陽明大學生物資訊研究所
2005 微陣列數據分析暑期課程

Microarray Data Analysis

Clustering and Visualization (II)

吳漢銘

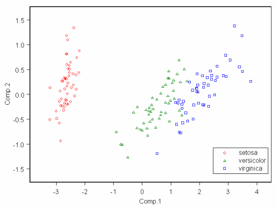
2005年7月29日

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

hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu

Outlines

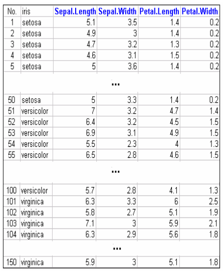
- Heat Map
- Hierarchical Clustering
    - Dendrogram
    - Single-linkage, complete-linkage, average-linkage, centroid-linkage, Ward's Method
- How Many Clusters?
- Generalized Association Plots (GAP)
- Generalization and Flexibility
- Visualization of Data Matrices
- Software: GAP



Dimension Free Data Visualization

全矩陣式資料視覺化

Data visualization techniques that can simultaneously visualize high dimensional (thousands) data structure without dimension reduction



No.	iris	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	setosa	5.1	3.5	1.4	0.2
2	setosa	4.9	3	1.4	0.2
3	setosa	4.7	3.2	1.3	0.2
4	setosa	4.6	3.1	1.5	0.2
5	setosa	5	3.6	1.4	0.2
...					
50	setosa	5	3.3	1.4	0.2
51	versicolor	7	3.2	4.7	1.4
52	versicolor	6.4	3.2	4.5	1.5
53	versicolor	6.9	3.1	4.9	1.5
54	versicolor	5.5	2.3	4	1.3
55	versicolor	6.5	2.8	4.6	1.5
...					
100	versicolor	5.7	2.8	4.1	1.3
101	virginica	6.3	3.3	6	2.5
102	virginica	5.8	2.7	5.1	1.9
103	virginica	7.1	3	5.9	2.1
104	virginica	6.3	2.9	5.6	1.8
...					
150	virginica	5.9	3	5.1	1.8

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.

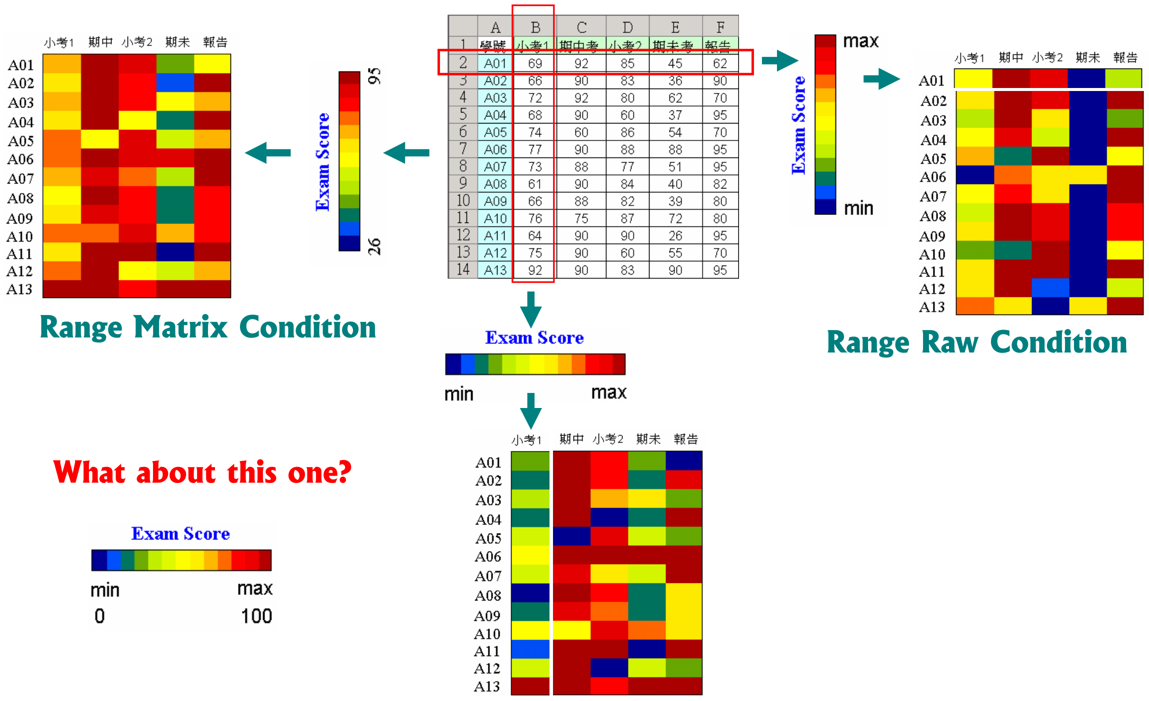
Data/Information Visualization

- 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.

Heat Map
(Data Image, Matrix Visualization)



