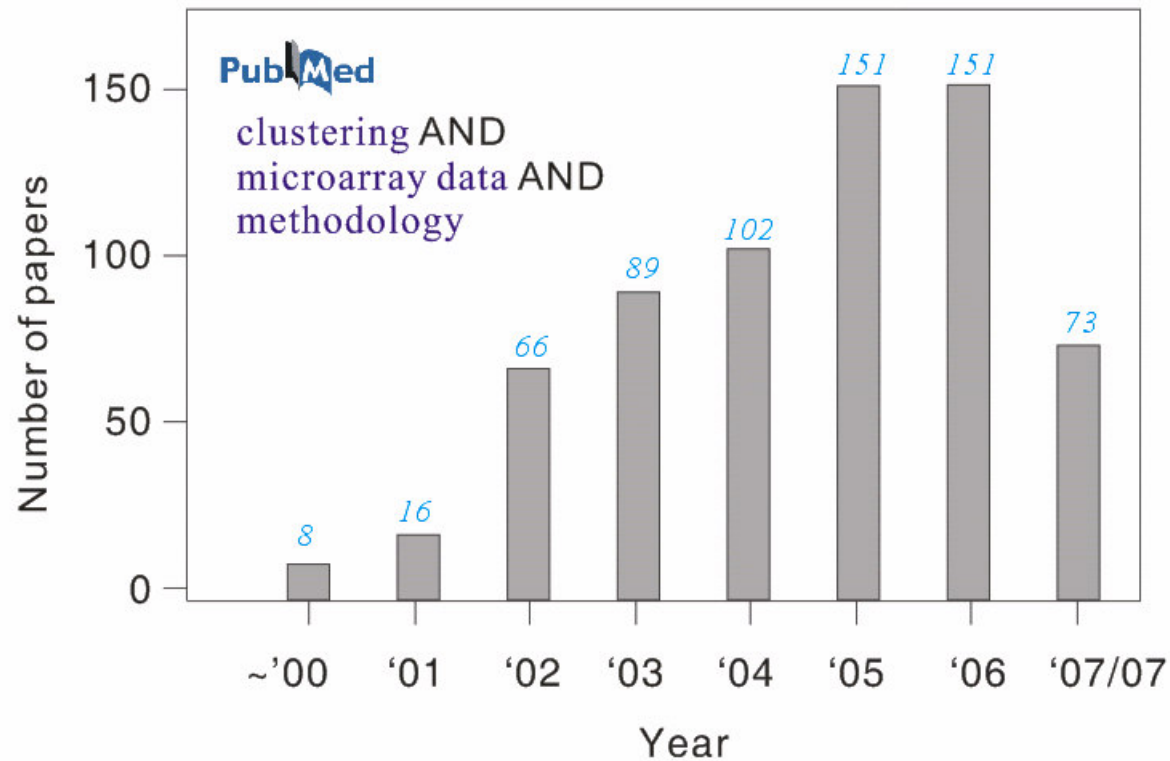
hmwu@stat.sinica.edu.tw
<http://idv.sinica.edu.tw/hmwu>



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

Clustering of Microarray Gene Expression Data

2 /61



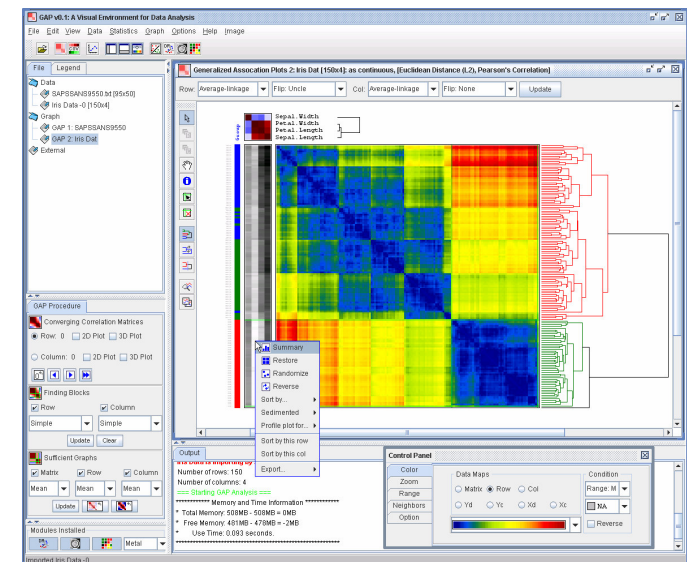
A continuing and active topic of research and application!

Outlines

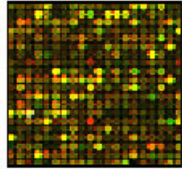
3 / 61

- ◆ **O**verview of Microarray Experiment
- ◆ **C**lustering Analysis and Visualization
- ◆ **D**istance and Similarity Measure
- ◆ **K**-Means Clustering
- ◆ **V**isualizing Clustering Results: Dimension Reduction Techniques
 - ◆ Principal Component Analysis (PCA)
 - ◆ Multidimensional Scaling (MDS)
- ◆ **C**lustering Analysis and Visualization
 - ◆ Self-Organizing Maps (SOM)
 - ◆ Heat Map
 - ◆ Hierarchical Clustering
- ◆ **G**eneralized Association Plots (GAP)
- ◆ **C**luster Validation
- ◆ **S**oftware

GAP: Demo



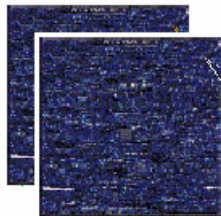
Overview of Microarray Experiment



cDNA Microarray Data

	A	B	C	D
1	UNIQUID	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

$\log_2(\text{Cy5}/\text{Cy3})$

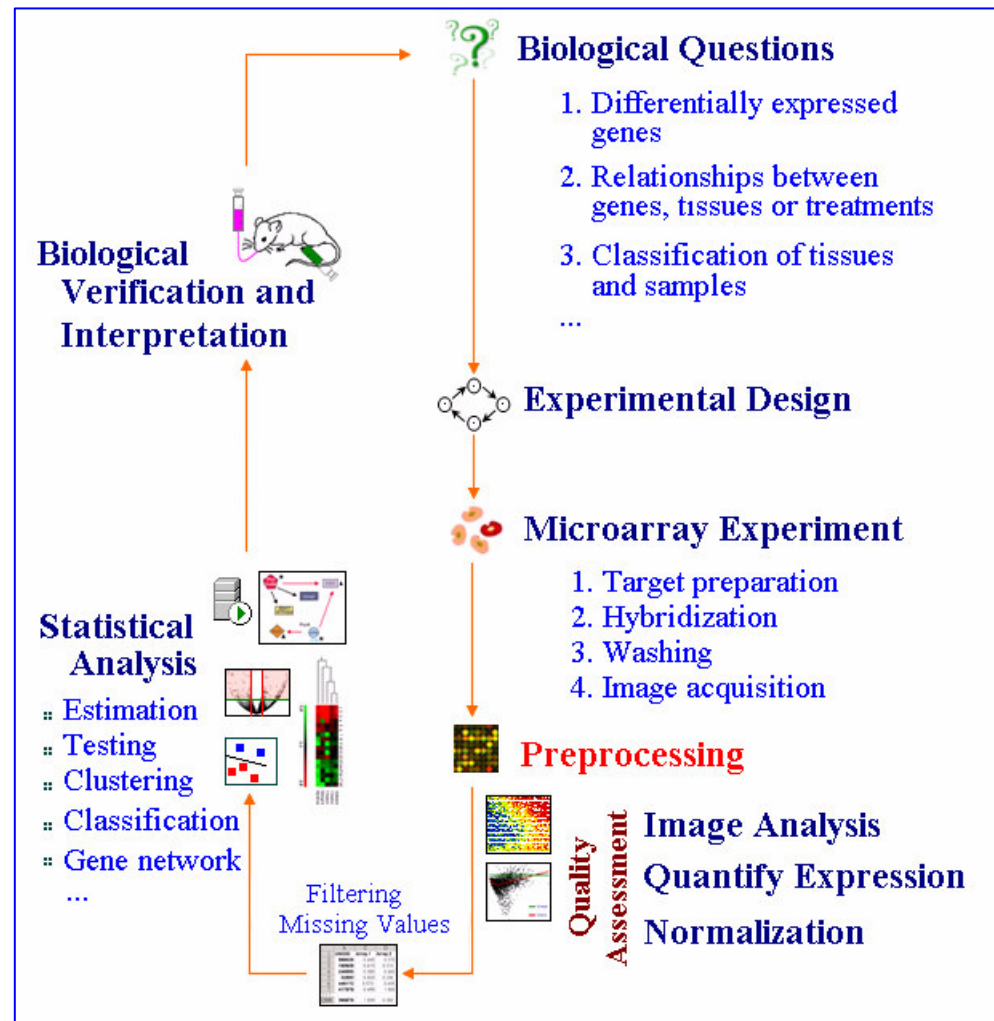


Oligonucleotide Array Data

	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 (Rho)	127.736	99.6895
4	104137_at	ATP-binding cassette, sub-family 4	109.522	65.2727
5	98458_at	baculoviral IAP repeat-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	69.688	54.7391

Expression index

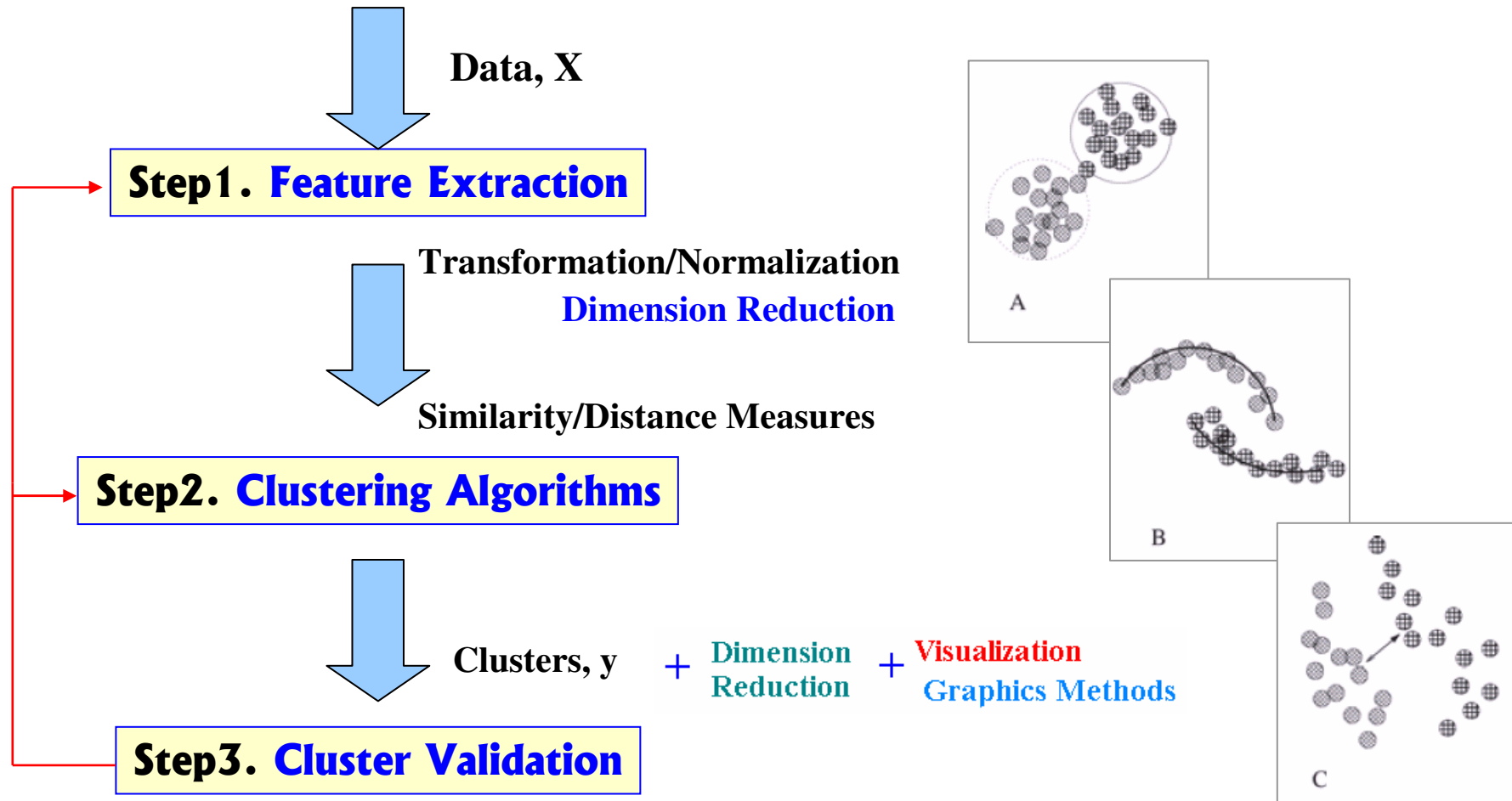
Microarray Life Cycle



Cluster Analysis (Unsupervised Learning)

5 / 61

Group a given collection of **unlabeled** patterns into **meaningful** clusters.



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.

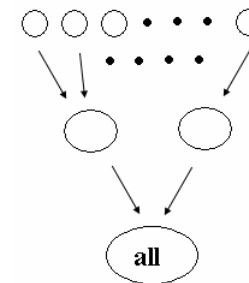
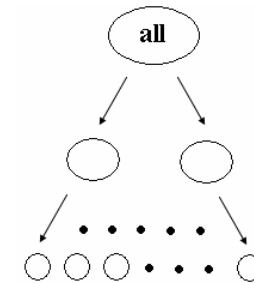
Clustering Analysis

6 / 61

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

7 /61

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.

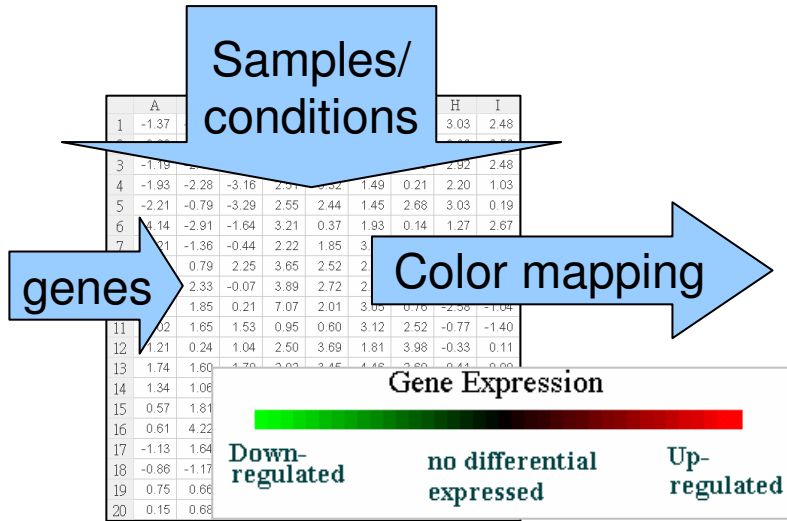
Data/Information Visualization

- ◆ 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.

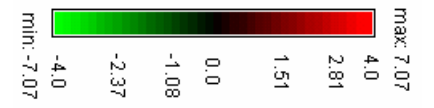
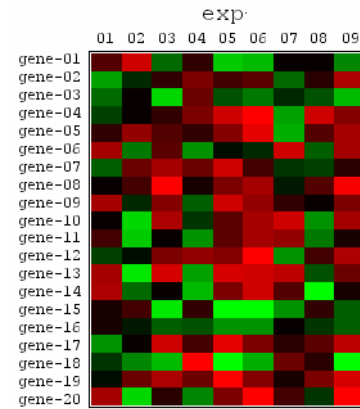
Visualization = Graphing for Data + Fitting + Graphing for Model

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

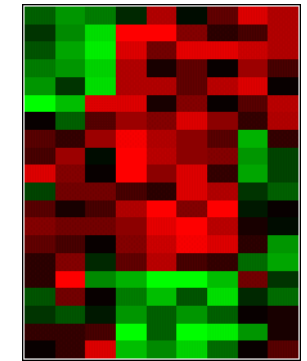
Visualizing Clustering Results: Heat Map



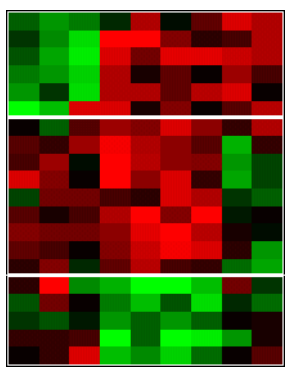
Without ordering



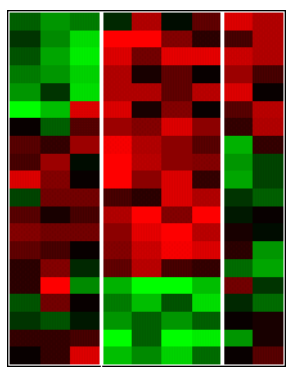
Ordering/
Seriation/
Clustering



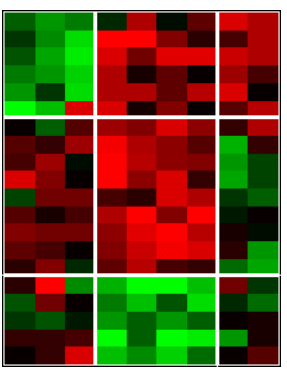
Gene-based
clustering



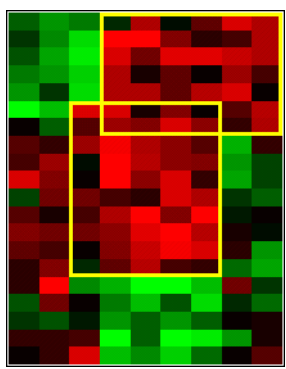
Sample-based
clustering



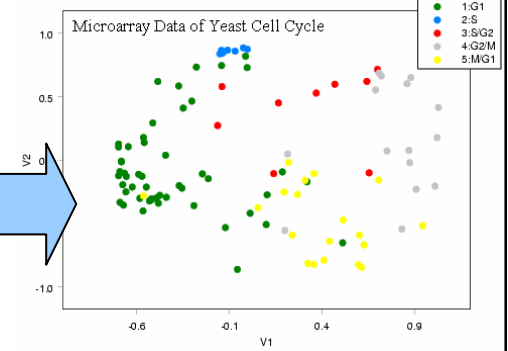
Two-way-based
clustering



Subspace
clustering



Dimension Reduction



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

e.g., Bi-clustering

Distance and Similarity Measure

Cov	x1	x2	x3	x4	x p
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
x p	-0.28	-0.23	-0.05	0.10	1.00

Proximity Matrix

Pearson Correlation Coefficient

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

Data Matrix

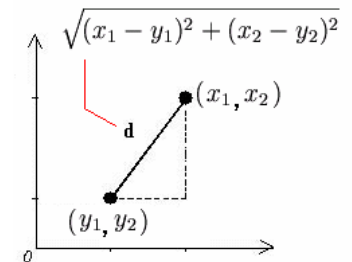
Data	x1	x2	x3	x4	...	x p
subject01	-0.48	-0.42	0.87	0.92	...	-0.18
subject02	-0.39	-0.58	1.08	1.21	...	-0.33
subject03	0.87	0.25	-0.17	0.18	...	-0.44
subject04	1.57	1.03	1.22	0.31	...	-0.49
subject05	-1.15	-0.86	1.21	1.62	...	0.16
subject06	0.04	-0.12	0.31	0.16	...	-0.06
subject07	2.95	0.45	-0.40	-0.66	...	-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

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

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

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

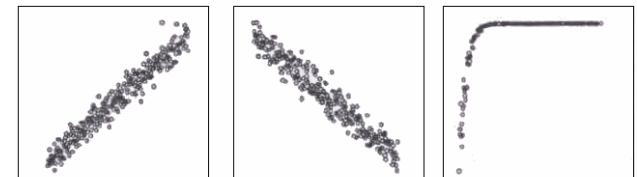
Two pairs of observation (x_i, y_i) and (x_j, y_j)

- 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: $(y_j - y_i) = 0$

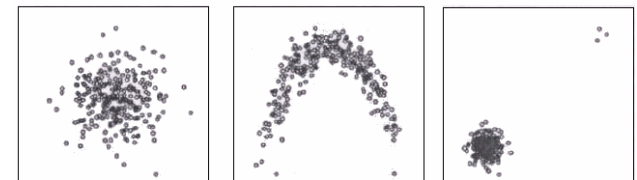
$$\tau = \frac{C - D}{\sqrt{C + D - E_y} \sqrt{C + D - E_x}}$$

- 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.

