

Introduction to **Generalized Association Plots (GAP)**

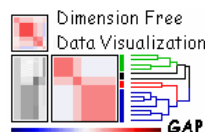
for Dimension-Free Data Visualization

Han-Ming Wu

**Institute of Statistical Science, Academia Sinica,
Taipei, Taiwan**

hmwu@stat.sinica.edu.tw

<http://www.sinica.edu.tw/~hmwu/>



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

2006/12/22

Outlines



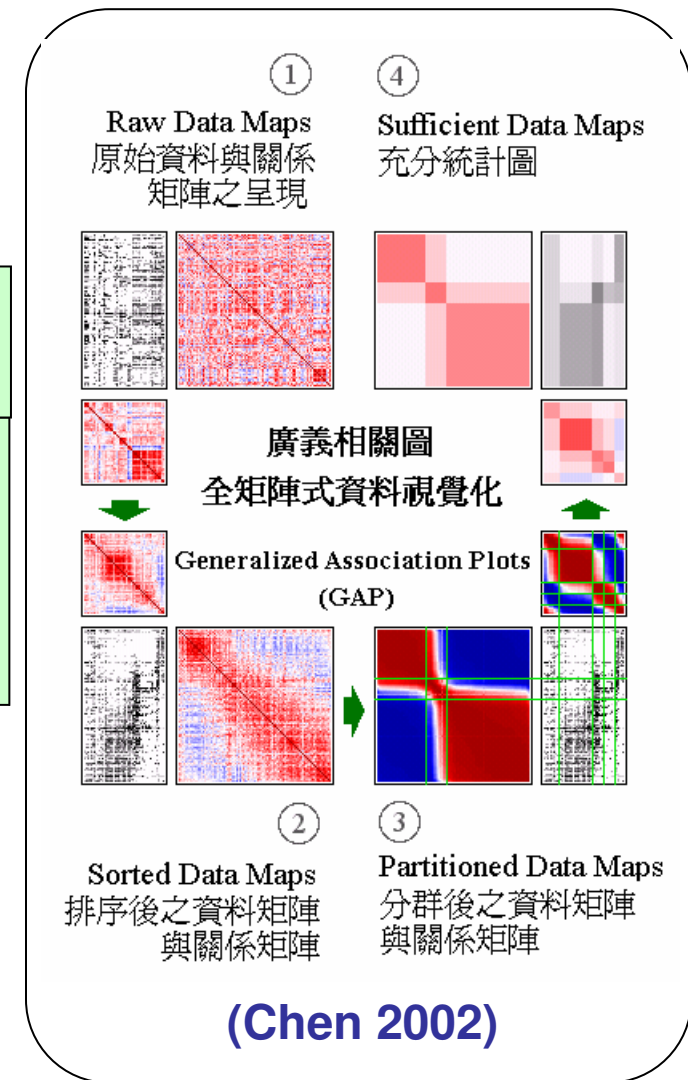
- Two Demo Datasets
- Four Steps of Generalized Association Plots (GAP)

Raw Data Matrix and Two Proximity Matrices

Presentation 呈現 → **Seriation** 排序 → **Partition** 分割 → **Sufficient** 充分

- Generalization and Flexibility
- Modules/Software/Conclusion

NOTE: Matrix Visualization (MV): reorderable matrix, the heatmap, color histogram, data image and matrix visualization.



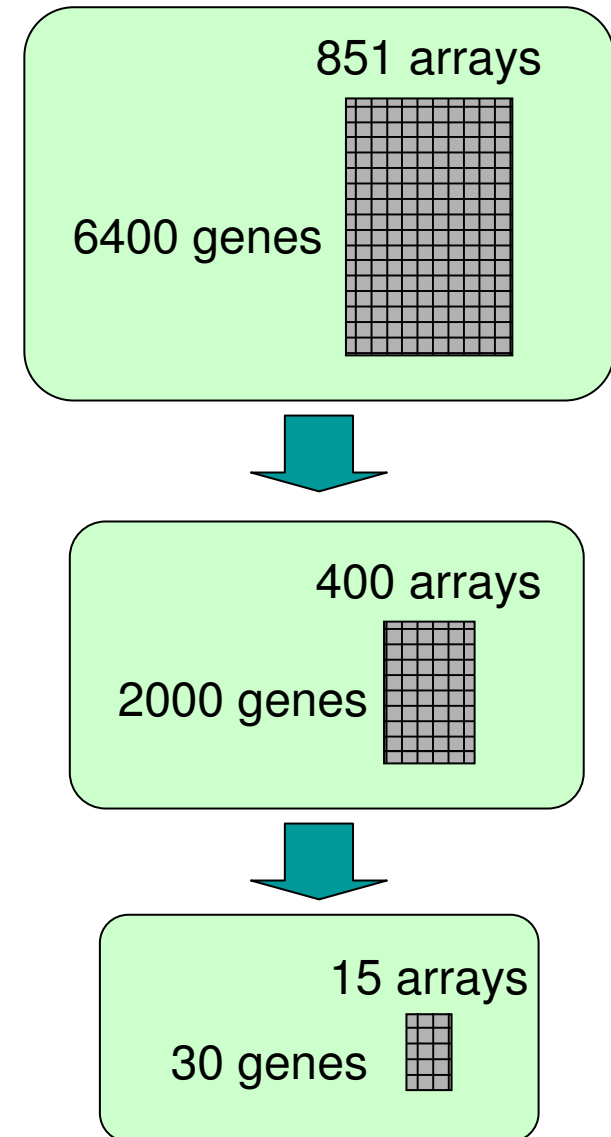
Microarray Gene Expression Data of Yeast Cell Cycle



yMGV: yeast Microarray Global Viewer



<http://transcriptome.ens.fr/ymgv/>



Psychosis Disorder Data

4/28

Hallucinations (AH1-6)