	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

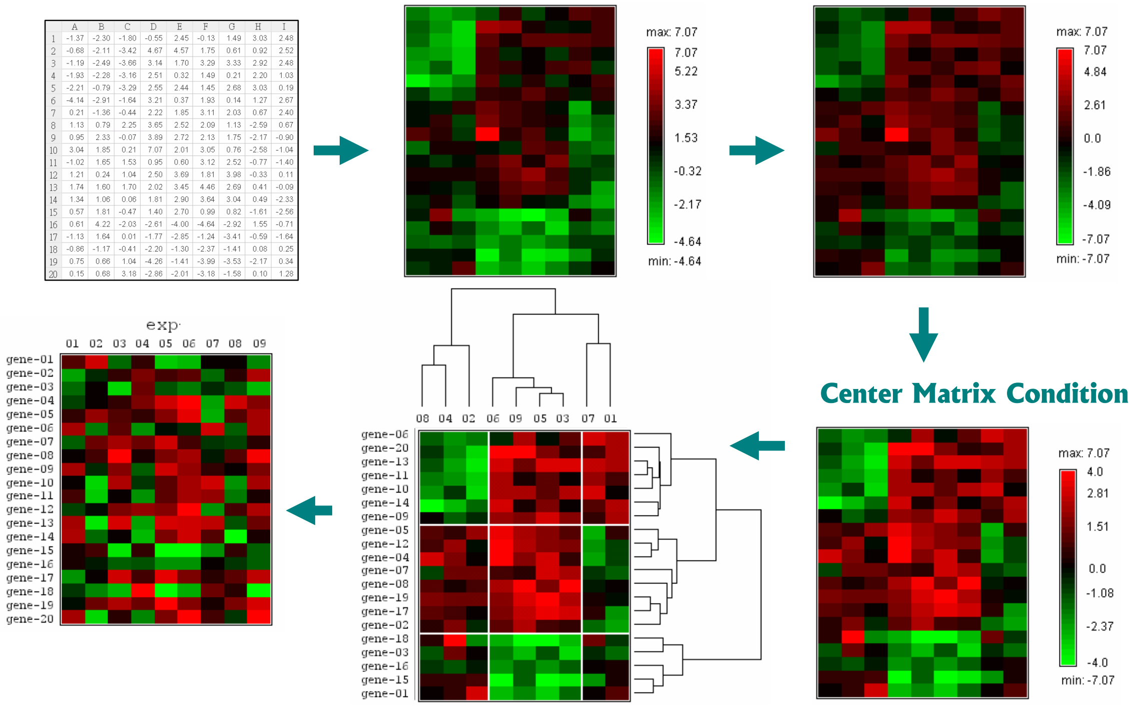
Range Matrix Condition

Range Raw Condition

What about this one?

Range Column Condition

Heat Map (conti.)



	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

Hierarchical Clustering

Hierarchical clustering can be perform using agglomerative and divisive approaches. 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, in each subsequent step a cluster is split off, until there are n clusters.
- **Agglomerative clustering:** all the objects start apart. There are n clusters at step 0, each object forms a separate cluster. In each subsequent step two clusters are merged, until only cluster is left.

Non-Hierarchical clustering

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

Hierarchical Clustering and Dendrogram

(Kaufman and Rousseeuw, 1990)

Example:

Average-Linkage

	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

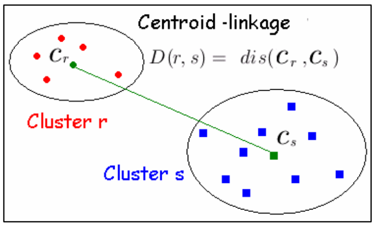
↓

	{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

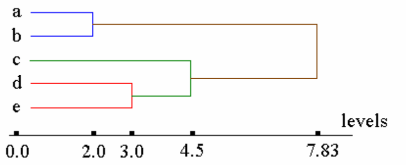
UPGMC (Unweighted Pair-Groups Method Centroid)

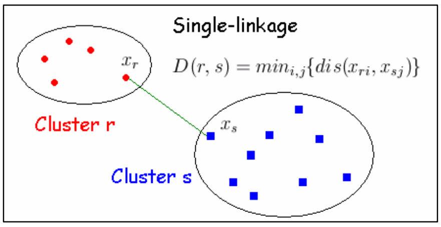


$$D(r,s) = dis(C_r, C_s)$$

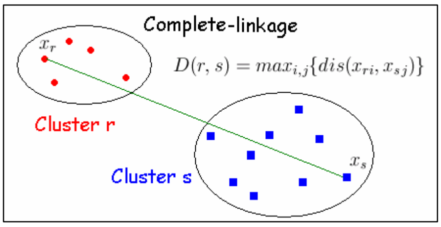
$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)]$$
$$= \frac{1}{2}(6+5) = 5.5$$

$$D(\{a,b\},\{d,e\})$$
$$= \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$
$$= \frac{1}{4}(10+9+9+8) = 9$$



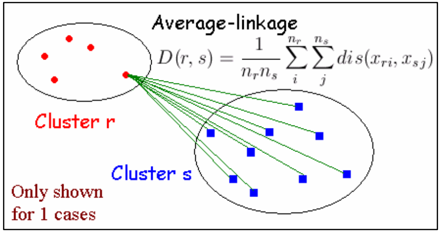


$$D(r,s) = min_{i,j}\{dis(x_{ri}, x_{sj})\}$$



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

UPGMA (Unweighted Pair-Groups Method Average)



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

Only shown for 1 cases

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-clustersums of squares.
- The within-cluster sums of squares that is minimized is also known as the error sums of squares (ESS).

Example:

Charles H. Romesburg (1984)

Toy Data

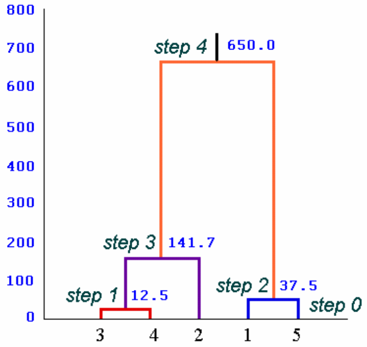
data	x1	x2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

step	Possible Partitions				ESS
1	(12)	3	4	5	?

$$\{\bar{12}\} = [\ (10+20)/2, (5+20)/2\ ]$$
$$= [\ 15, 12.5]$$

$$\begin{aligned} ESS &= wss\{12\} + wss\{3\} + wss\{4\} + wss\{5\} \\ &= ss(1, \{\bar{12}\}) + ss(2, \{\bar{12}\}) \\ &= (10-15)^2 + (5-12.5)^2 + (20-15)^2 + (2-12.5)^2 \\ &= 162.5 \end{aligned}$$

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



Exercise

Toy Data

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

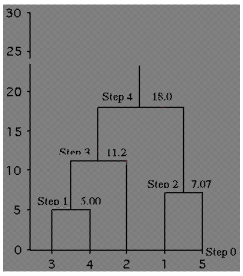
0.00				
18.00	0.00			
20.60	14.10	0.00		
22.40	11.20	5.00	0.00	
7.07	18.00	25.00	25.50	0.00