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



(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

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

K-Means Clustering

12 / 61

- 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

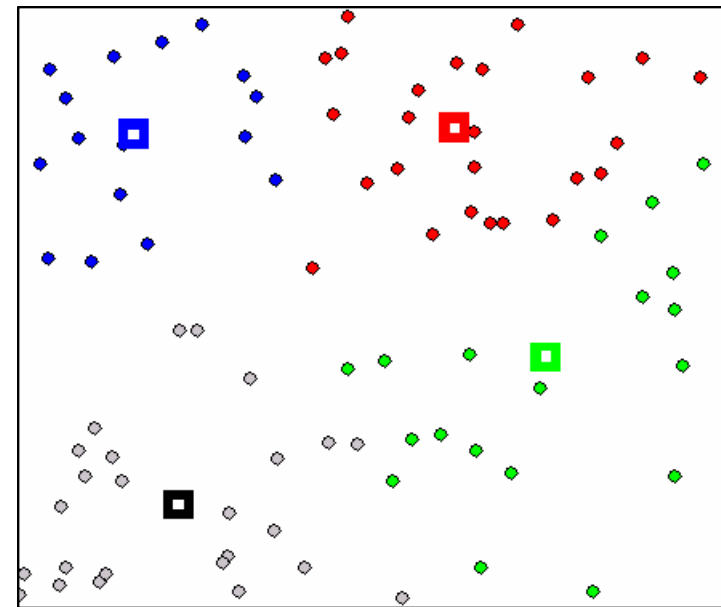
$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

Converged

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.



Visualizing Clustering Results:

Dimension Reduction Techniques

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

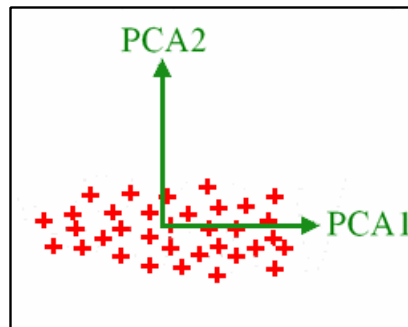
Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.

Principal Component Analysis (PCA)

14 / 61

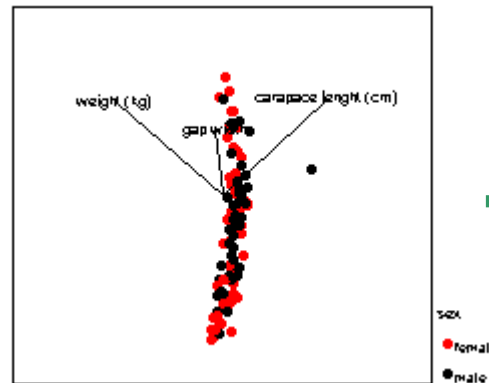
(Pearson 1901; Hotelling 1933; Jolliffe 2002)

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



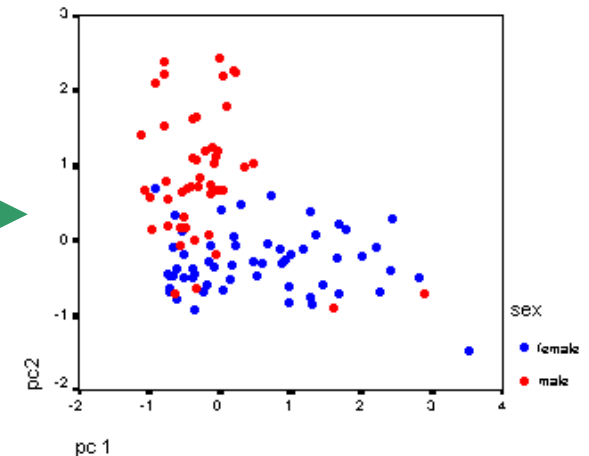
$$PCA_1 = a_1 X + b_1 Y$$

$$PCA_2 = a_2 X + b_2 Y$$



$$PCA_1 = a_1 X + b_1 Y + c_1 Z$$

$$PCA_2 = a_2 X + b_2 Y + c_2 Z$$



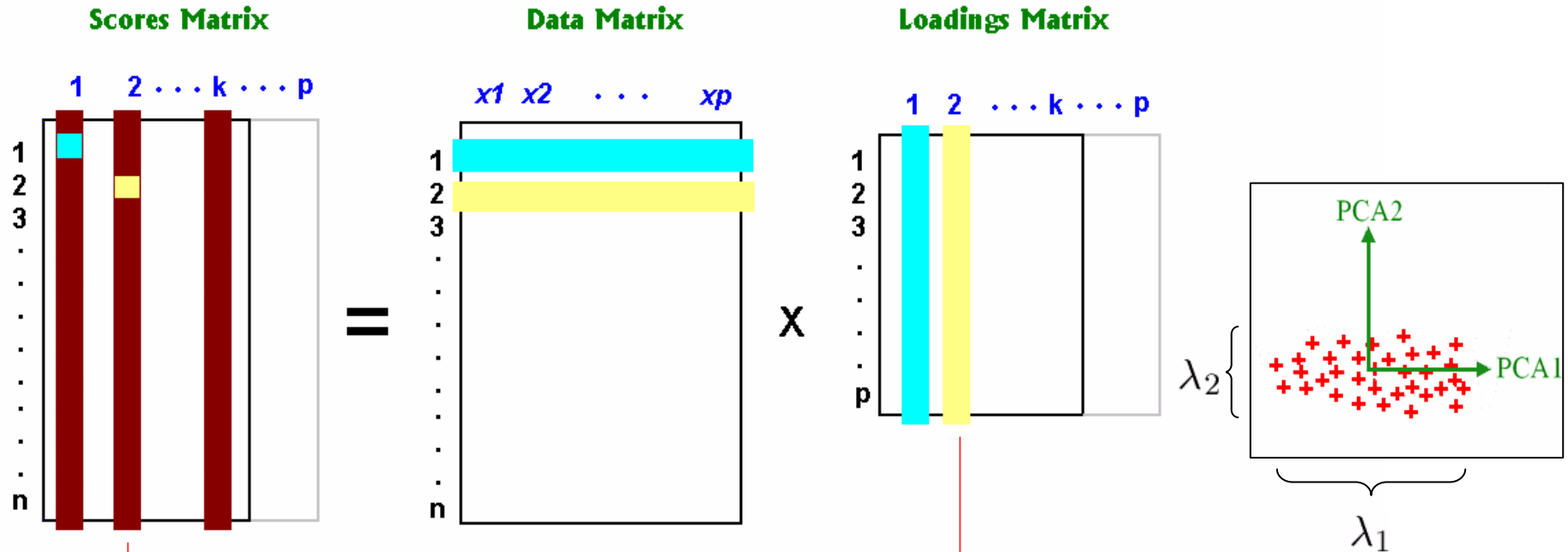
Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

$$PCA_1 = a_{11} X_1 + a_{12} X_2 + \dots + a_{1p} X_p$$

$$PCA_2 = a_{21} X_1 + a_{22} X_2 + \dots + a_{2p} X_p$$

PCA: Loadings and Scores

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



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

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

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

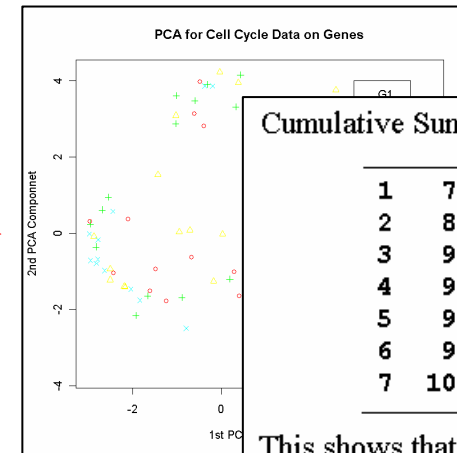
PCA (conti.)

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

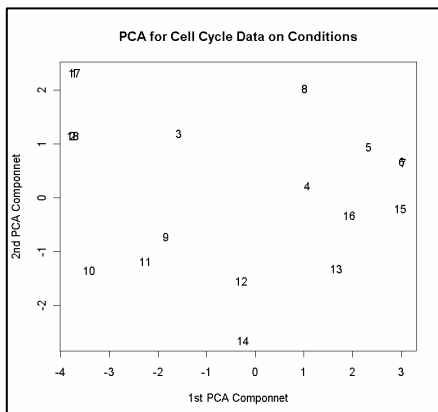


Cumulative Sum of the Variances:

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

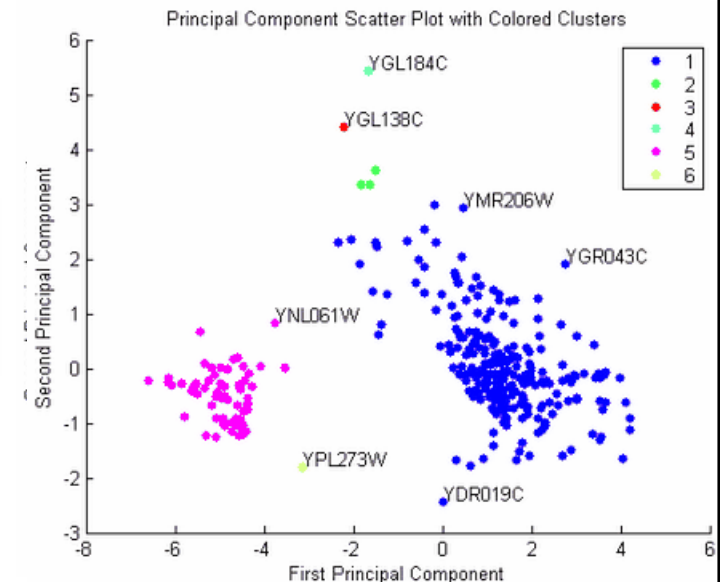
This shows that almost 90% of the variance is accounted for by the first two principal components.

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	1.31

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.

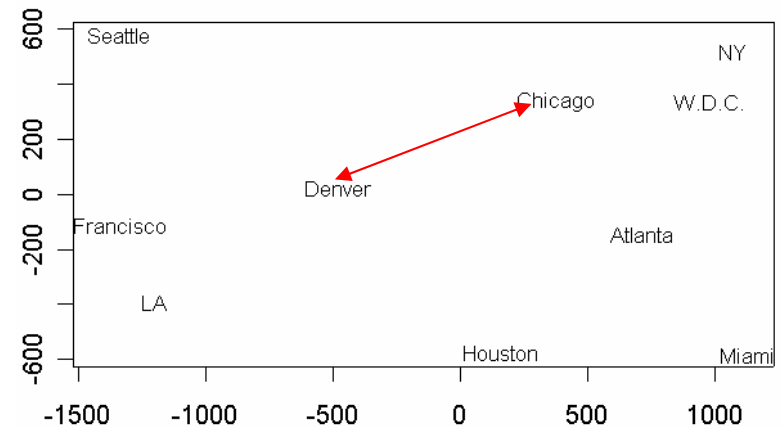
http://www.lib.utexas.edu/maps/united_states.html

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 k , compute points x_1, \dots, x_n in k -dimensional space such that the object function is minimized.

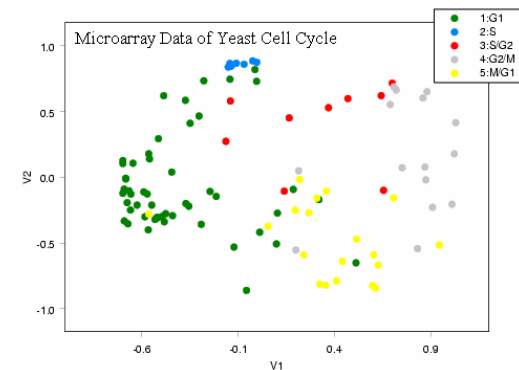
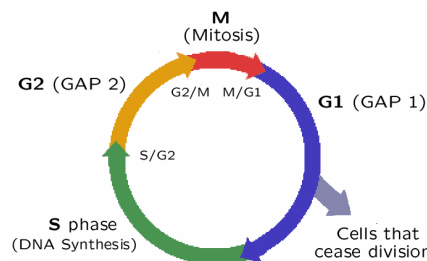
$$Stress = \sqrt{\frac{\sum_{i,j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

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.

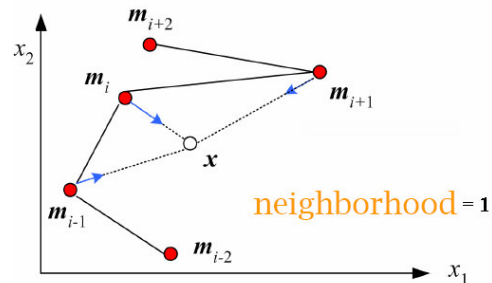


Clustering Analysis and Visualization

- ◆ **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.



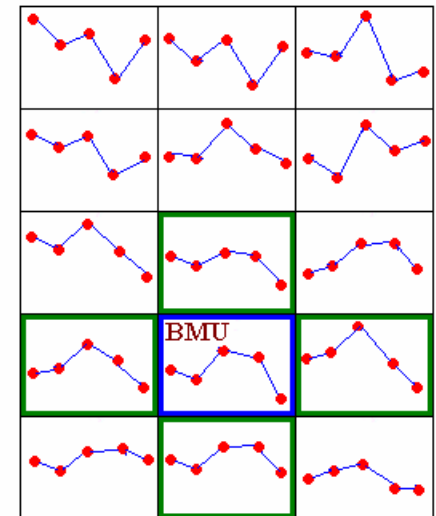
- SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by **clustering**, and to construct a nonlinear projection of the data onto a **low-dimensional display**.

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

Learning process:

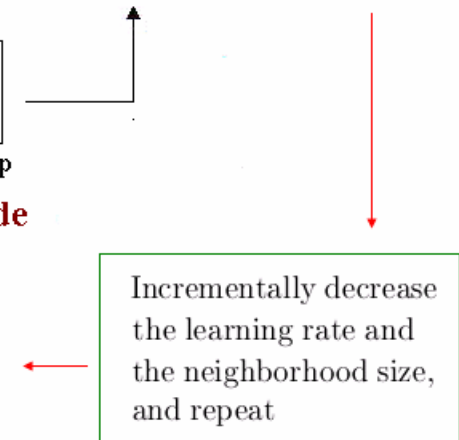
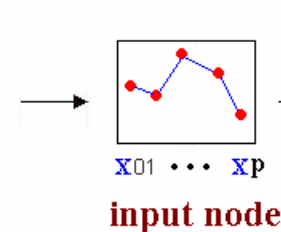
$$w_i(t+1) = \begin{cases} w_i(t) + h_{ci}(t)[x(t) - w_i(t)] & i \in N_c(t) \\ w_i(t), & \text{o.w.} \end{cases}$$

5 x 3 output node



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



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:

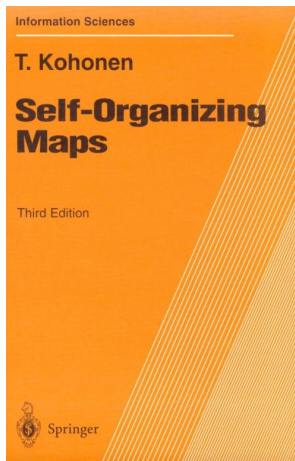
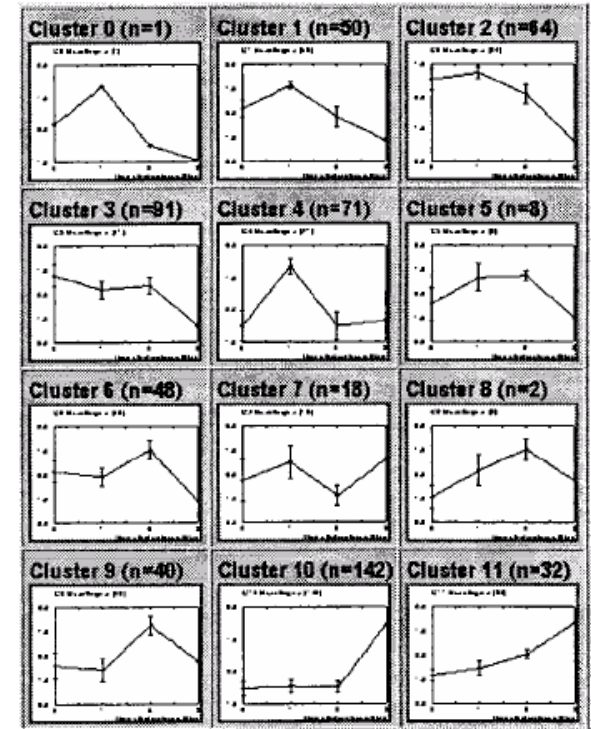
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

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.

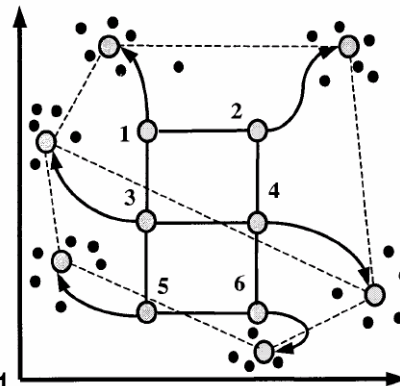
Step 2: Incrementally decrease the learning rate and the neighborhood size, and repeat Step 1.