Behavior (BE1-4)

Delusions (DL1-12)

Thought disorder (TH1-8)

The Andreasen's Positive and Negative Symptom Table

Scale for Assessment of Positive Symptoms
(**SAPS**): **30 items**,
4 subgroups.

Scale for Assessment of Negative Symptoms
(**SANS**): **20 items**,
5 subgroups.

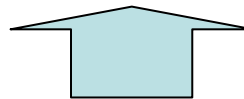
Expression (NA1-7)

Speech (NB1-4)

Hygiene (NC1-3)

Activity (ND1-4)

Inattentiveness (NE1-2)



69 schizophrenic

26 bipolar disorders

95
Subjects

50 Variables

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	
1	AH1	AH2	AH3	AH4	AH5	AH6	DL1	DL2	DL3	DL4	DL5	DL6	DL7	DL8	DL9	DL10	DL11	DL12	BE1	BE2	BE3	BE4	TH1	TH2	TH3	TH4	TH5	TH6	TH7	TH8	NA1		
2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
3	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
4	5	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
5	5	5	5	4	5	5	0	5	5	5	1	5	5	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5		
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
7	2	0	2	0	0	2	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	0	
8	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
9	3	1	1	0	0	2	0	0	0	0	1	3	1	3	0	0	2	5	0	3	3	0	0	0	0	0	0	0	0	0	0	0	
10	5	5	1	0	0	4	0	0	0	0	0	3	4	4	0	4	0	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
12	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
13	5	0	1	0	0	5	0	0	4	5	0	0	0	0	0	0	0	0	1	3	4	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	5	0	5	0	0	0	0	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	0	0	0	0	0	0	3	1	5	5	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	5	5	5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	1	1	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	5	5	4	0	0	5	5	0	0	0	4	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	4	2	4	0	0	0	3	0	0	4	0	0	0	0	0	0	0	0	0	1	3	4	0	2	2	0	0	0	0	0	0	0	
22	3	3	3	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	4	3	3	4	0	0	0	0	0	0	
23	2	3	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	4	4	4	0	0	0	3	0	0	0	0	4	4	0	0	3	0	0	0	2	2	3	3	2	0	0	0	0	0	0	0	0	0
25	3	2	1	0	0	0	1	0	0	2	2	0	2	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	1	1	1	0	0	0	0	0	5	5	0	0	0	0	0	0	0	0	0	2	4	0	3	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
28	4	4	4	0	0	4	0	0	4	4	0	0	0	0	0	0	0	0	0	3	2	4	3	3	0	3	3	0	0	0	0	0	0
29	1	1	1	0	0	0	0	1	3	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
31	3	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	2	2	1	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	4	4	1	0	0	3	4	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	4	4	0	0	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	2	2	0	0	0	0	4	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	4	0	4	0	0	0	0	0	0	2	0	0	5	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Raw Data

Raw Data Matrix

All the symptoms are recorded on a six point scale (**0-5**).



The 1st Step of GAP

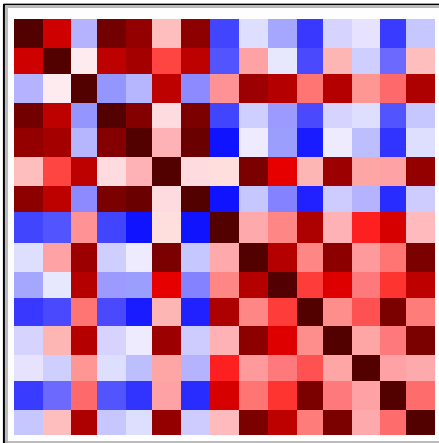
Presentation of Raw Data Matrix

- **Data Transformation**
- **Selection of Proximity Measures**
- **Color Spectrum**
- **Display Conditions**