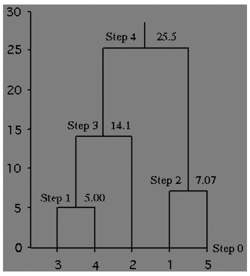
Single-Linkage

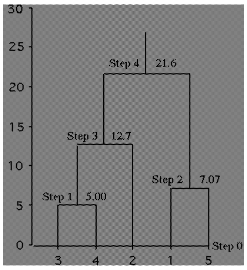
Complete-Linkage

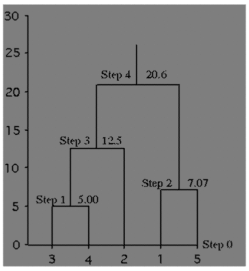
Average-Linkage

Centroid-Linkage

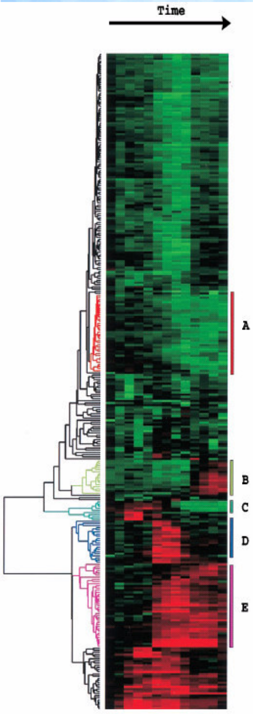








Display of Genome-Wide Expression Patterns



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

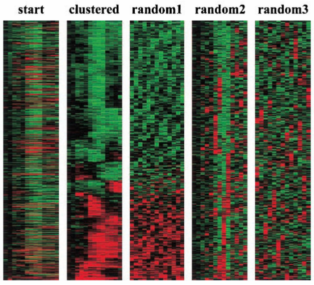
Cluster analysis and display of genome-wide expression patterns

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).



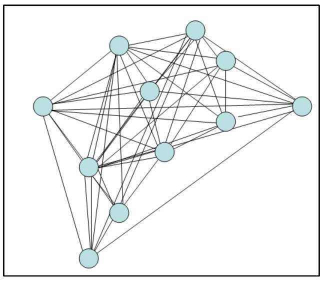
How Many Clusters?

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

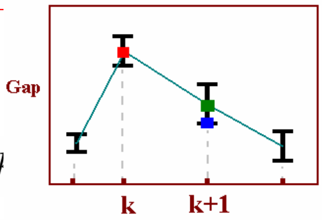


Within-Cluster Sum of Squares

$$D_r = \sum_{i\in C_r}\sum_{j\in C_r}\|x_i - x_j\|^2$$

$$W_k = \sum_{r=1}^{k}\frac{1}{2n_r}D_r$$

$\mathrm{Gap}_n($



): CH(k)

KL(k)

90): $s(i)$

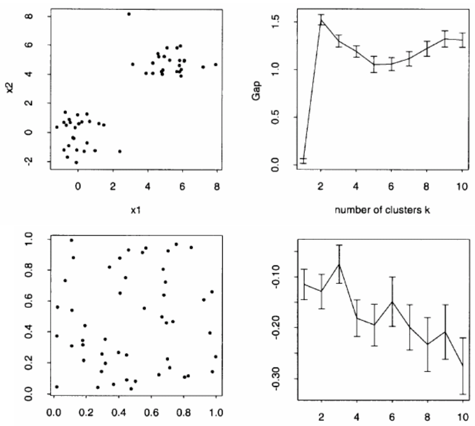
(W_k)

Computational Implementation

choose the number of clusters via

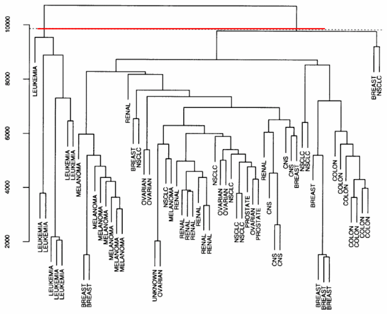
$$\hat{k} = \text{ smallest } k \text{ such that}$$

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$$

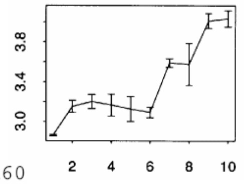


application to hierarchical clustering and DNA microarray data

6834×64 matrix



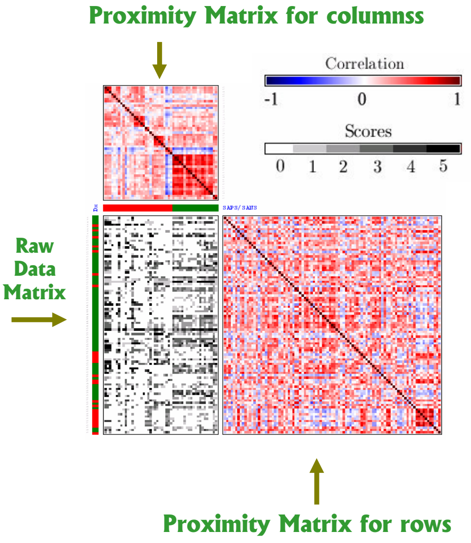
http://www-genome.stanford.edu/nci60

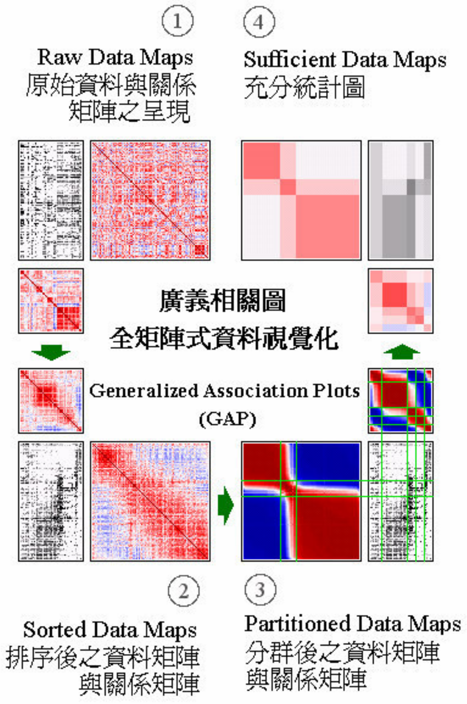


Generalized Association Plots (GAP)