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

HL-60 4×3 SOM 567 genes



1995, 1997, 2001



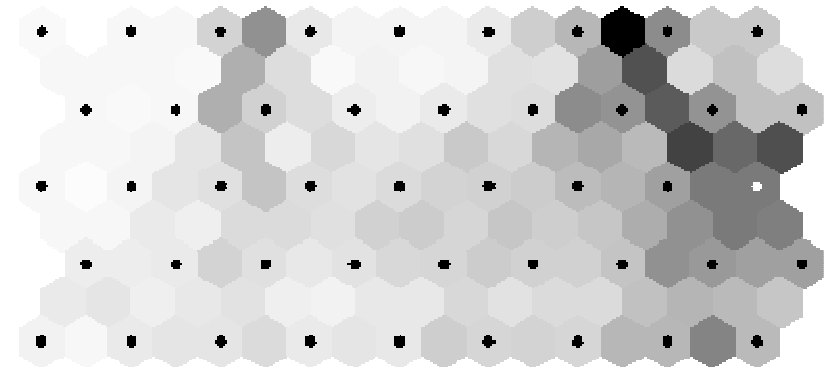
Macrophage Differentiation in HL-60 cells

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.

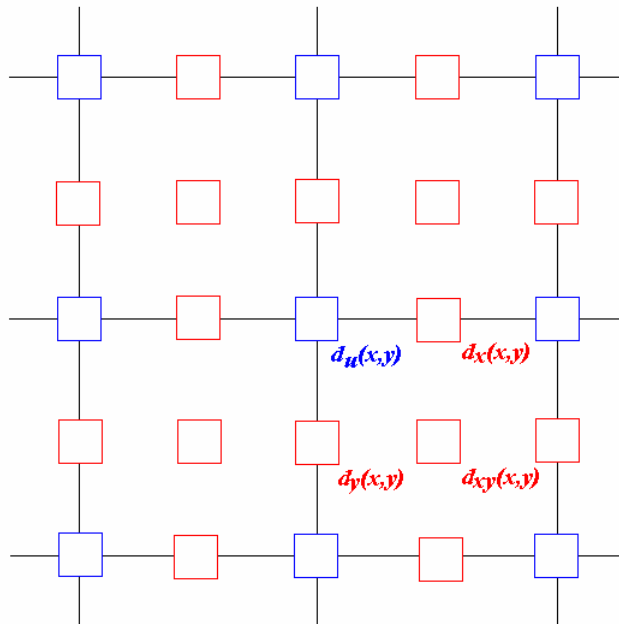
U-matrix: Unified Matrix Method

(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 $(2n_x - 1) \times (2n_y - 1)$.

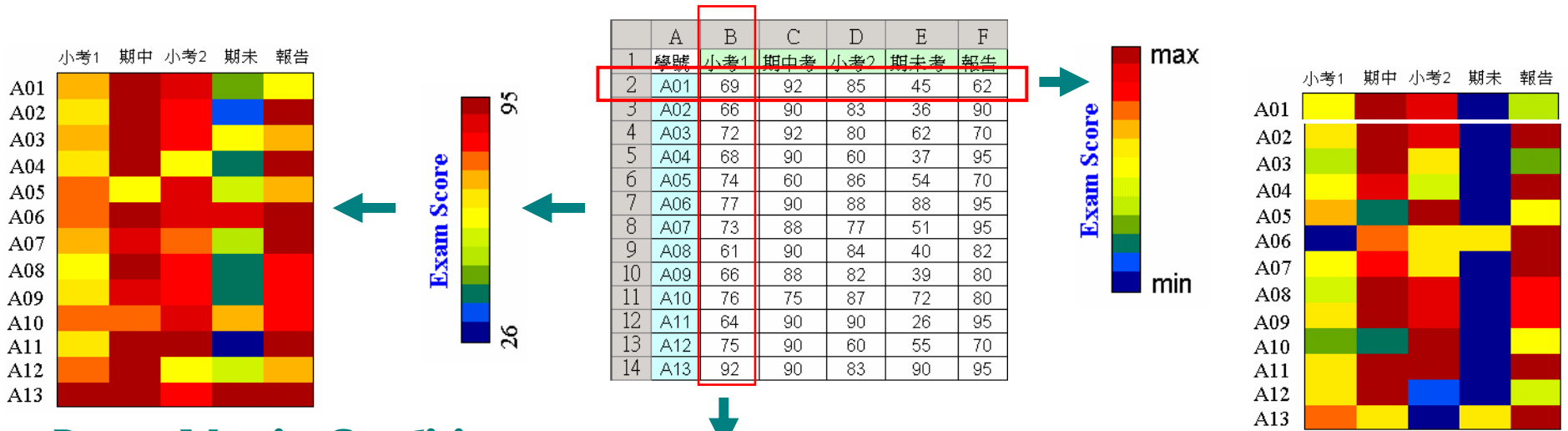
$$d_x(x, y): \|b(x, y) - b(x + 1, y)\| = \sqrt{\sum_i [w_i(x, y) - w_i(x + 1, y)]^2}$$

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

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

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

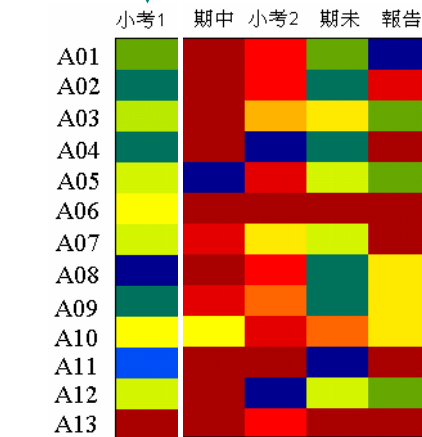
Heat Map: Data Image, Matrix Visualization



Range Matrix Condition

Range Raw Condition

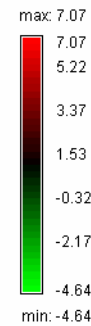
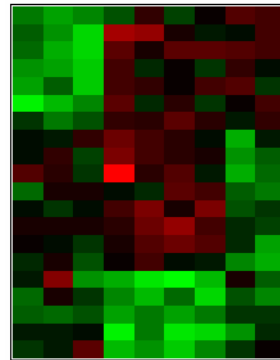
What about this one?



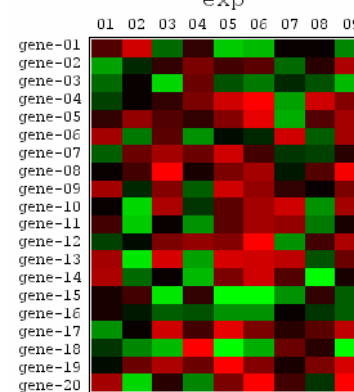
Range Column Condition

Heat Map: Display Conditions

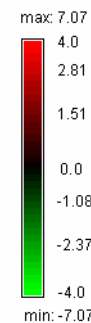
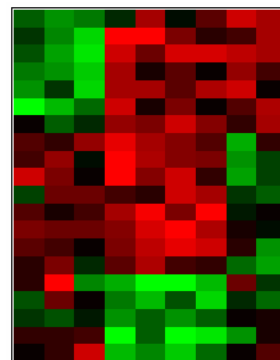
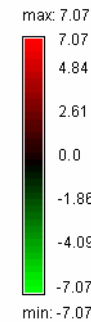
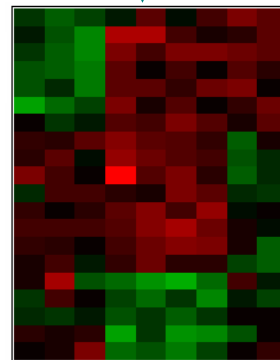
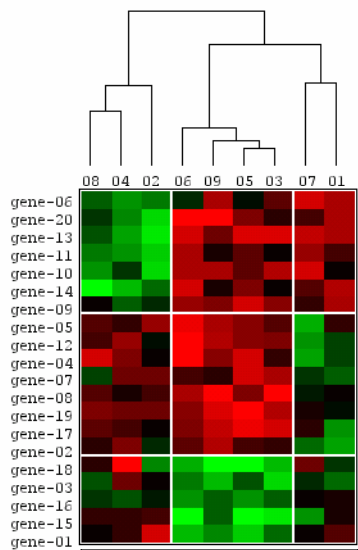
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	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



Without ordering



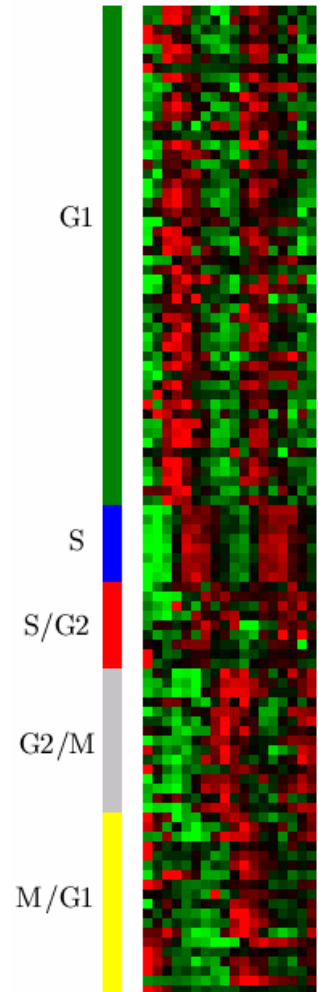
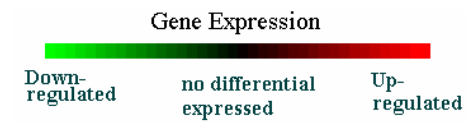
Center Matrix Condition



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

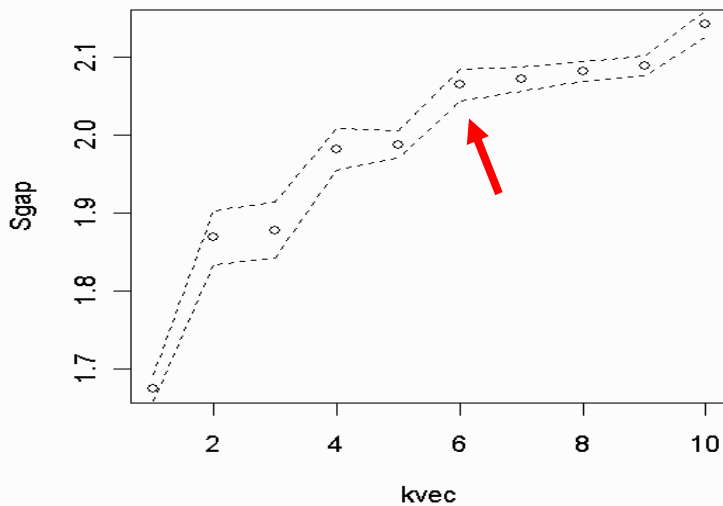
■ Data

Baseline: Culture Medium (CM-00h)

OH-04h, OH-12h, OH-24h

CA-04h, CA-24h

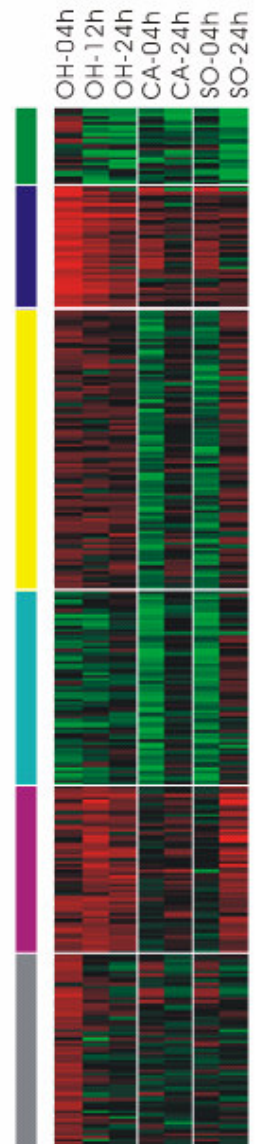
SO-04h, SO-24h



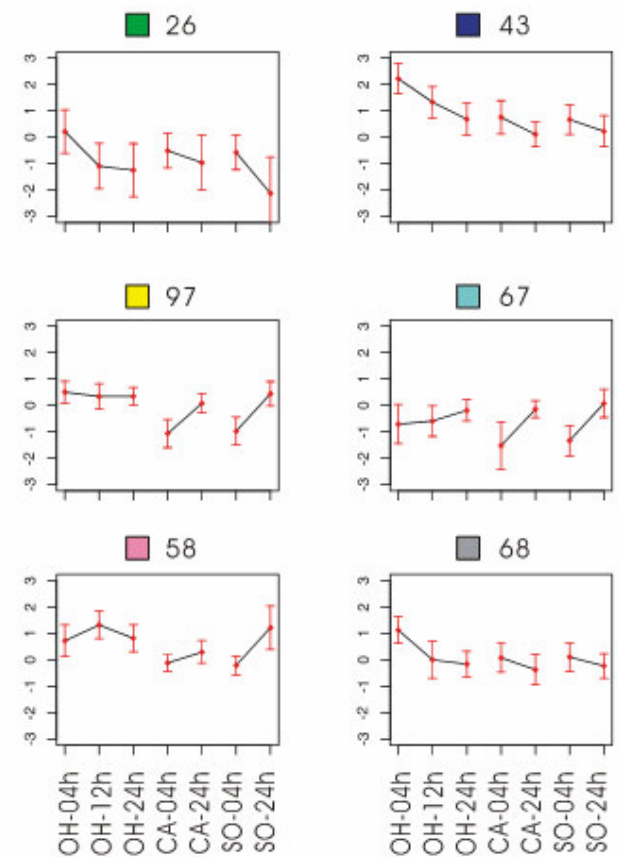
J. R. Statist. Soc. B (2001)
63, Part 2, pp. 411-423

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

Robert Tibshirani, Guenther Walther and Trevor Hastie
Stanford University, USA



K-means Clustering



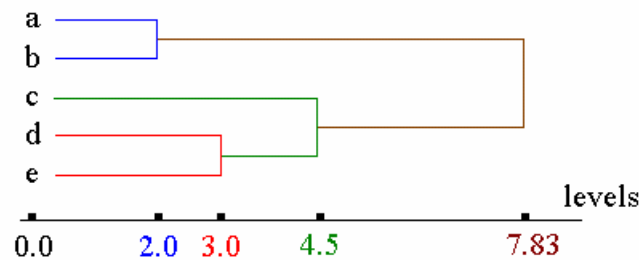
Hierarchical Clustering and Dendrogram

(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

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

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

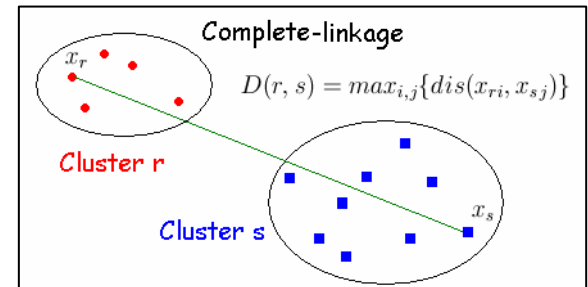
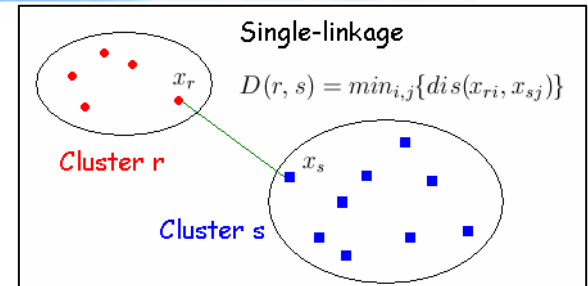
	{a, b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

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

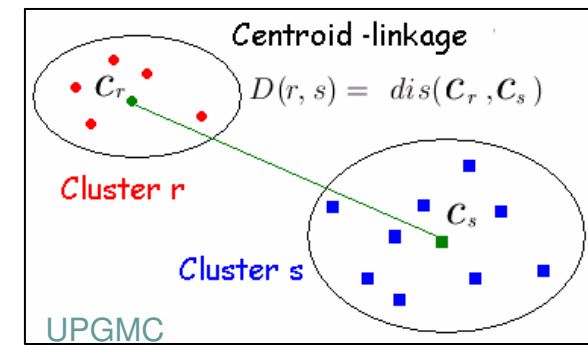
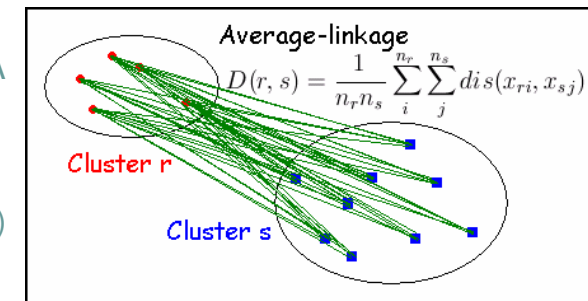
$$= \frac{1}{4}[D(a, d) + D(a, e) + D(b, d) + D(b, e)]$$

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

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



UPGMA
(Unweighted
Pair-Groups
Method
Average)



Ward's Method

- The Ward's method does not compute distances between clusters.
- It forms clusters by maximizing **within-clusters homogeneity**.
- The within-group (i.e., within-cluster) **sum of squares** is used as the measure of homogeneity.
- The Ward's method tries to minimize the total within-group or within-cluster sum of squares.
- Clusters are formed at each step such that the resulting cluster solution has the fewest within-cluster sums of squares.
- The within-cluster sums of squares that is minimized is also known as the **error sums of squares (ESS)**.

Toy Data

data	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

step Possible Partitions ESS

1	(12)	3	4	5	?
---	------	---	---	---	---

$$\{\bar{12}\} = [(10 + 20)/2, (5 + 20)/2]$$

$$= [15, 12.5]$$

$$ESS = wss\{12\} + wss\{3\} + wss\{4\} + wss\{5\}$$

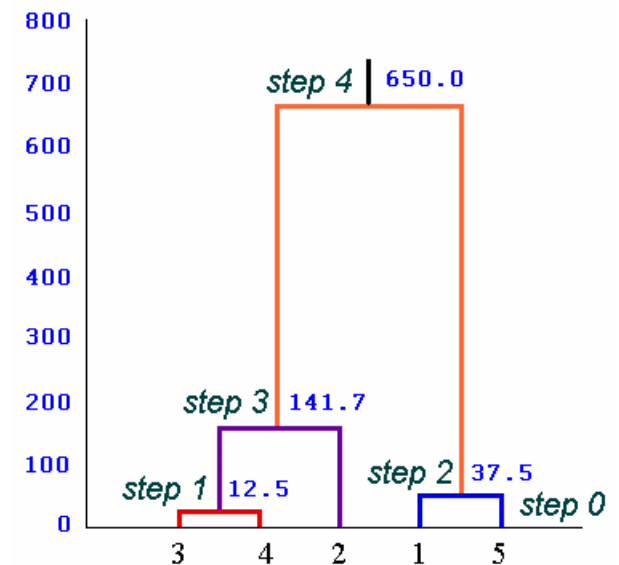
$$= ss(1, \{\bar{12}\}) + ss(2, \{\bar{12}\})$$

$$= (10 - 15)^2 + (5 - 12.5)^2 + (20 - 15)^2 + (2 - 12.5)^2$$

$$= 162.5$$

step Possible Partit.