(Chen, 2002)

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

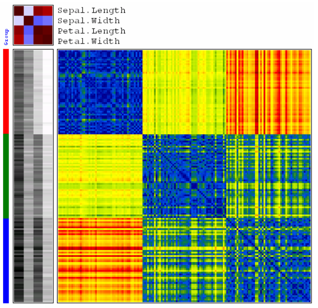


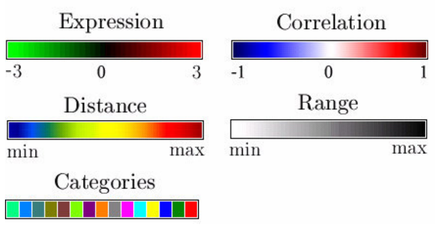


Presentation of Raw Data Matrix

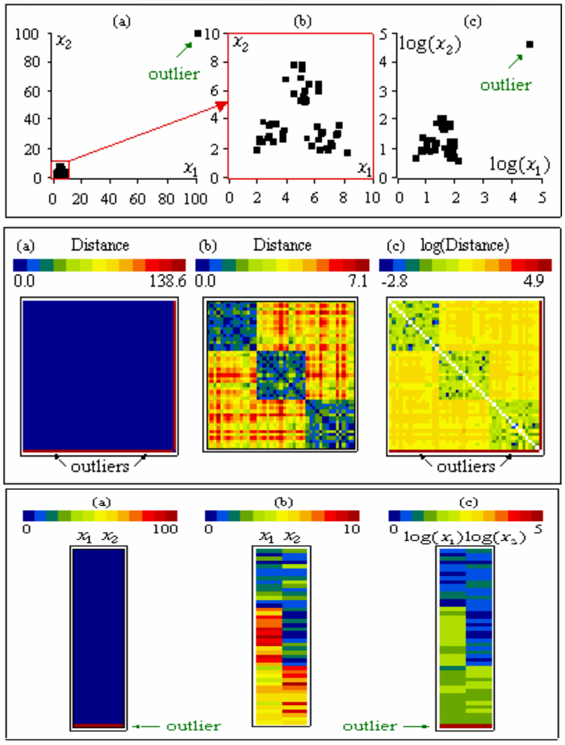
Image source: Dr. Chen Chun-houh's Silde

1. Color spectrum
2. Variable transformation





"Resolution" of a Statistical Graph



Selection of Proximity Measures

Dissimilarity/Similarity Measure for Quantitative Data

Dissimilarity	Formula
Minkowski	$d(i,j)=\left(\sum_k \lvert x_{ik}-x_{jk}\rvert^p\right)^{1/p}$
Canberra	$d(i,j)=\sum_k \lvert x_{ik}-x_{jk}\rvert \big/ \lvert x_{ik}+x_{jk}\rvert$
Soergel	$d(i,j)=\sum_k \lvert x_{ik}-x_{jk}\rvert \big/ \sum_k \max(x_{ik},x_{jk})$
Divergence	$d(i,j)=\sum (x_{ik}-x_{jk})^2 \big/$
Bary-Curtis	$d(i$
Wave-Hedges	$d(i$
Bhattacharyya	$d(i$

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

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

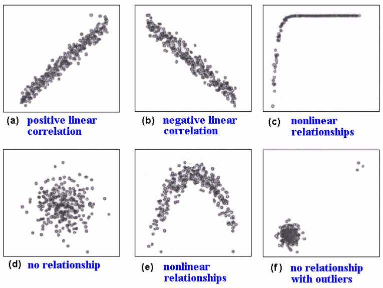
- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

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

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

Kendall's tau

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


(a) positive linear correlation (b) negative linear correlation (c) nonlinear relationships

(d) no relationship (e) nonlinear relationships (f) no relationship with outliers

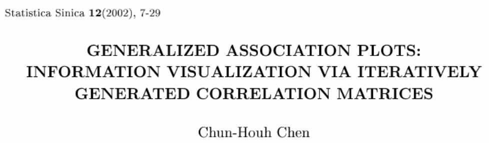
Data	Pearson's rho	Spearman's rho	Kendall's tau
(a)	0.98	0.98	0.87
(b)	-0.98	-0.98	-0.87
(c)	0.50	0.99	0.98
(d)	-0.02	-0.03	-0.02
(e)	-0.06	-0.02	-0.02
(f)	0.68	0.00	0.00

- ■ 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).

Concept of Relativity of a Statistical Graph

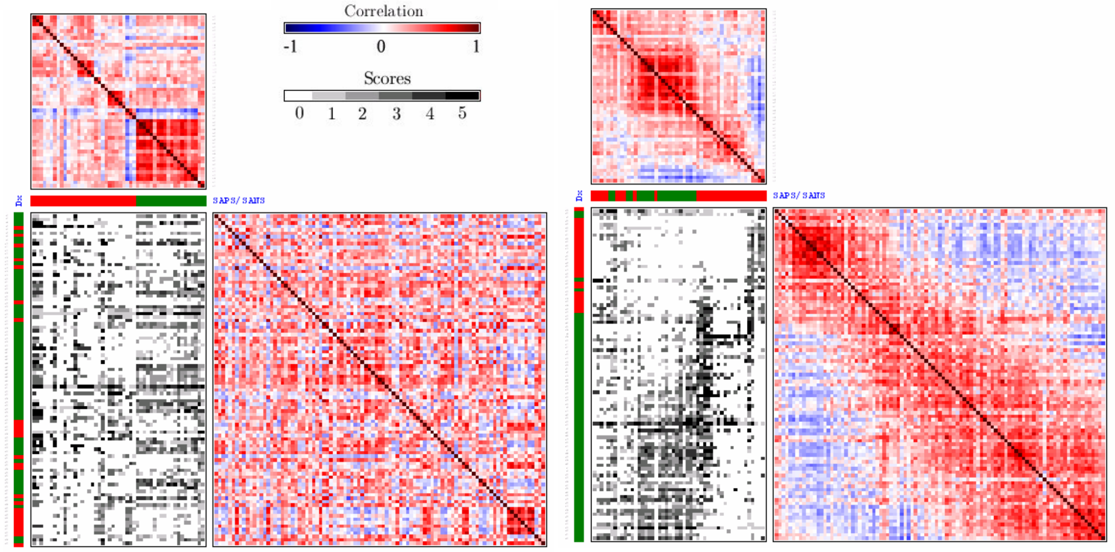
Placing similar (different) objects at closer (distant) positions