1	(12)	3	4	5	162.5
	(13)	2	4	5	212.5
	(14)	2	3	5	250.0
	(15)	2	3	4	25.0
	(23)	1	4	5	100.0
	(24)	1	3	5	62.5
	(25)	1	3	4	162.5
	(34)	1	2	5	12.5*
	(35)	1	2	4	312.5
	(45)	1	2	3	325.0
2	(34)	(12)	5	175.0	
	(34)	(15)	2	37.5*	
	(34)	(25)	1	175.0	
	(134)	2	5	316.7	
	(234)	1	5	116.7	
	(345)	1	2	433.3	
3	(234)	(15)	141.7*		
	(125)	(34)	245.9		
	(1345)	2	568.8		
4	(12345)		650.0		

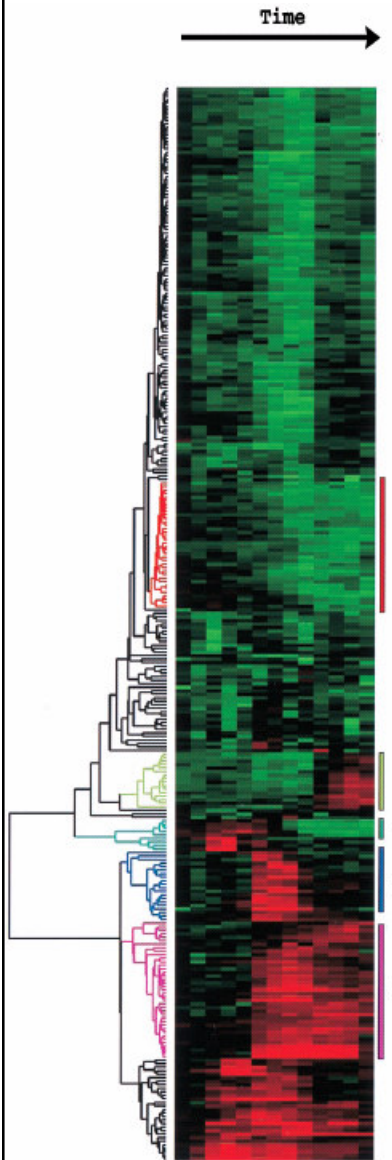


Example:

Charles H. Romesburg (1984)

Display of Genome-Wide Expression Patterns

28 / 61



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

Cluster analysis and display of genome-wide expression patterns

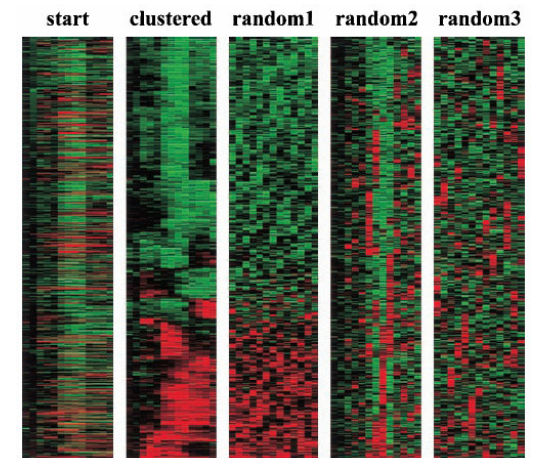
MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

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

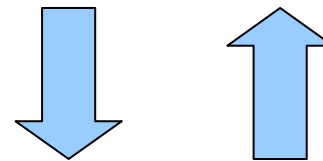
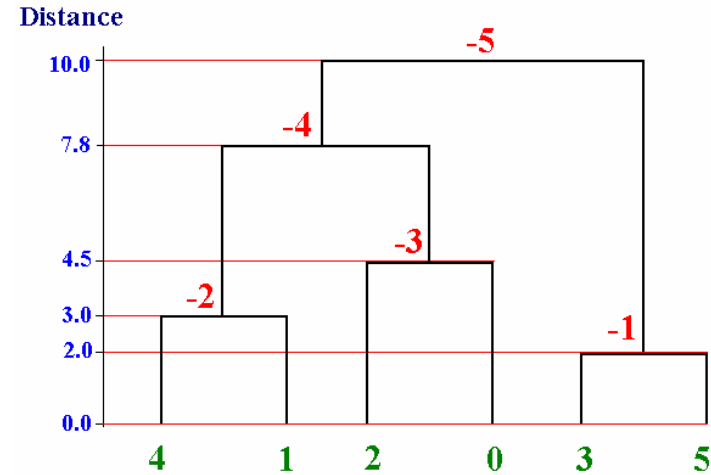
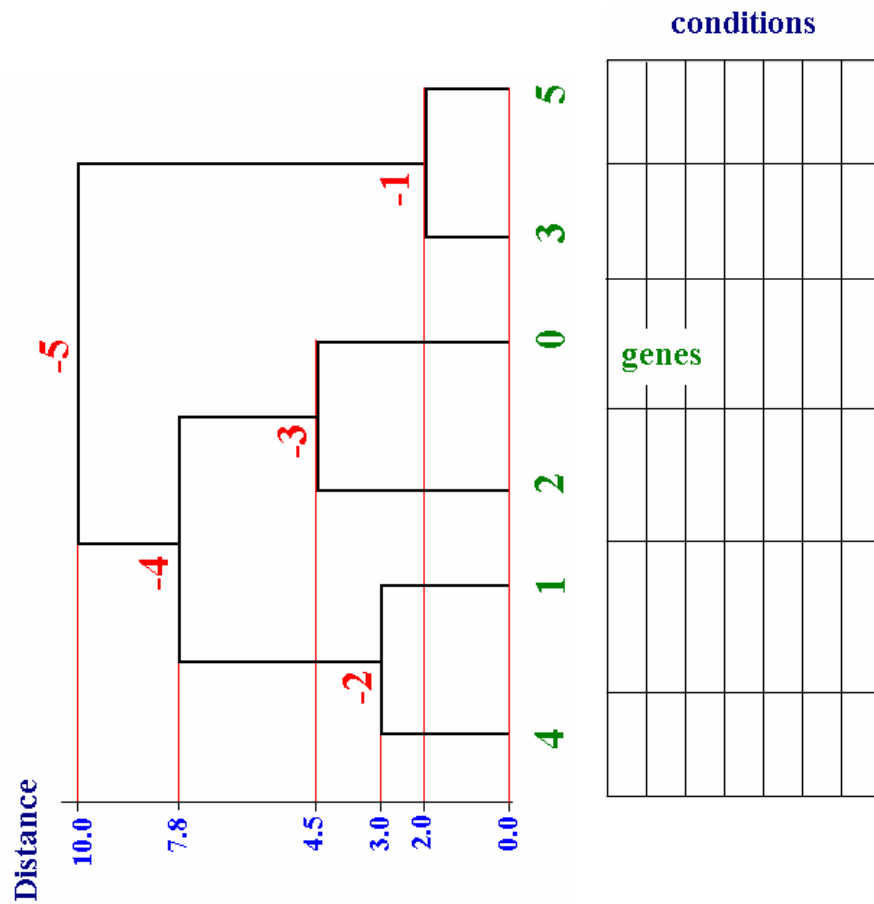
human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated 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>.

Software: Cluster and TreeView

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).



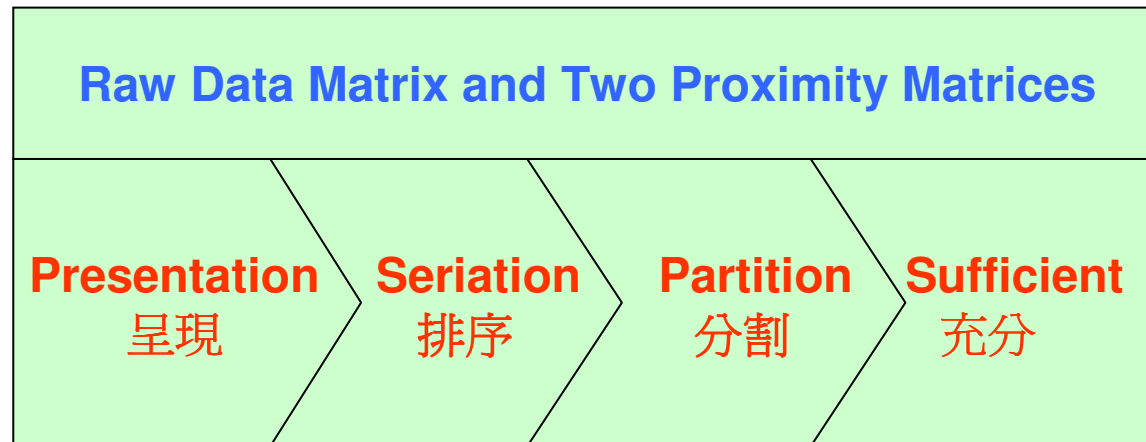
Dendrogram and Tree Storage



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

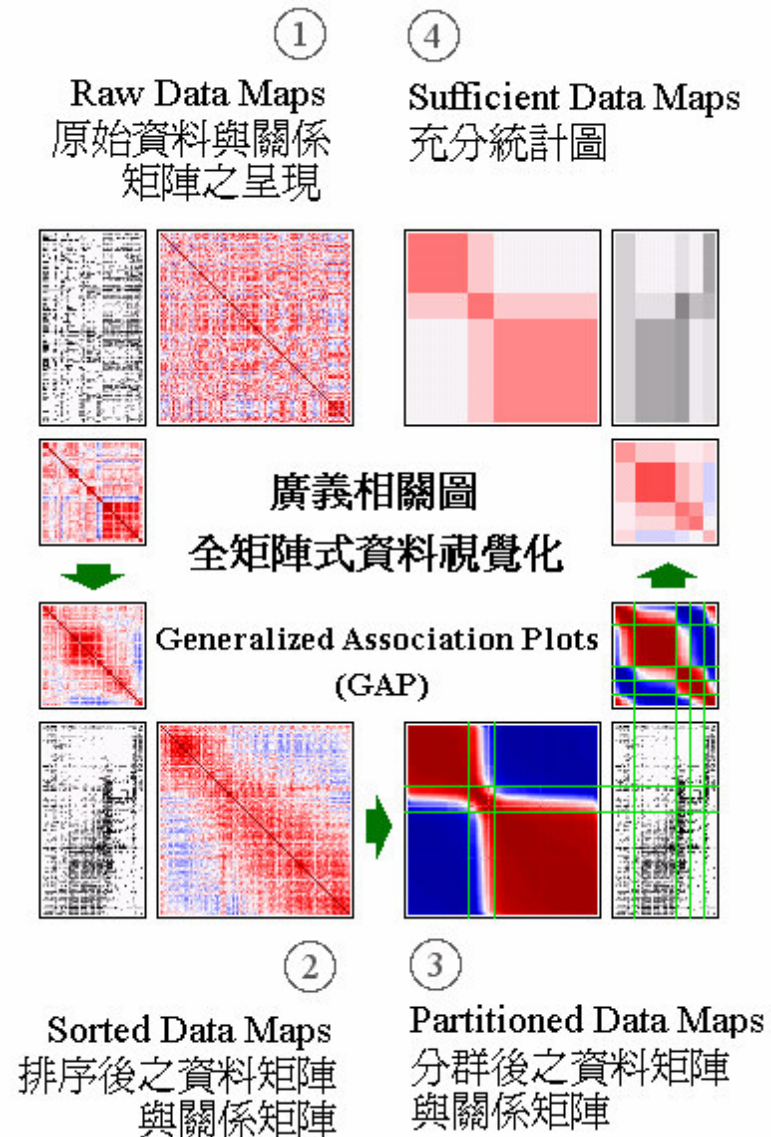
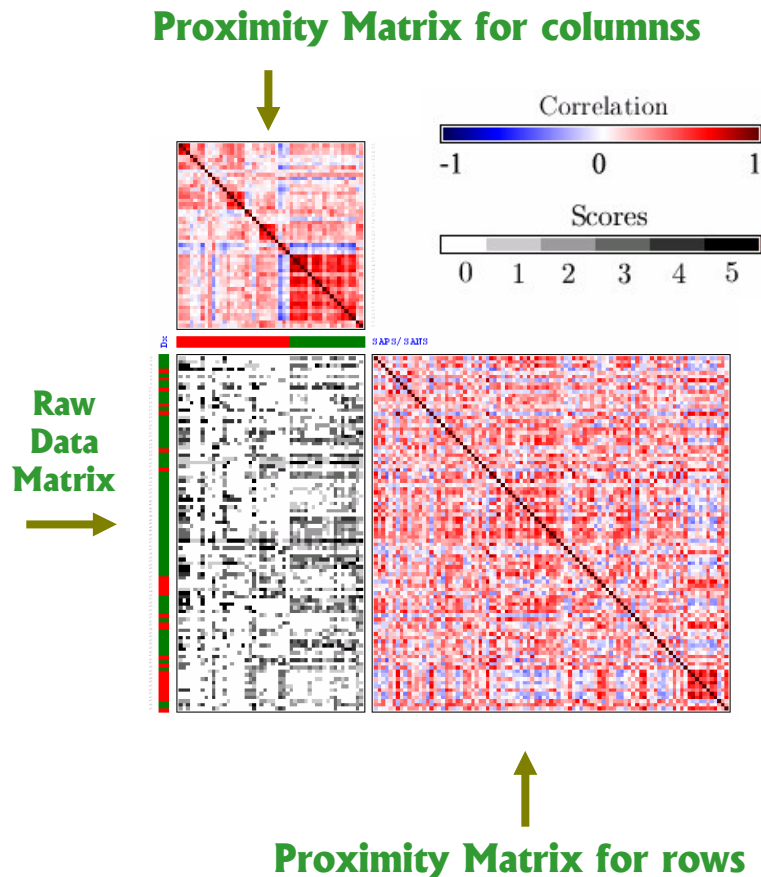
Generalized Association Plots



Generalized Association Plots (GAP)

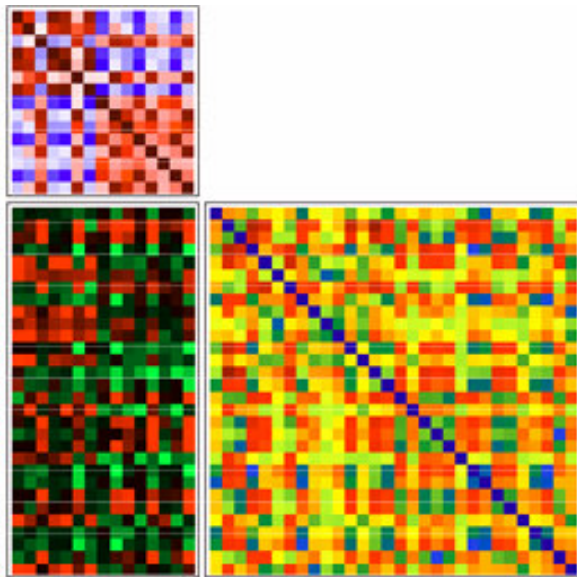
(Chen, 2002)

- 95 patients: 69 schizophrenic and 26 bipolar disorders
- SAPS: 30 items, SANS: 20 items
- Six point scale (0-5).



Presentation of Raw Data Matrix

1. Color spectrum
2. Variable transformation
3. Selection of proximity



“Resolution”
of a
Statistical
Graph

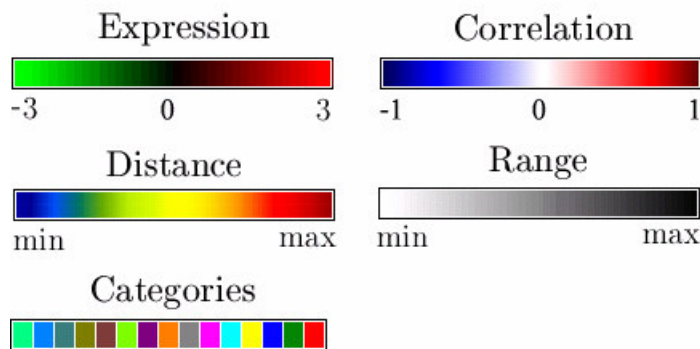
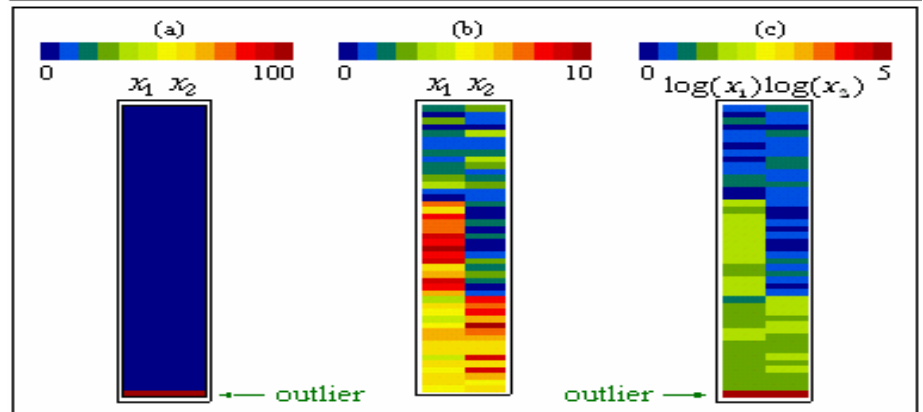
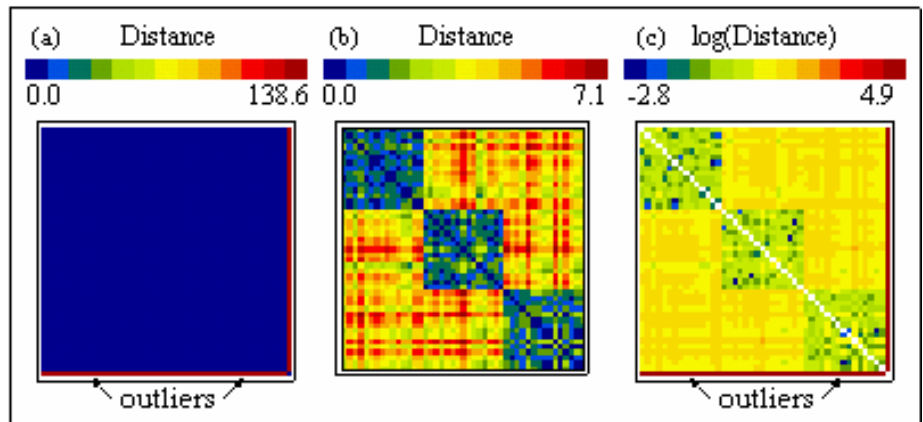
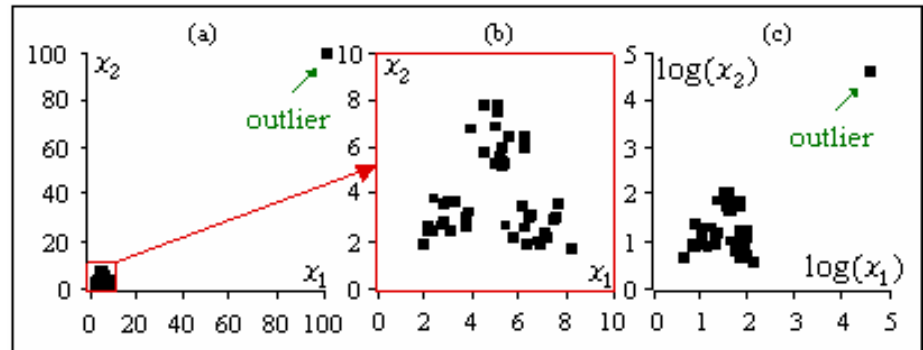