Statistica Sinica **12**(2002), 7-29

GENERALIZED ASSOCIATION PLOTS:
INFORMATION VISUALIZATION VIA ITERATIVELY
GENERATED CORRELATION MATRICES

Chun-Houh Chen



Seriation Problem

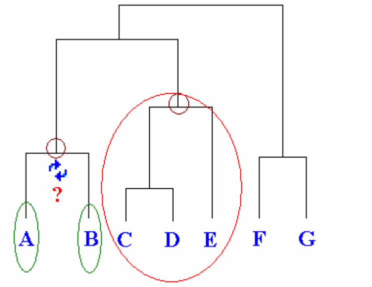
Cluster Software (Eisen et al 1998):

(1) Based on average expression level

(2) Using the results of a one-dimensional SOM

Alon et al (1999):

Based on similarity to their parent's siblings

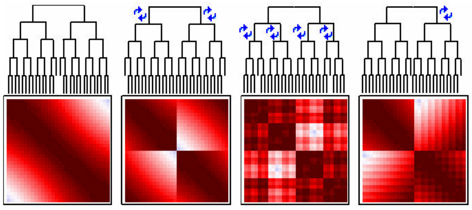


if d(A, {C,D, E}) < d(B, {C,D, E}) then flip

Bar-Joseph et al (2001)

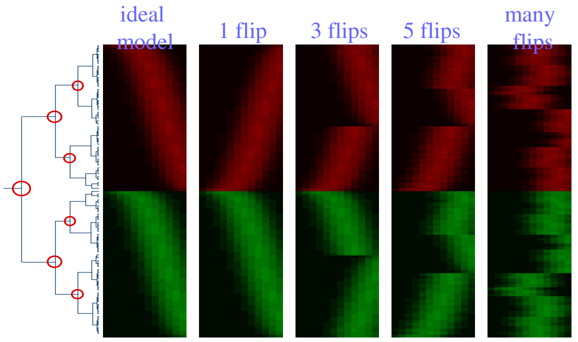
Ziv Bar-Joseph, David K. Gifford, and Tommi S. Jaakkola, (2001), **Fast Optimal Leaf Ordering** for Hierarchical Clustering. Bioinformatics 17(Suppl. 1):S22–S29.

Tree seriation for proximity matrices



Different Seriations

Generated from Identical Tree Structure

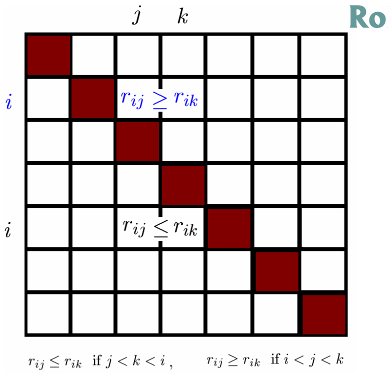


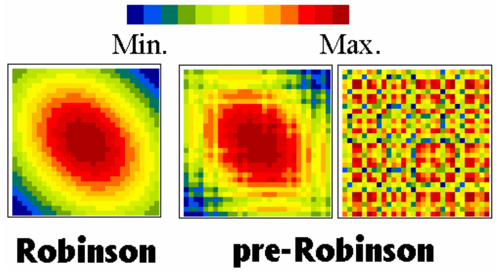
Tree seriation for raw data matrices

Criteria for a "good" Permutation

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

Robinson Form


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



Global/local Criterion:

Anti-Robinson Measurements

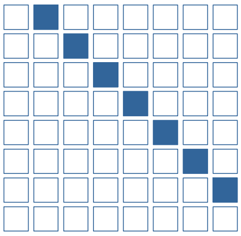
permuted matrix, $D = [d_{ij}]$

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

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

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

Local criterion: Minimal Span Loss Function



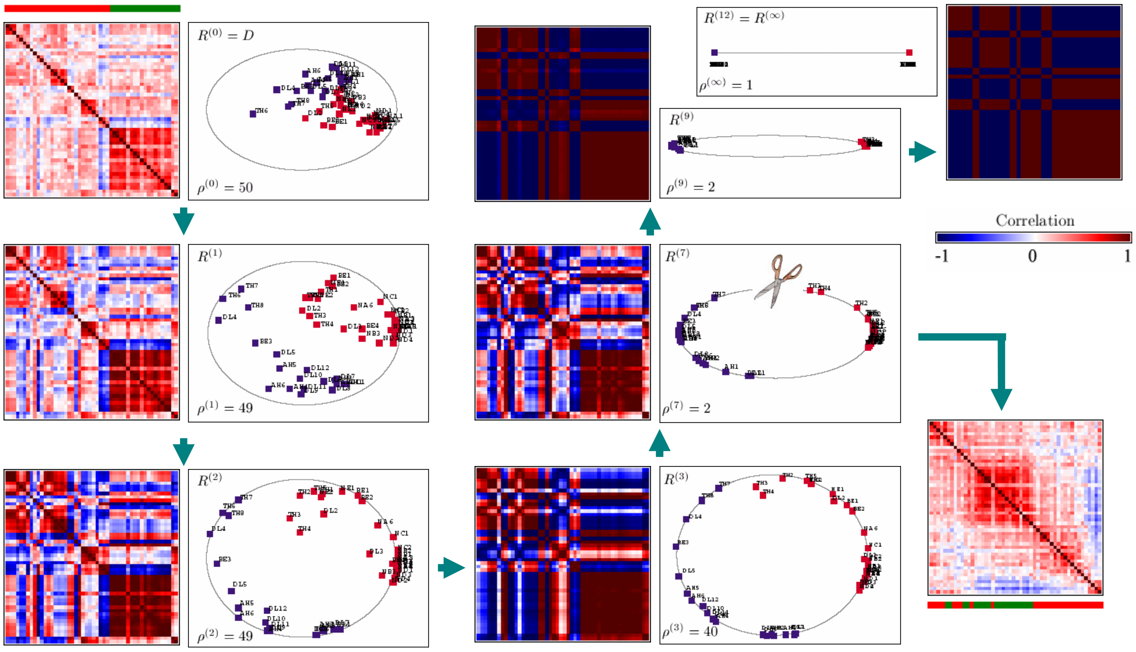
$$MS = \sum_{i=1}^{n-1} d_{i,i+1}$$

Further Reading

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

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

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

GAP Elliptical Seriation — Michael Eisen Tree Seriation

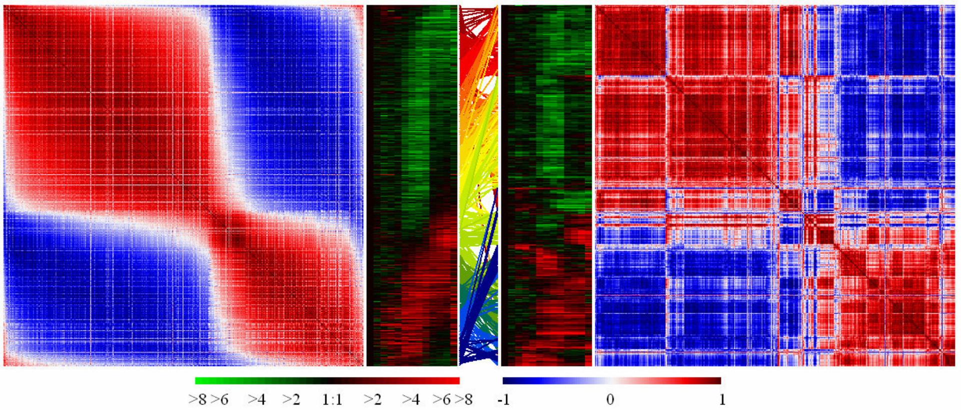
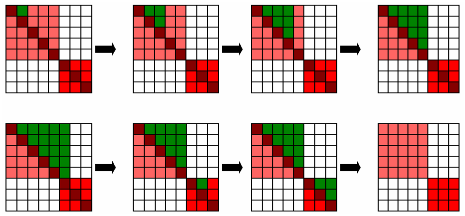


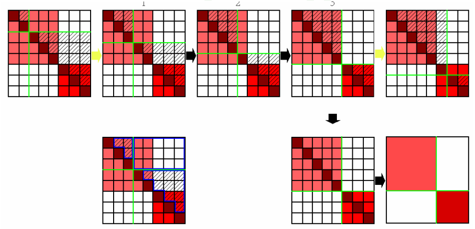
Image source: Dr. Chen Chun-houh's slide

Partitions of Permuted Matrix Maps

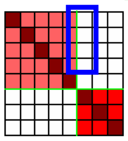
One-Way block Searching

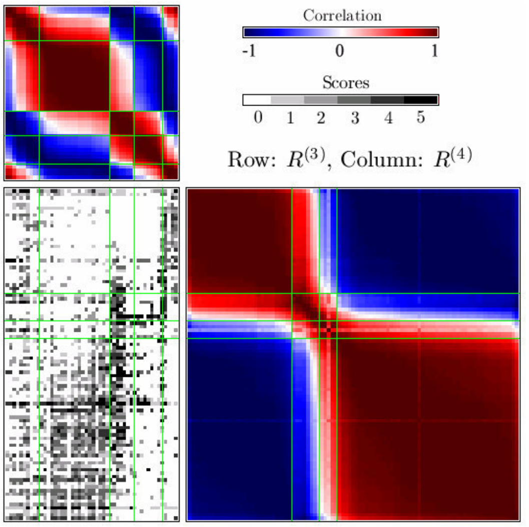


Within-Sum-of-Square Approach

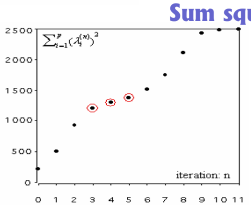


Two-Sample Problem



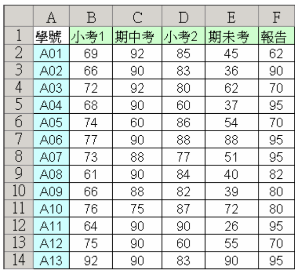


Sum squared eigenvalues (sum squared correlations)



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

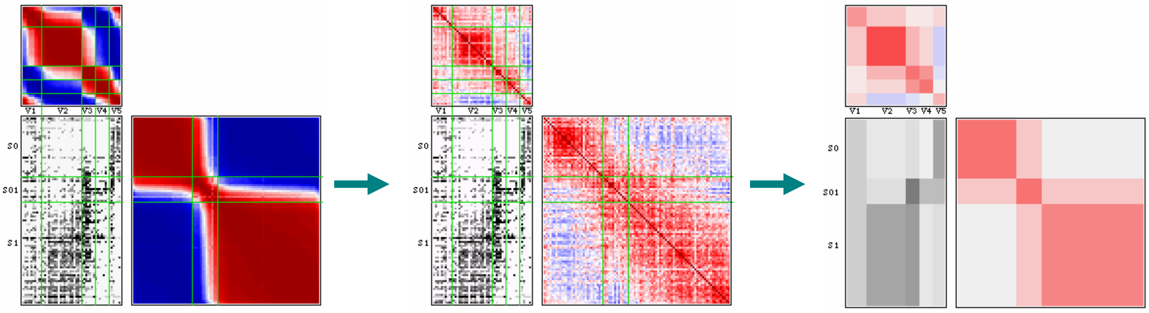


	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

Sufficient Statistic

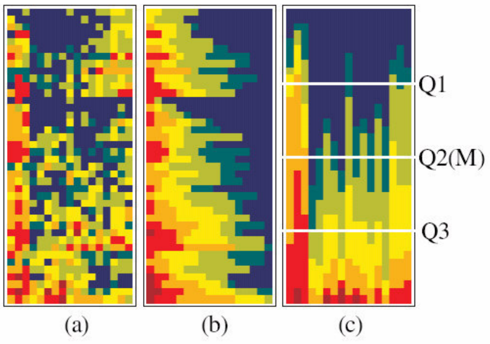
	小考1	期中考	小考2	期末考	報告
平均	71.77	86.54	80.38	53.46	83

70		小考1	期中考	小考2	期末考	報告
	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



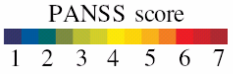
Generalization and Flexibility

Sedimented MV for patients and symptoms.