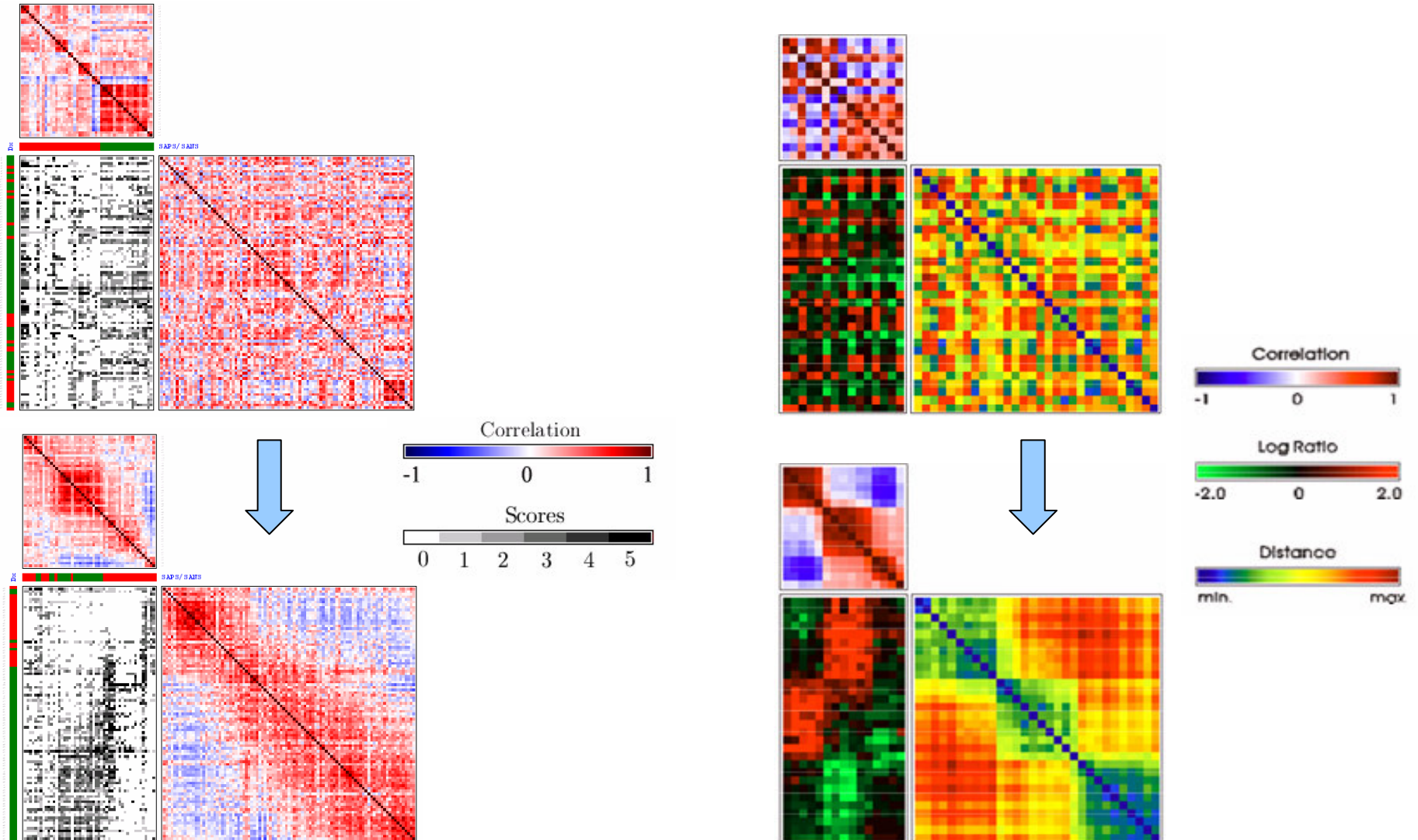
Image source: Dr. Chen Chun-houh's Slide



Concept of Relativity of a Statistical Graph

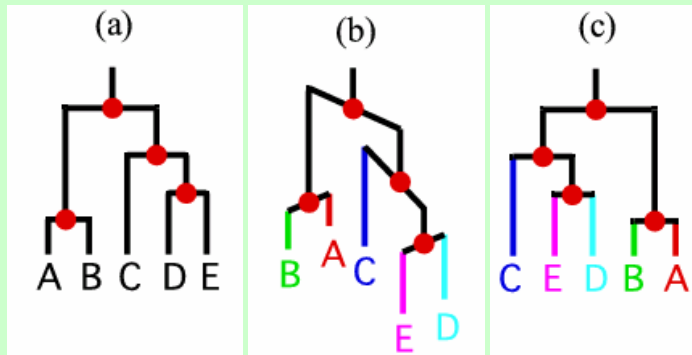
33 / 61

Placing similar (different) objects at closer (distant) positions

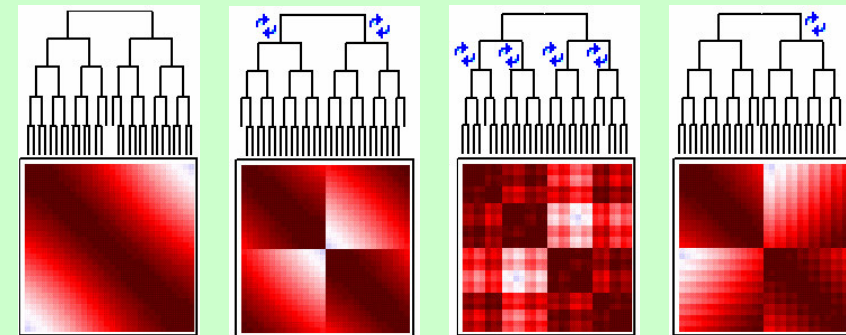


Seriation Problem for Hierarchical Clustering

Tree seriation

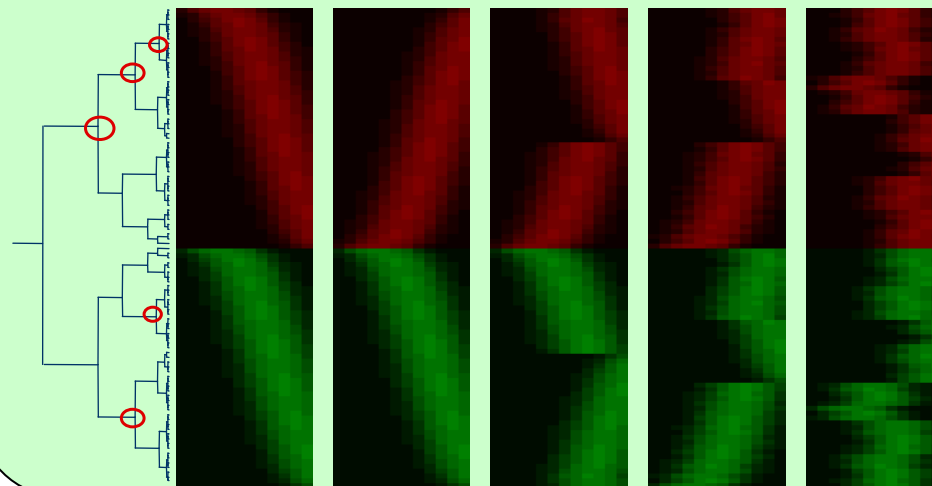


Tree seriation for proximity matrices



Tree seriation for raw data matrices

ideal model 1 flip 3 flips 5 flips many flips

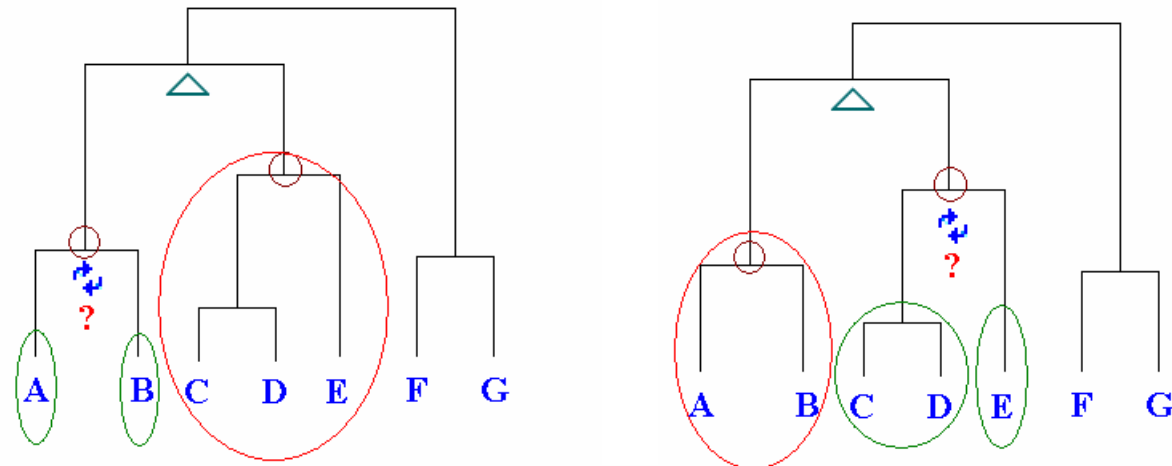


**Different Seriations
Generated from Identical
Tree Structure**

Internal Tree Flips

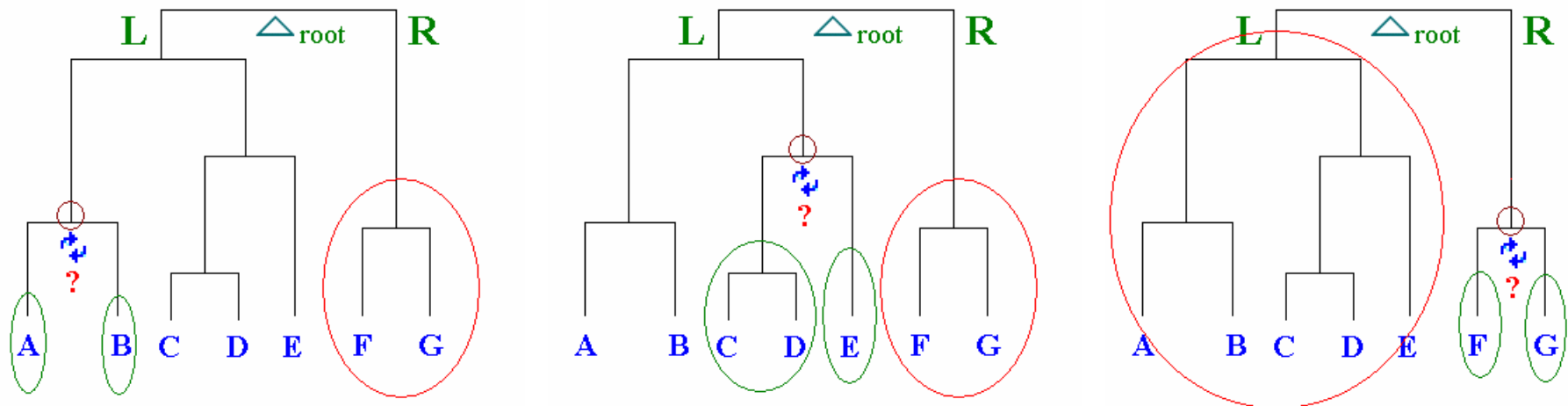
35 / 61

Uncle Approach



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

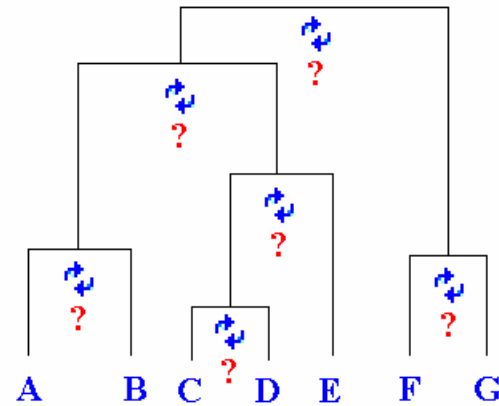
GrandPa Approach



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

36 / 61



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.

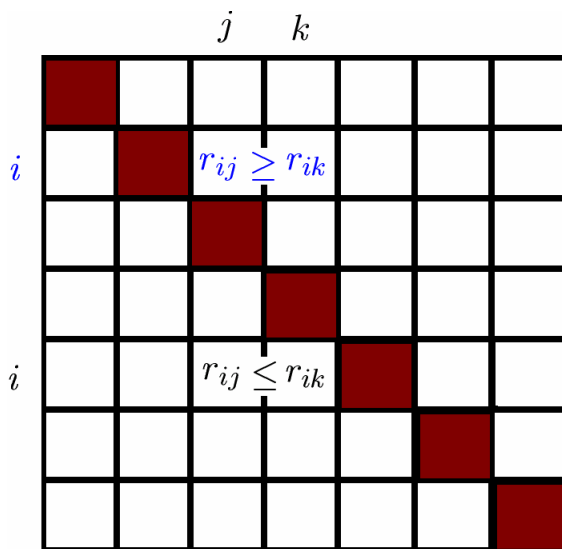
Criteria for a “good” Permutation

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

Global/local Criterion: Anti-Robinson Measurements

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

Robinson Form

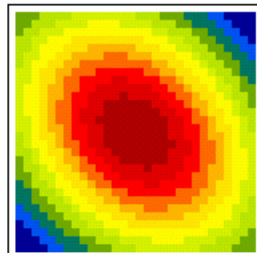


$r_{ij} \leq r_{ik}$ if $j < k < i$, $r_{ij} \geq r_{ik}$ if $i < j < k$

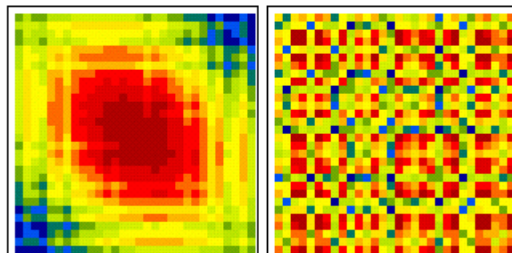


Min.

Max.



Robinson



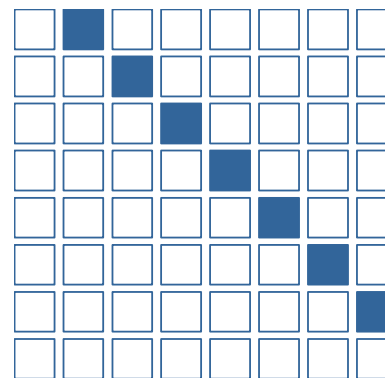
pre-Robinson

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

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

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

Local criterion: Minimal Span Loss Function



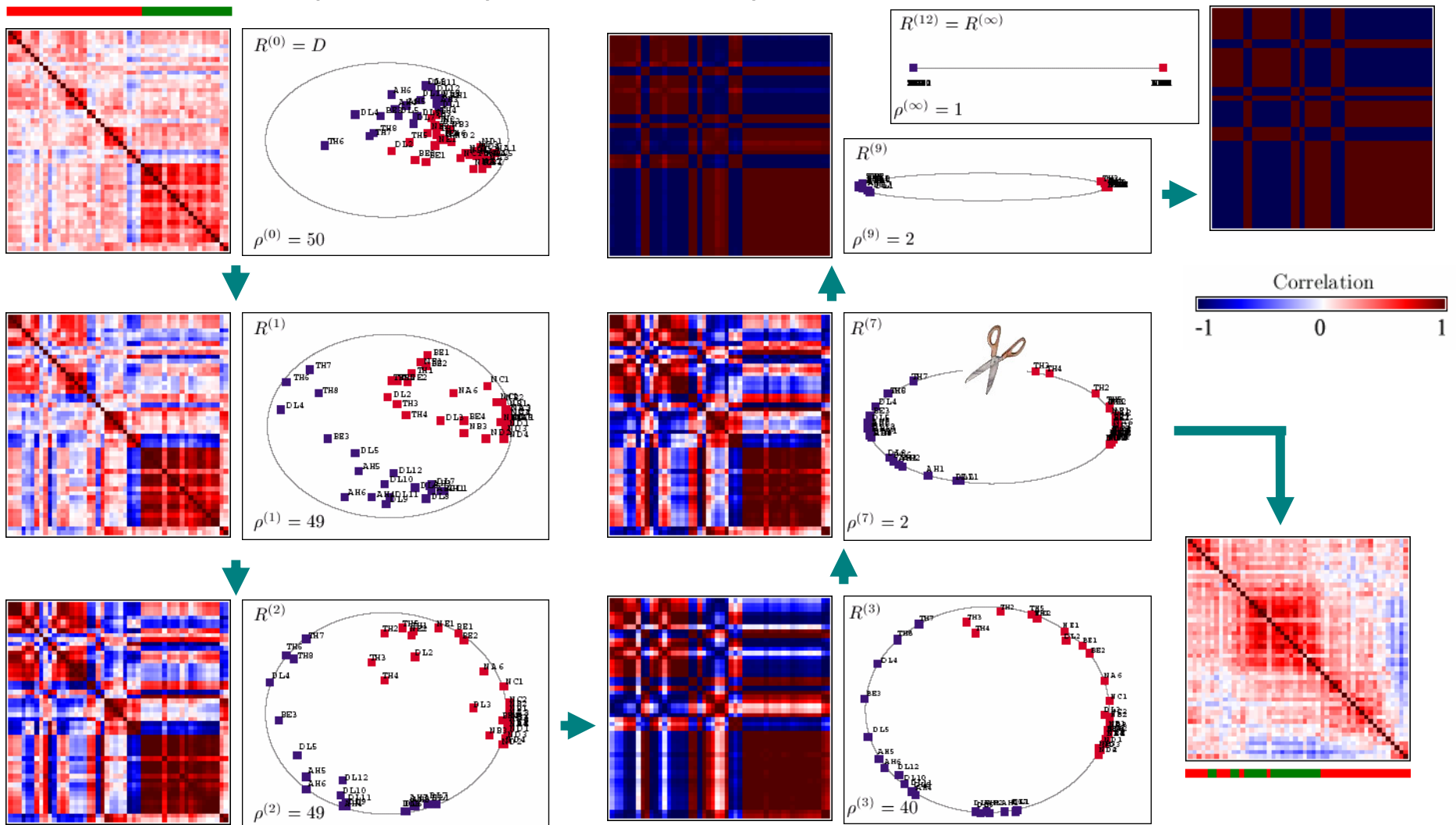
$$MS = \sum_{i=1}^{n-1} d_{i,i+1}$$

Further Reading

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

GAP Rank-Two Elliptical Seriation

- Seriation Algorithms with Converging Correlation Matrices
- When the sequence reaches an iteration with rank two, the p objects fall on an ellipse and have unique relative position on the ellipse.