(a) (b) (c)

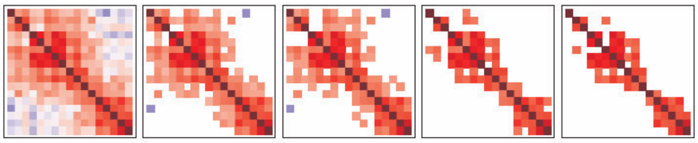
Image source: Chen etal 2004



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.

Sectional MV for the permuted correlation coefficient map.



p-value < 1   < 0.10   < 0.05   < 0.01   < 0.005

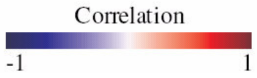


Image source: Chen etal 2004

Visualization of Data Matrices

Simple ← Information Visualization of Data Matrices → Difficult

Continuous	Ordinal	Binary	Categorical
(Gene/Time)	(Patient/Symptom)	(Mouse/Tumor)	(Subject/SNP)

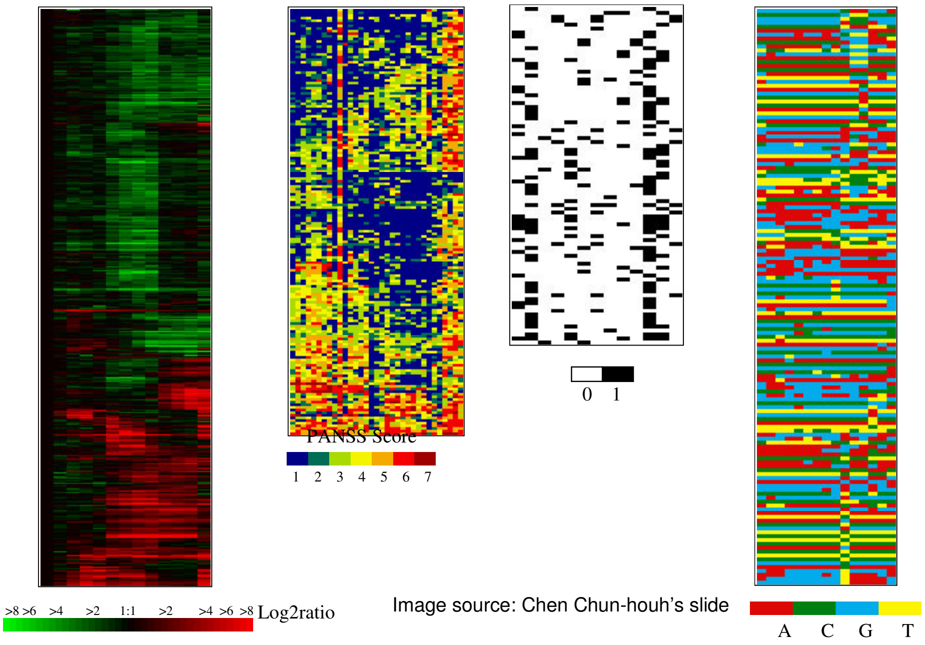
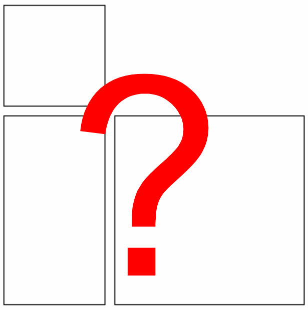


Image source: Chen Chun-houh's slide

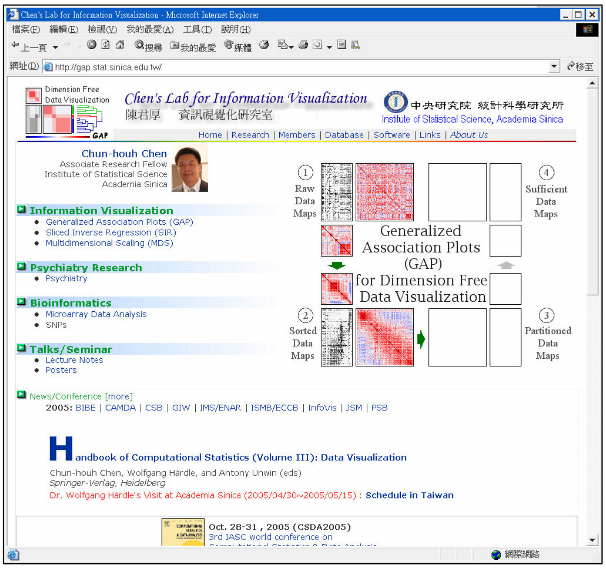
Chen's Lab for Information Visualization



- 類別型 (categorical) 資料之全矩陣視覺化
- 條件式（變項校正）全矩陣視覺化
- 全矩陣視覺化之遺漏值 (missing value) 處理
- 多時點（相同變項）資料之全矩陣視覺化
- 多條件（不同變項）資料之全矩陣視覺化
- 相依 (dependent) 或群集 (clustered) 資料之全矩陣視覺化
- 巨量資料之全矩陣視覺化

Web site

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


Chen's Lab for Information Visualization - 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://gap.stat.sinica.edu.tw/
移至
網際網路

Dimension Free Data Visualization GAP

Chen's Lab for Information Visualization
陳君厚 資訊視覺化研究室

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

Home | Research | Members | Database | Software | Links | About Us

Chun-houh Chen
Associate Research Fellow
Institute of Statistical Science
Academia Sinica

Information Visualization
- Generalized Association Plots (GAP)
- Sliced Inverse Regression (SIR)
- Multidimensional Scaling (MDS)