Global vs Local Seriation

39 / 61

GAP Elliptical Seriation

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

GAP Elliptical Seriation

Michael Eisen Tree Seriation

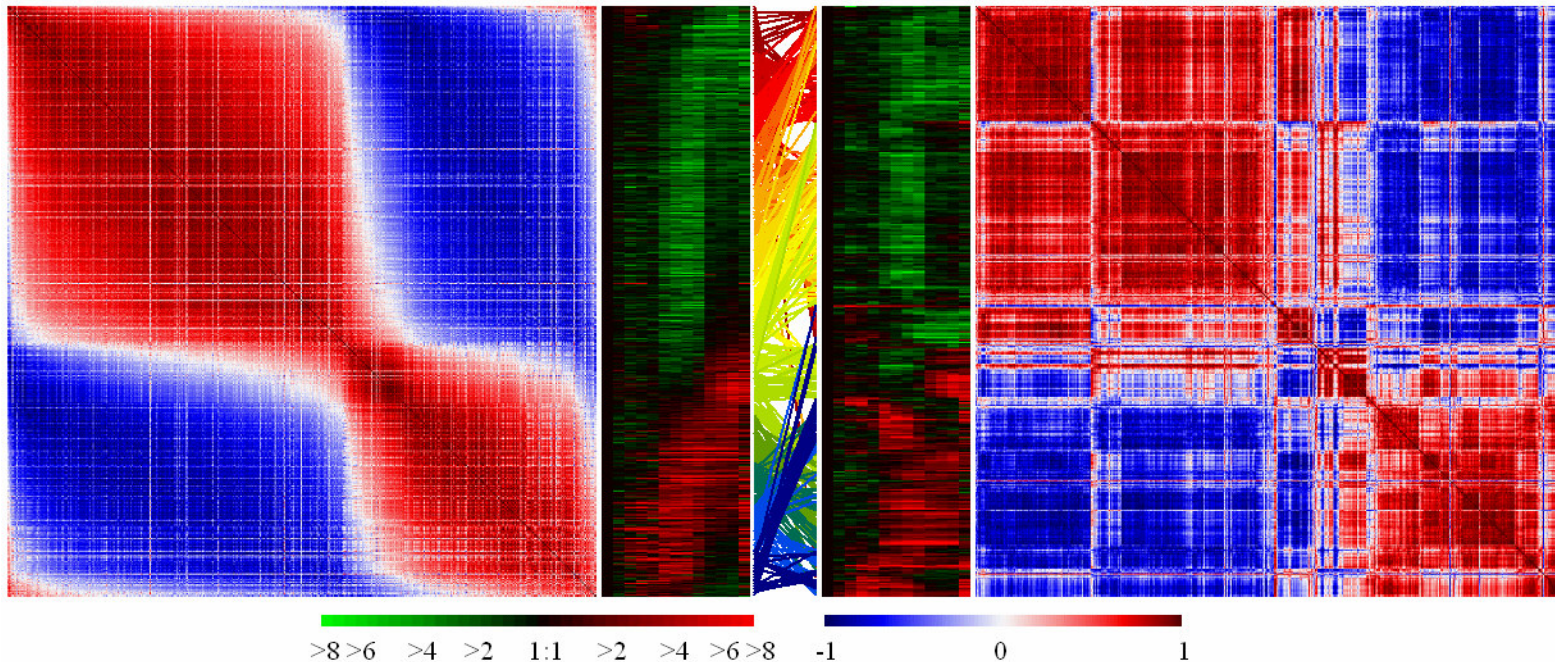
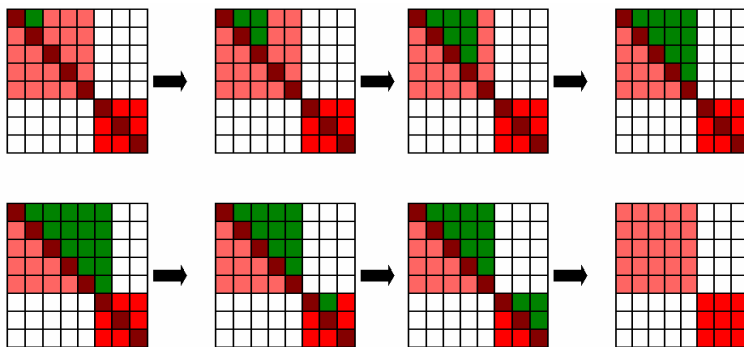


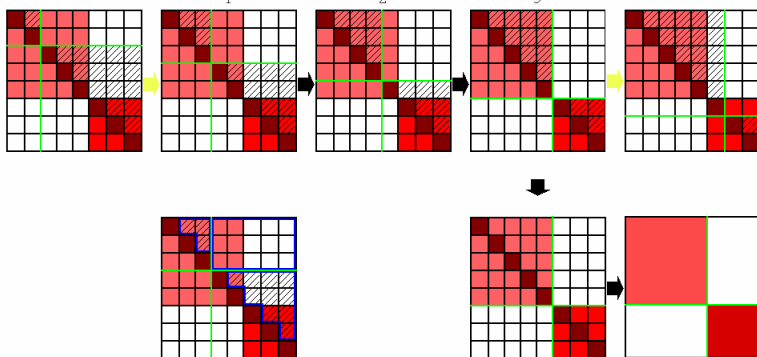
Image source: Dr. Chen Chun-houh's slide

Partitions of Permuted Matrix Maps

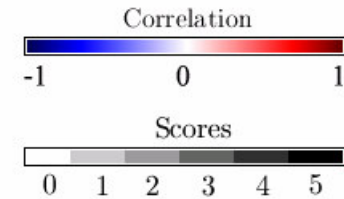
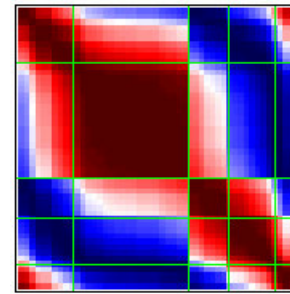
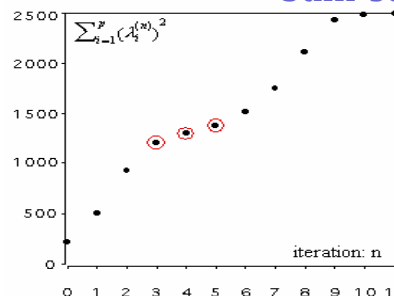
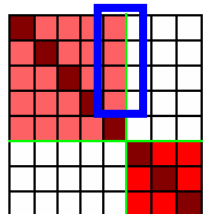
One-Way block Searching



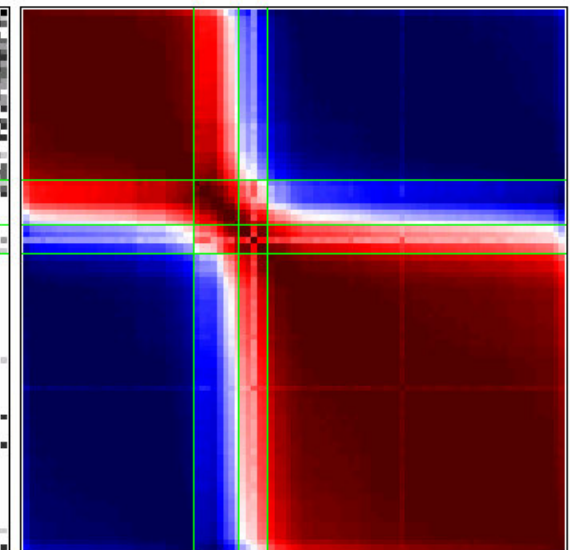
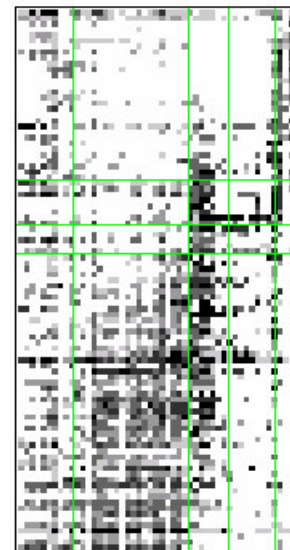
Within-Sum-of-Square Approach



Two-Sample Problem



Row: $R^{(3)}$, Column: $R^{(4)}$



Sum squared eigenvalues (sum squared correlations)

Further Reading

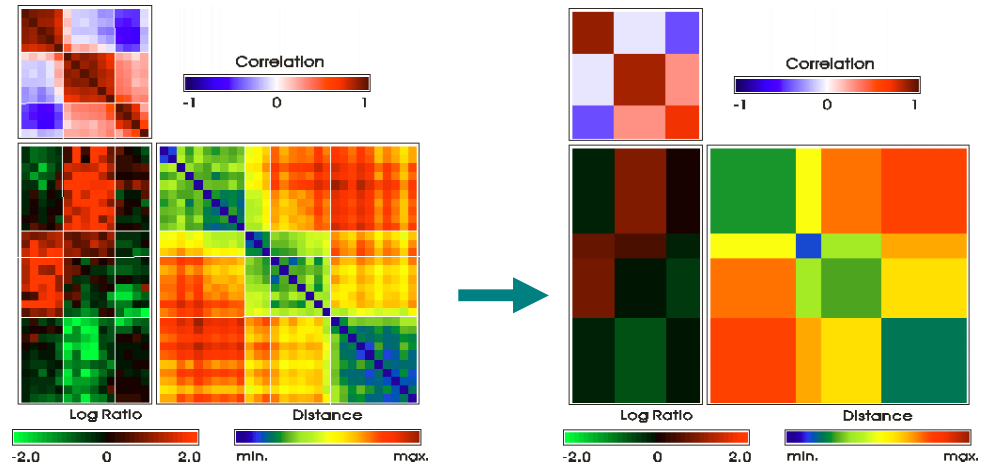
J. A. Hartigan. Direct clustering of a data matrix. Journal of the American Statistical Association, 67(337):123-129, March 1972.
 Duffy, D. & Quiroz, A. (1991), 'A permutation-based algorithm for block clustering', J. of Classification 8, 65--91.

Sufficient Graph

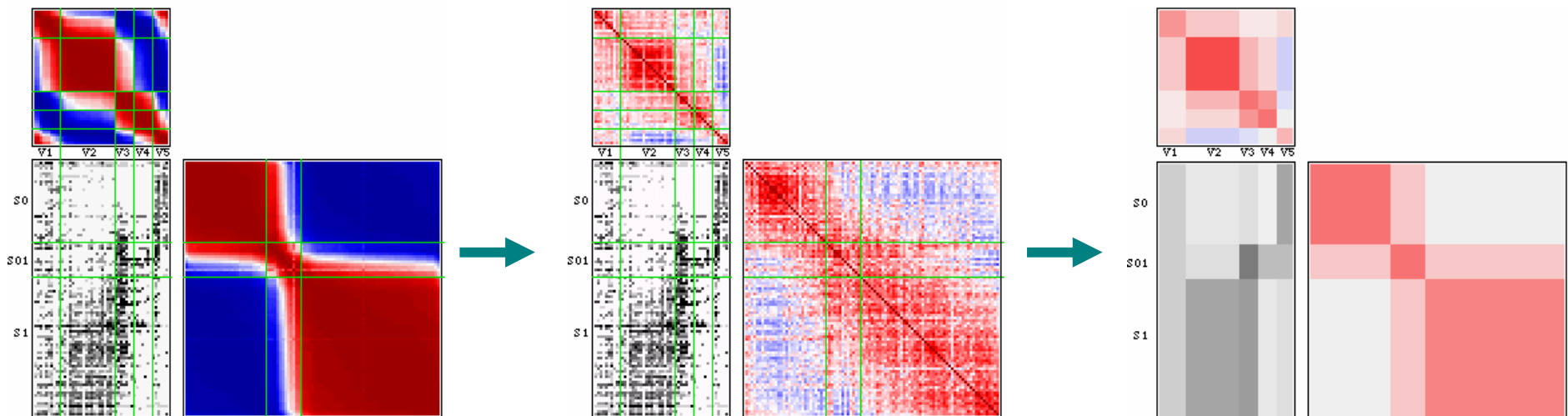
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

**Sufficient
Statistic**

	小考1	期中考	小考2	期末考	報告
平均	71.77	86.54	80.38	53.46	83
低平均	65.67	81.83	73.67	53.67	72
高平均	77.83	90.67	86.67	53.67	94.17



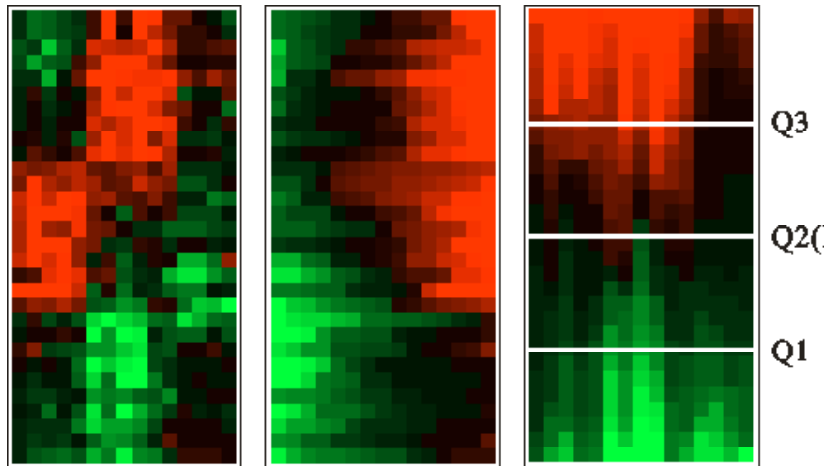
70



Generalization and Flexibility

42 / 61

Sedimented MV for patients and symptoms.



The sediment MV for patients: express severity structure.

The sediment MV for symptoms: this is a side-by-side bar-chart and box-plot which displays the distribution structure for all symptoms simultaneously.

Image source: Chen et al 2004

Sectional MV for the permuted correlation coefficient map.

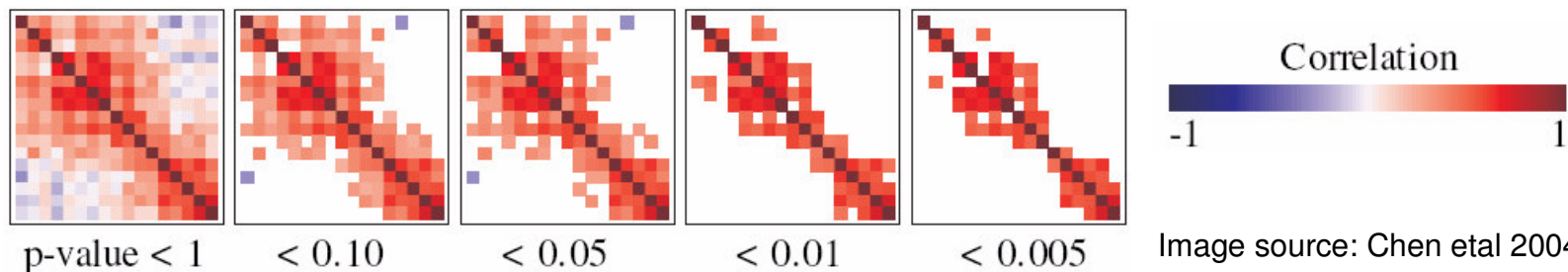
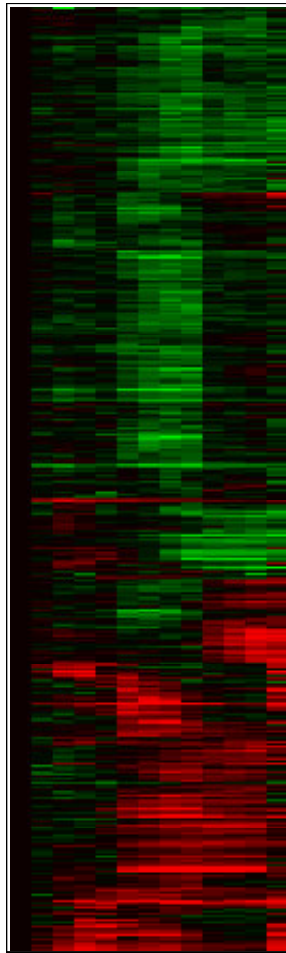


Image source: Chen et al 2004

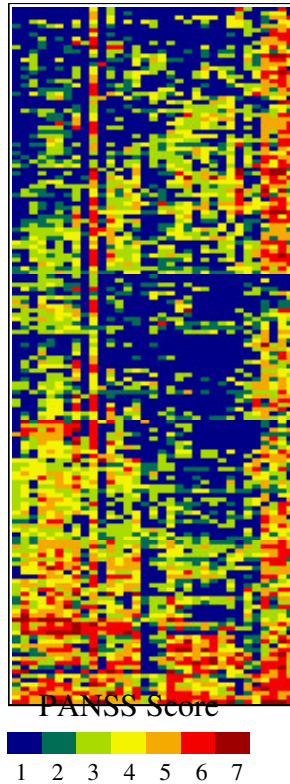
Visualization of Data Matrices

Simple ← Information Visualization of Data Matrices → Difficult

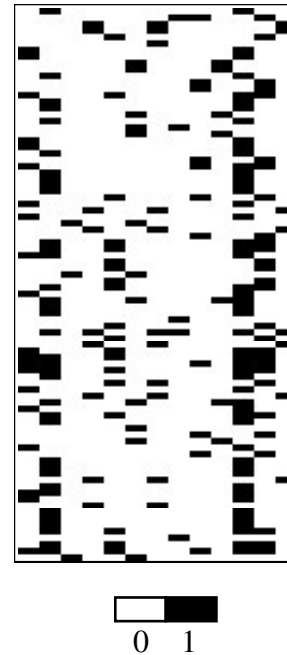
Continuous
(Gene/Time)



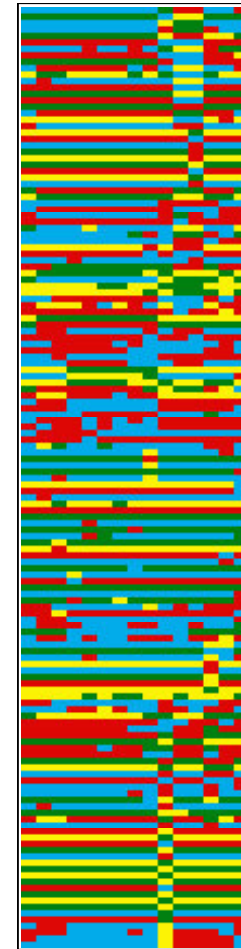
Ordinal
(Patient/Symptom)



Binary
(Mouse/Tumor)



Categorical
(Subject/SNP)



>8 >6 >4 >2 1:1 >2 >4 >6 >8 Log2ratio

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:

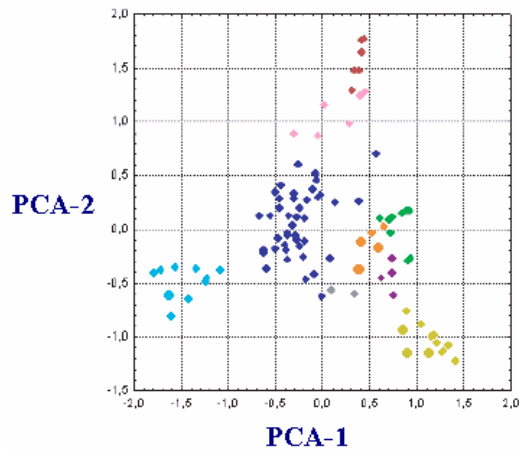
Help to decide the **number of clusters** in the data.

Choosing the Number of Clusters

45 / 61

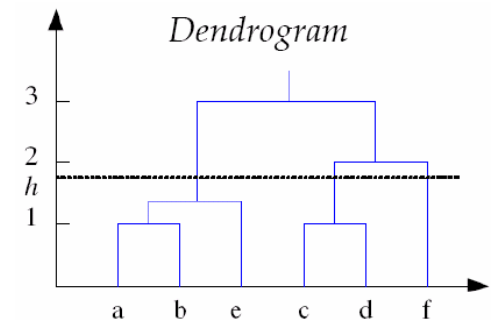
(1) K is defined by the application.

(2) Plot the data in two PAC dimensions.

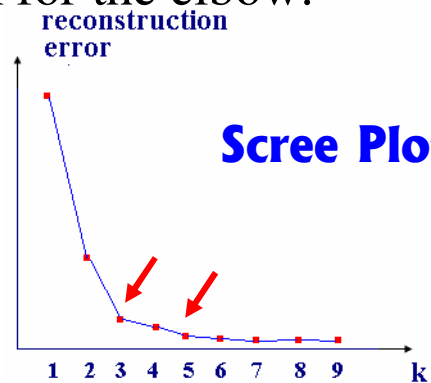
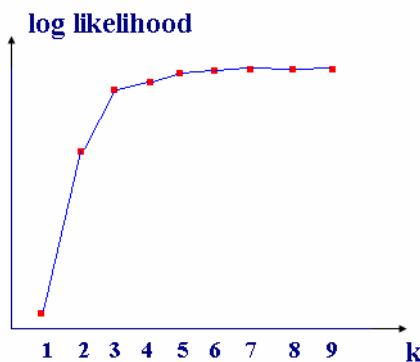


(e.g., k-means:
within-cluster sum of
squares)

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



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



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

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

46 / 61

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 Bioinformatics 2

More than 30 papers for Microarray!

with? BMC

2004

- Daxin Jiang, Chun Tang and Aiqiong 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 (1)

Statistical Evaluation: Internal Measures

- ◆ Connectivity
- ◆ Dunn index
- ◆ Silhouette Width
- ◆ Within-cluster Variance

Compactness

Homogeneity

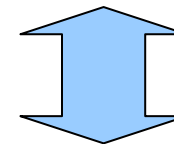
Separation

Statistical Evaluation: Stability

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

Full data

	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	0.18
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	-2.58
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



Compare two clusterings

	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

Cluster Validation Index (2)

48 / 61

■ Agreement with Reference Partition

- ◆ Rand index
- ◆ Jaccard coefficient,
- ◆ Minikowski Measure
- ◆ **Adjusted Rand index**

See also

clValid: an R package for cluster validation.

■ Biological Evaluation

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

Datta and Datta (BMC Bioinformatics, 2006)

Biological Evaluation: Homogeneity

4.1 Biological Homogeneity Index (BHI)

- $\mathcal{B} = \{B_1, \dots, B_F\}$: 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 ,
- $I(B^i = B^j) = \begin{cases} 1, & \text{if } B^i \text{ and } B^j \text{ match,} \\ 0, & \text{otherwise.} \end{cases}$
- Given statistical clustering partition $\mathcal{C} = \{C_1, \dots, C_K\}$ and set of biological classes $\mathcal{B} = \{B_1, \dots, B_F\}$, the BHI is defined as

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

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

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

Biological Evaluation: Stability

50 / 61

4.2 Biological Stability Index (BSI)

- The BSI is defined as

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

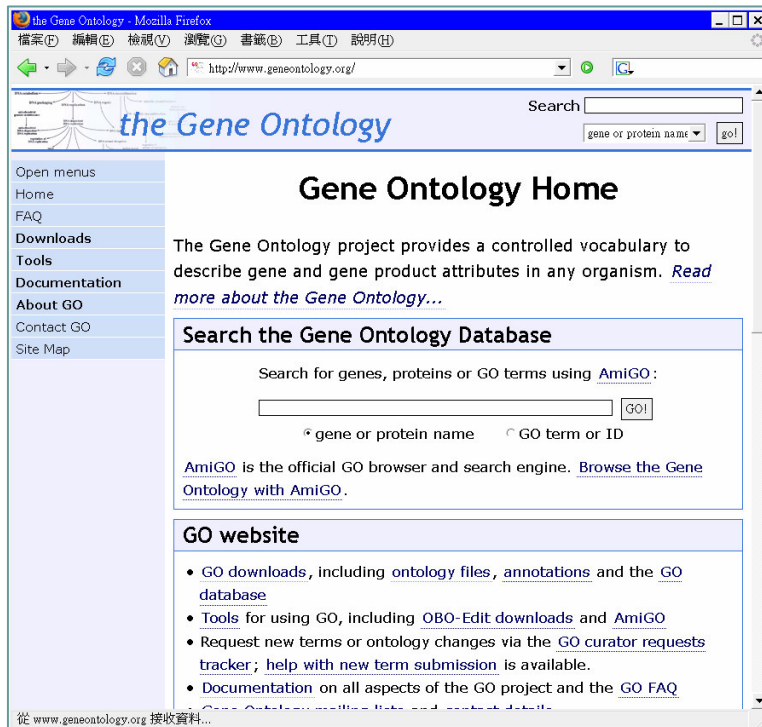
- $C^{i,0}$: the statistical cluster containing observation i based on all the data.
- $C^{j,\ell}$: the statistical cluster containing observation j when column ℓ is removed.
- Range $[0, 1]$: maximum.
- Susmita Datta and Somnath Datta, (2006), Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, *BMC Bioinformatics* 7:397.

Obtain Functional Categories (Annotation)

51 / 61

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

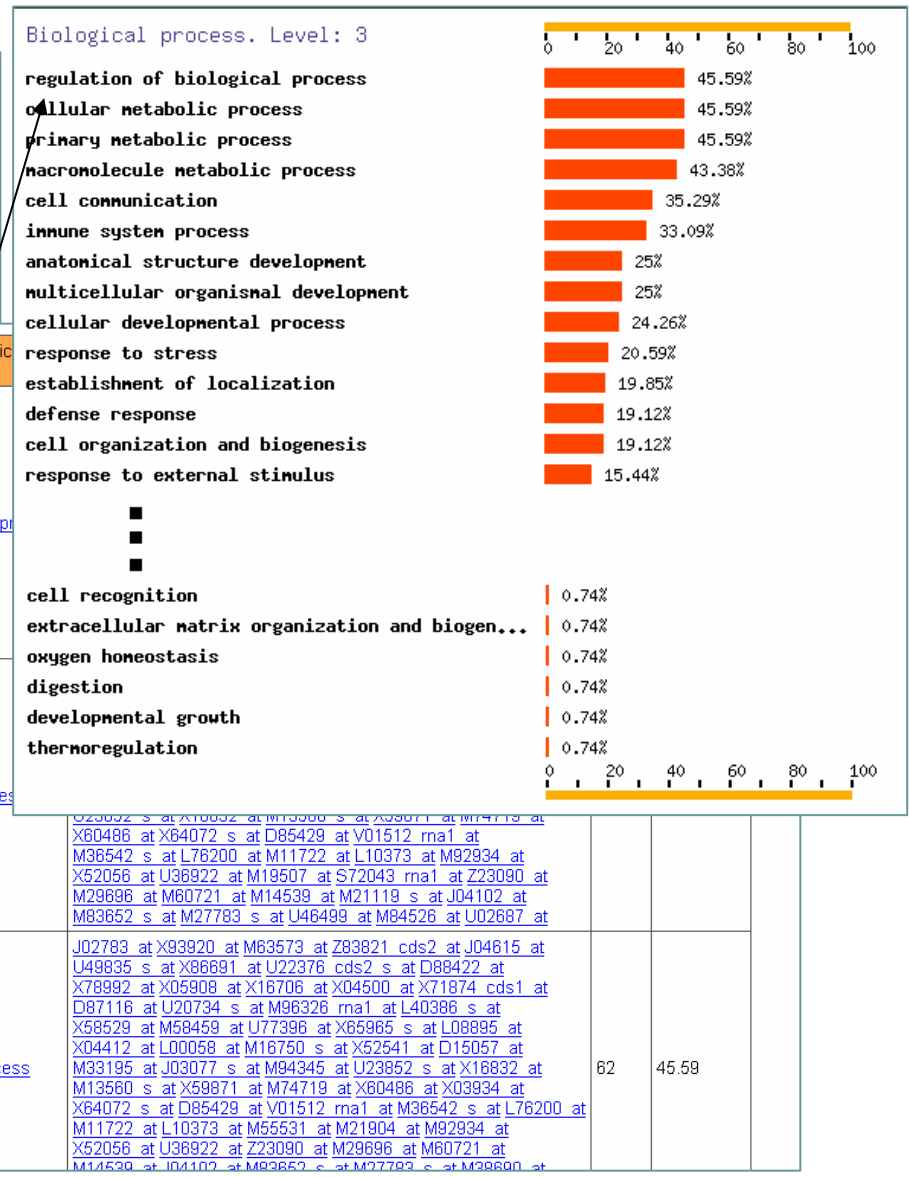


GO: Gene Ontology

- 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)

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

The screenshot shows the FatiGO web interface. It includes a search bar, a dropdown menu for 'Organism', a 'List of genes' input field, and a 'Functional annotation' section with radio buttons for 'Gene Ontology: biological process', 'Gene Ontology: molecular function', and 'Gene Ontology: cellular component'. Below the search options, there are 'References' and a 'Run' button.



The ontologies are used to categorize gene products.

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

Software

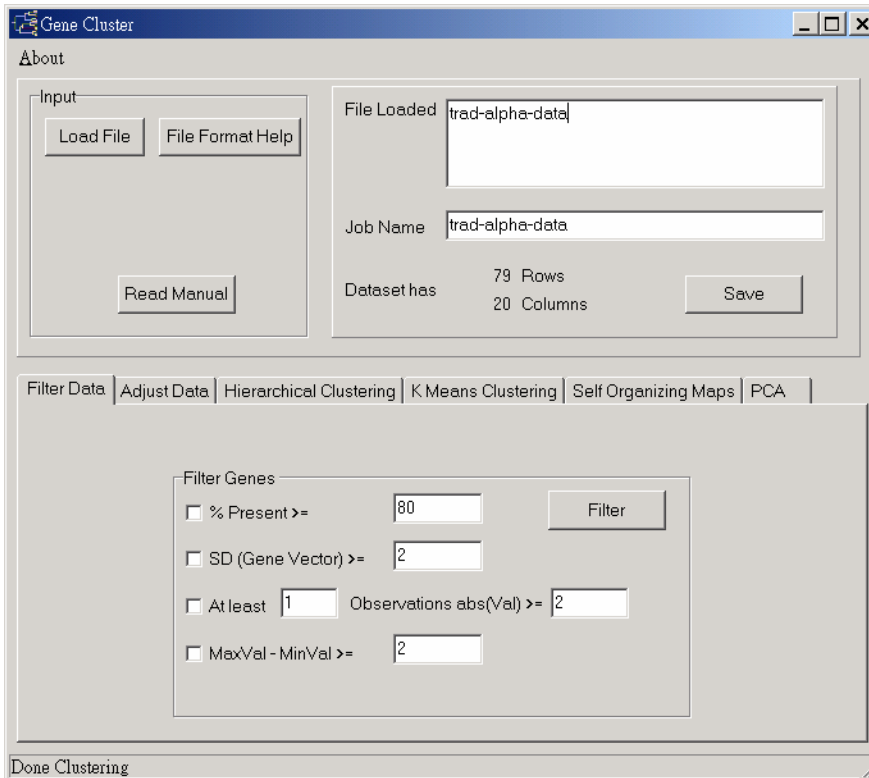
53 / 61

- Cluster and TreeView
- Bioconductor: Limma, LimmaGUI, LimmaAffy, gclus
- PermutMatrix
- GAP (Generalized Association Plots)

- GeneSpring GX v7.3

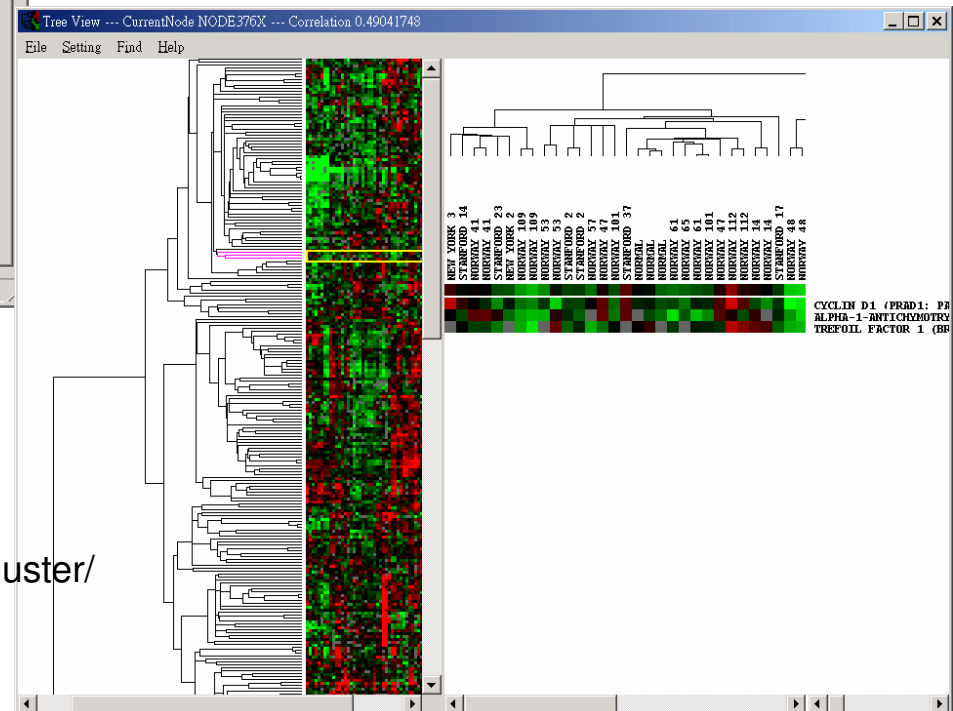
Cluster and TreeView

54 / 61



<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.



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](#)

The Bioconductor

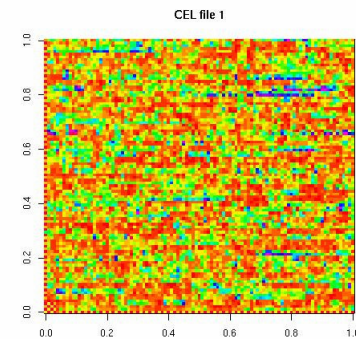
version 1.6

<http://www.bioconductor.org>

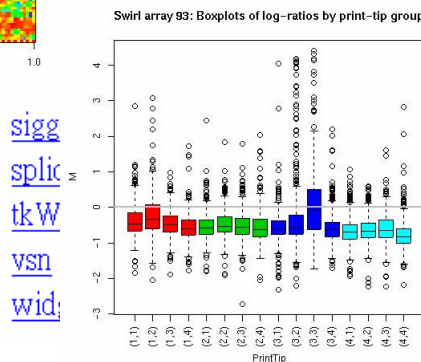
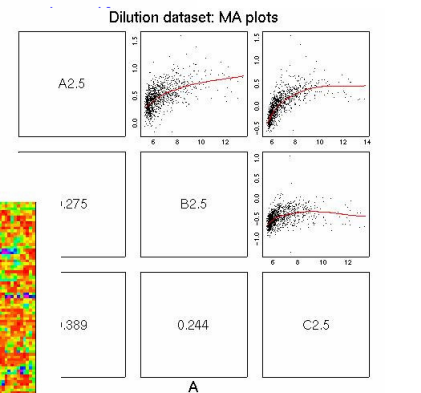


R version 2.1.1 (2005-06-20)

<http://www.r-project.org>

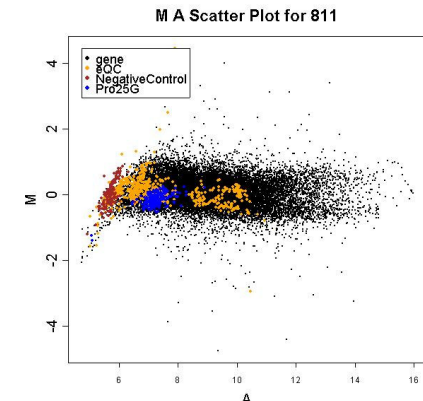
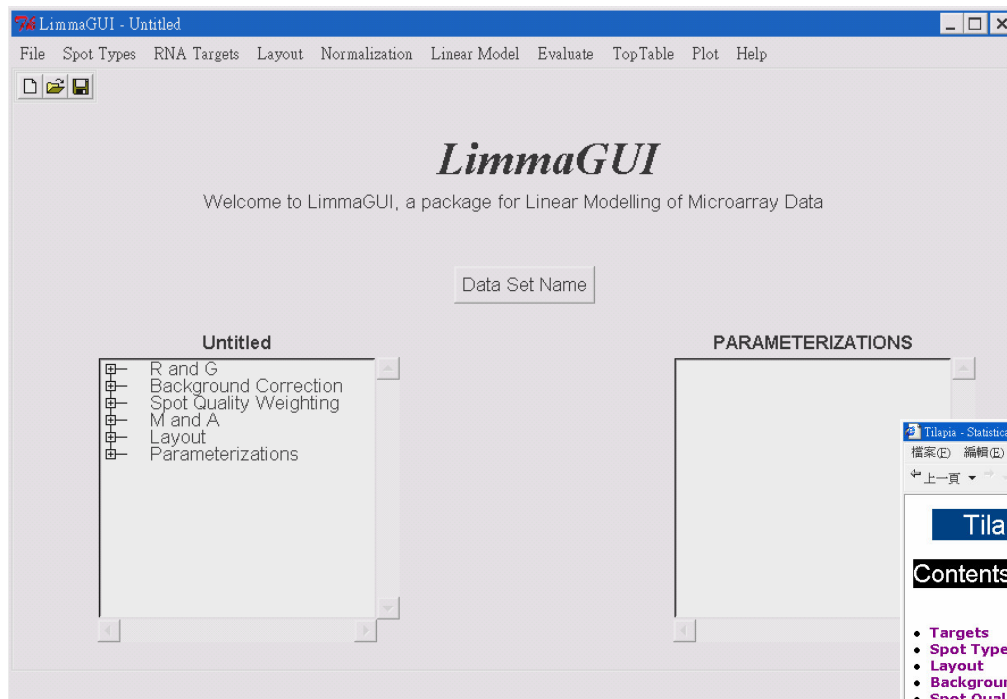


[daMA](#)
[edd](#)
[externalVector](#)
[factDesign](#)
[gcrma](#)



Limma, LimmaGUI, AffyLmGUI

56 / 61



Contents

- Targets
- Spot Types
- Layout
- Background Correction
- Spot Quality Weighting
- Raw M A Plots
- Raw Print-Tip Group Loess M A Plots
- M Box Plot for each Slide
- Spot Types Included In Linear Model
- Normalization Used In Linear Model
- Design Matrix
- Complete Tables of Genes Ranked in order of Evidence for Differential Expression
- M A Plots (with fitted M values)

RNA Targets

SlideNumber	Name	FileName	Cy3	Cy5
1	T060404	a0604_060404_TilapiaGH_gpr_treatment_control		
2	T070704	a0707_070704_TilapiaGH_gpr_treatment_control		
3	T072004	a0720_072004_TilapiaGH_gpr_control_treatment		
4	T080504	a0805_080504_Tilapia_gpr_control_treatment		
5	T081204	a0812_081204_TilapiaGH_gpr_control_treatment		

Spot Types

SpotType	ID	Name	Color
1	gene	*	black
2	Blank	Blank	orange

Limma: Linear Models for Microarray Data

<http://bioinf.wehi.edu.au/limma/>

LimmaGUI: a menu driven interface of Limma

<http://bioinf.wehi.edu.au/limmaGUI>

- Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Chapter 23. (To be published in 2005)
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, No. 1, Article 3.