Psychiatry Research
- Psychiatry

Bioinformatics
- Microarray Data Analysis
- SNPs

Talks/Seminar
- Lecture Notes
- Posters

① Raw Data Maps
② Sorted Data Maps
③ Partitioned Data Maps
④ Sufficient Data Maps

Generalized Association Plots (GAP) for Dimension Free Data Visualization

News/Conference [more]

2005: BIBE | CAMDA | CSB | GIW | IMS/ENAR | ISMB/ECCB | InfoVis | JSM | PSB

Handbook of Computational Statistics (Volume III): Data Visualization

Chun-houh Chen, Wolfgang Härdle, and Antony Unwin (eds)
Springer-Verlag, Heidelberg
Dr. Wolfgang Härdle's Visit at Academia Sinica (2005/04/30~2005/05/15) : **Schedule in Taiwan**

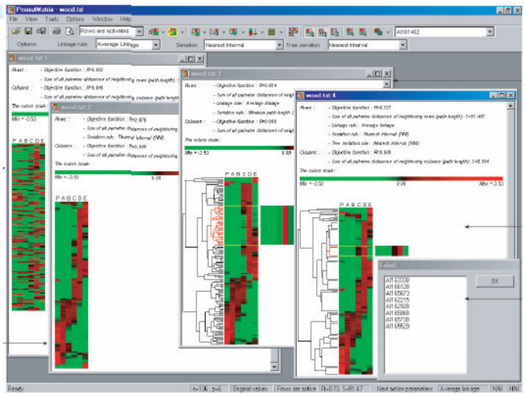
Oct. 28-31 , 2005 (CSDA2005)
3rd IASC world conference on

Software

- **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.

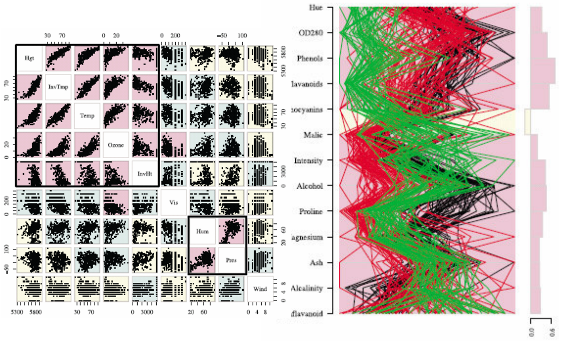


- **gclus: Clustering Graphics**

  (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



Software: GAP

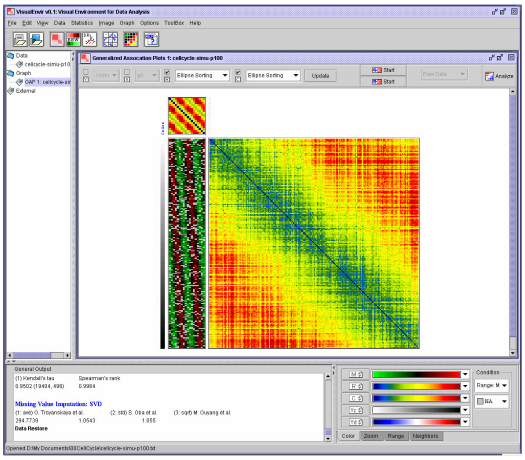
- **Generalized Association Plots**
  - Various seriation algorithms (Clustering Analysis)
  - Various display conditions
- **GAP with a Covaraite Adjusted**
  - Within And Between Analysis (WABA).
  - Partial Correlation Analysis.
- **GAP with Nonlinear Association Analysis**
  - ISOMAP
  - Kernel Transformation
- **GAP with Missing Values Imputation**
  - Row means, Columns means
  - Regression methods
  - KNN (KNNImpute)
  - SVD (SVDImpute)
  - GAPImpute
- **Statistical Plots**
  - Histogram, 2D Scatterplot, 3D Scatterplot (Rotatable)

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

Expected to release on 15th Dec, 2005.

iasc

Dec. 15-17, 2005 (IASC-ARS 2005)
The 5th Asian Conference on Statistical Computing, IASC
The University of Hong Kong, Hong Kong


VisualEnvir v0.1: Visual Environment for Data Analysis
File Edit View Data Statistics Image Graph Options ToolBox Help
Data
cellcycle-simu-p10
Graph
GAP 1: cellcycle-si
External
Generalized Assocation Plots 1: cellcycle-simu-p100
Order
g0
Ellipse Sorting
Ellipse Sorting
Update
Start
Start
Raw Data
Analyze
Order
M
R
C
Yc
Yd
Condition
Range: M
NA
Color Zoom Range Neighbors

```
General Output
(1) Kendall's tau        Spearman's rank
0.9502 (19404, 496)      0.9964

Missing Value Imputation: SVD
(1: ave) O. Troyanskaya et al.     (2: std) S. Oba et al.     (3: sqrt) M. Ouyang et al.
294.7739                 1.0543               1.055
Data Restore
```

Opened D:\My Documents\00CellCycle\cellcycle-simu-p100.txt

Reference

- Chen, C. H. (2002), Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices, Statistica Sinica, 12, 7-29.
- Chen, C. H., Hwu, H. G., Jang, W. J., Kao, C. H., Tien, Y. J., Tzeng, S., and Wu, H. M. (2004). “Matrix Visualization and Information Mining,” Proceedings in Computational Statistics 2004 (Compstat 2004), 85-100, Physika Verlag, Heidelberg.
- Hartigan, J. (1972), Direct Clustering of a Data Matrix. Journal of the American Statistical Association, 67(337):123-129.
- Hartigan, J. (1975), Clustering Algorithms, John Wiley and Sons, New York.
- Jacoby, W. G. (1998), Statistical Graphics for Visualizing Multivariate Data, Thousand Oaks, Calif. : Sage Publications.
- Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
- Minnotte , M. C. and West, R. W., (1999), "The Data Image: a Tool for Exploring High Dimensional Data Sets,". 1998 Proceedings of the ASA Section on Statistical Graphics, in press.
- Jain, A.K., Murty M.N., and Flynn P.J. (1999): Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3, 264-323. http://citeseer.ist.psu.edu/jain99data.html