Gclus, PermutMatrix

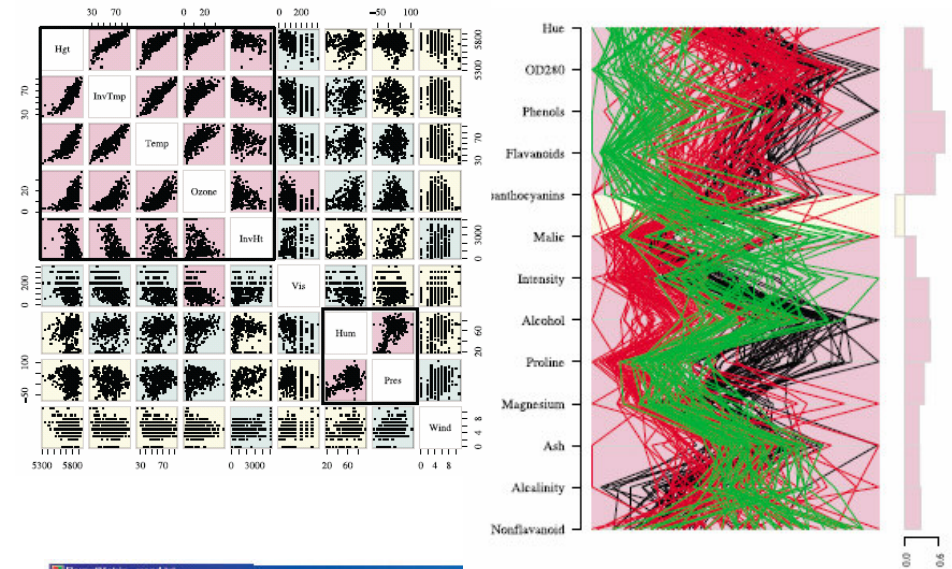
57 / 61

■ gclus: Clustering Graphics

(R package)

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

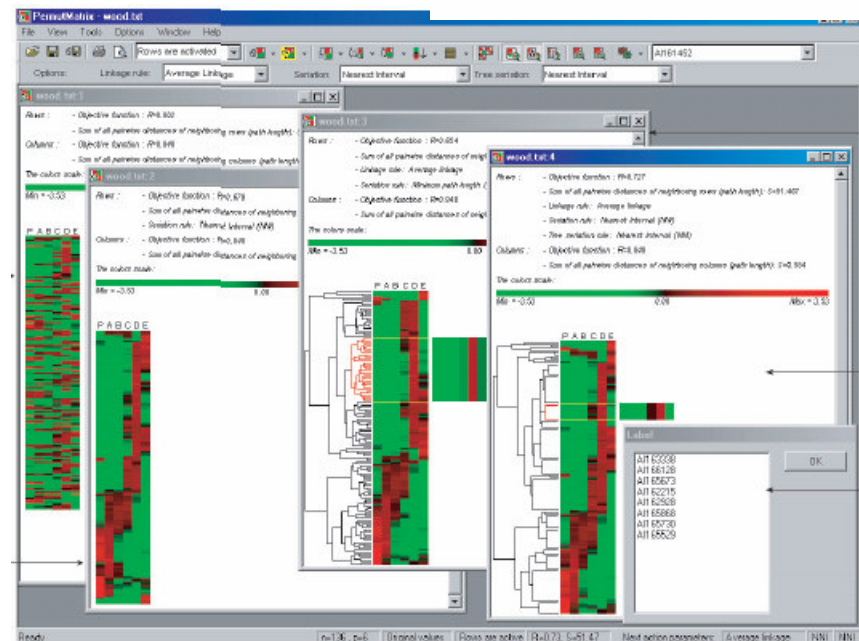
Catherine B. Hurley, (2004), Clustering Visualizations of Multidimensional Data, Journal of Computational & Graphical Statistics, Vol. 13, No. 4, pp.788-806



■ 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.



Generalized Association Plots

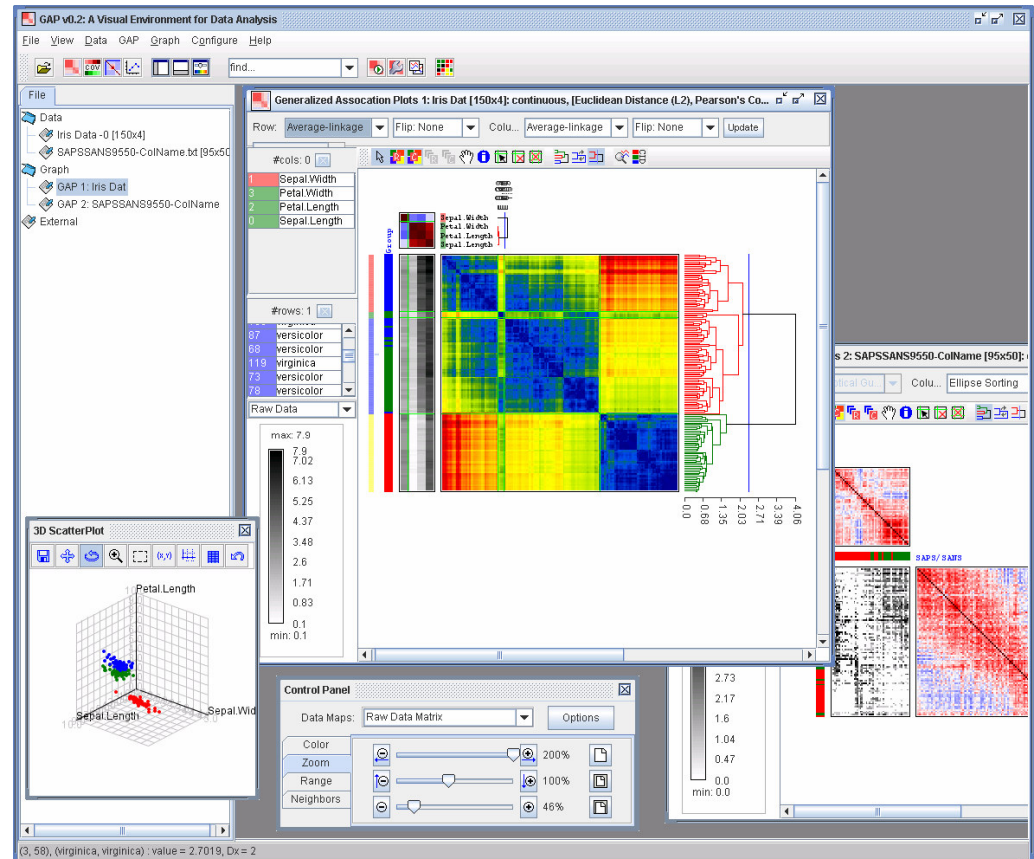
- Input Data Type: continuous or binary.
- Various seriation algorithms and **clustering analysis**.
- Various display conditions.
- Modules:
GAP with Covaraite Adjusted,
Nonlinear Association Analysis,
Missing Value Imputation.

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.

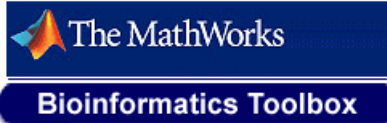
Wu, H. M., Tien, Y. J. and Chen, C. H. (2006). GAP: a Graphical Environment for Matrix Visualization and Information Mining.



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

Matlab: Bioinformatics ToolBox

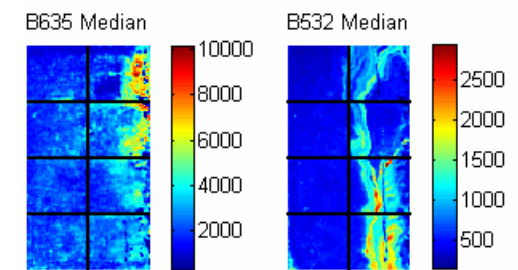
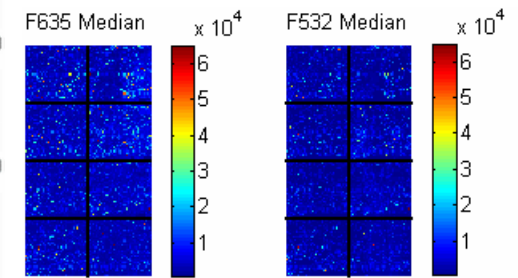
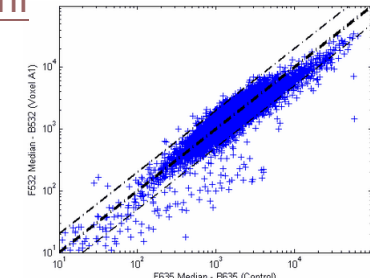
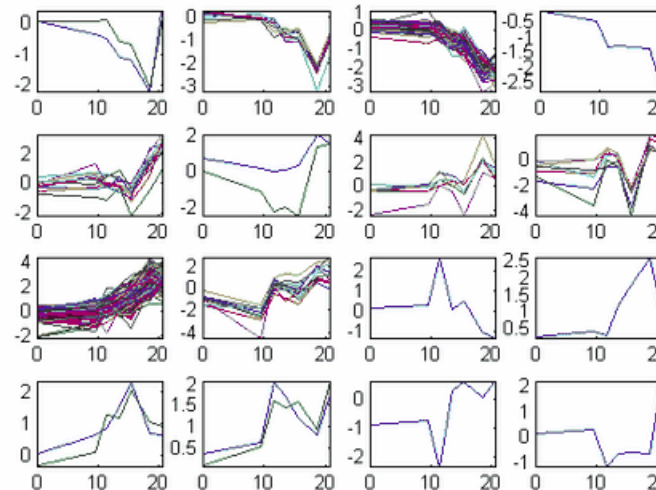
59 / 61



<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.

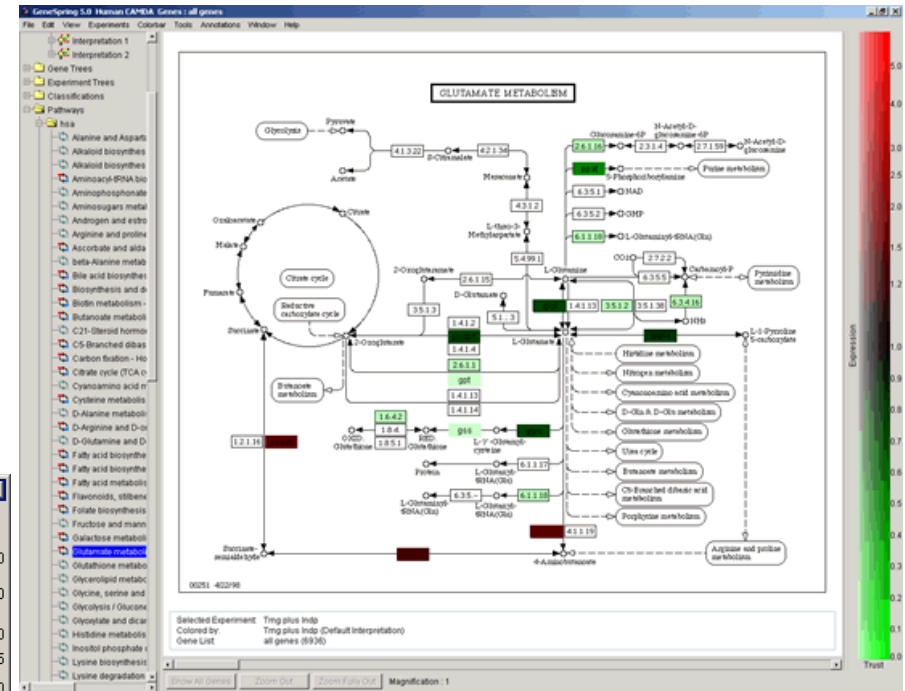
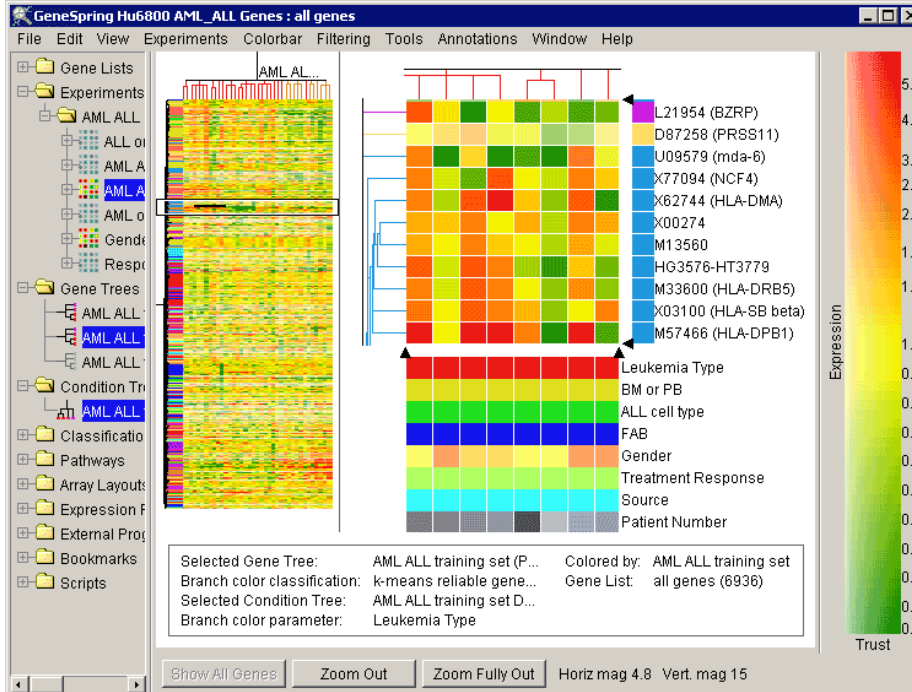
Hierarchical Clustering of Profiles



GeneSpring GX v7.3.1

60 / 61

- 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



Images from
<http://www.silicongenetics.com>



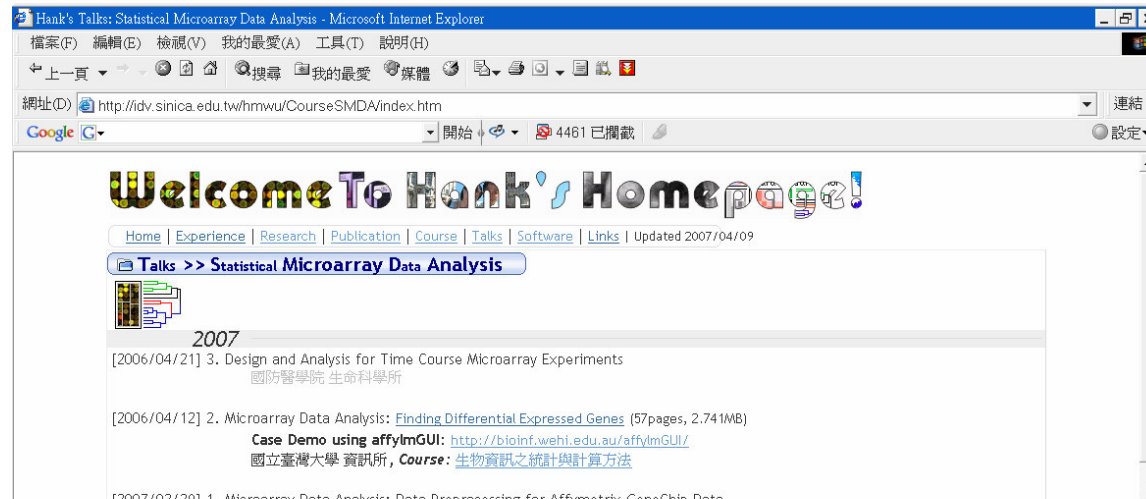
2004 Articles Citing GeneSpring®

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

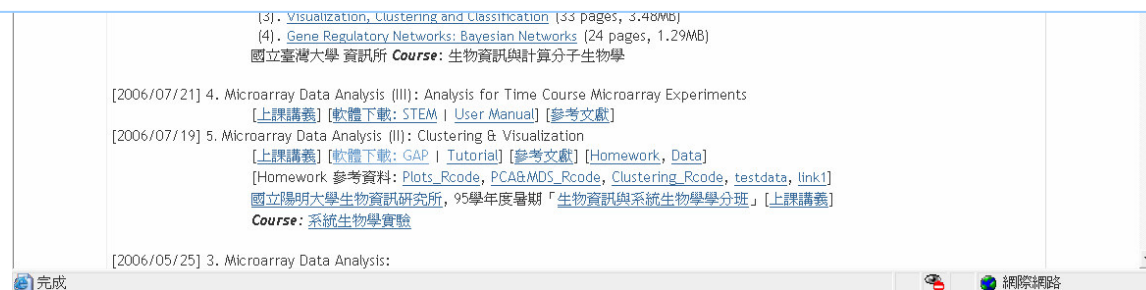
More than 700 papers

Questions?

61 / 61



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