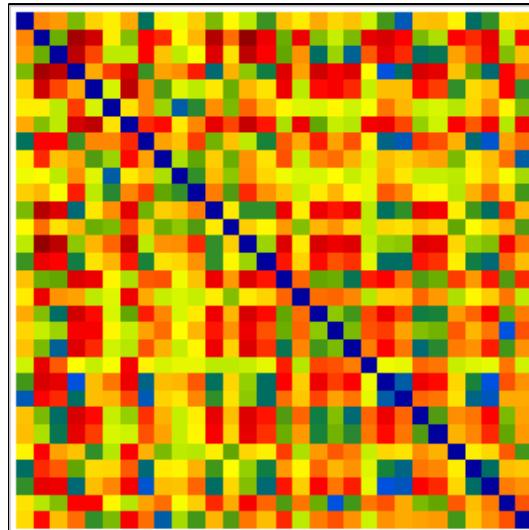
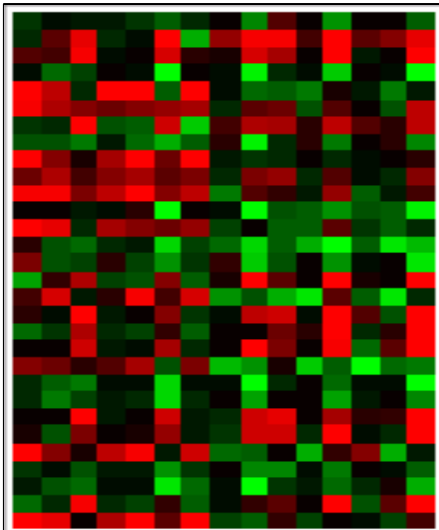
Presentation of Raw Data Matrix



Proximity Matrix
for Columns



Proximity Matrix
for Rows



(1) Selection of Proximity Measures

Pearson Correlation Coefficient

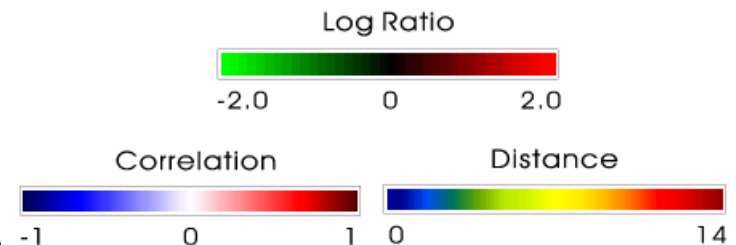
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Euclidean Distance

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

Other Similarity/Dissimilarity
Measures

(2) Color Spectrum



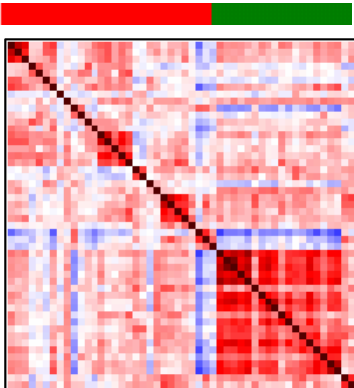
Presentation of Raw Data Matrix



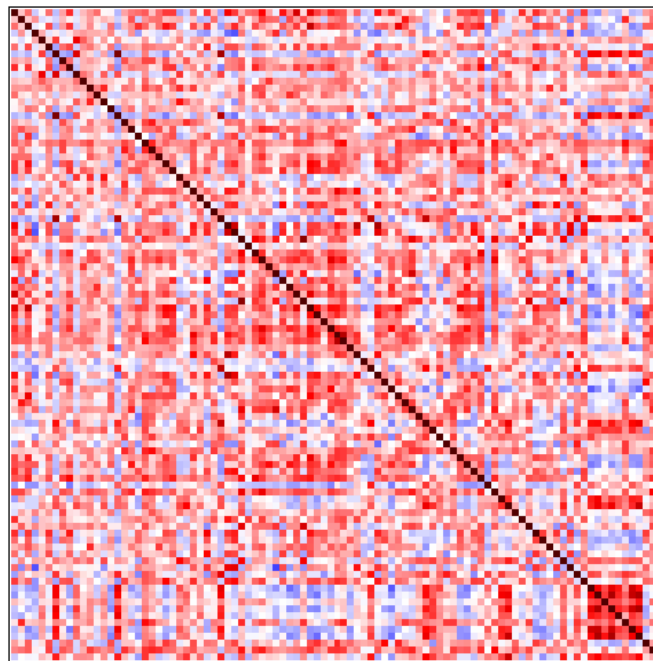
(1) Selection of Proximity Measures

Pearson Correlation Coefficient

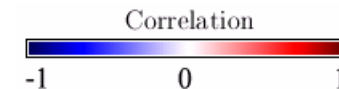
Correlation Matrix
for Variables



Correlation Matrix for Subjects



(2) Color Spectrum



Symptoms

■ SAPS

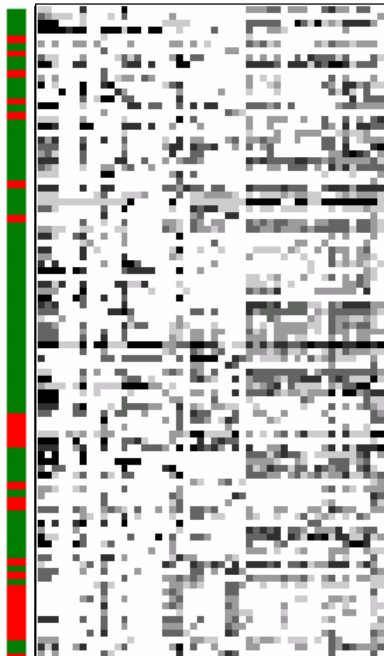
■ SANS

Patients

■ Schizophrenic

■ Bipolar disorders

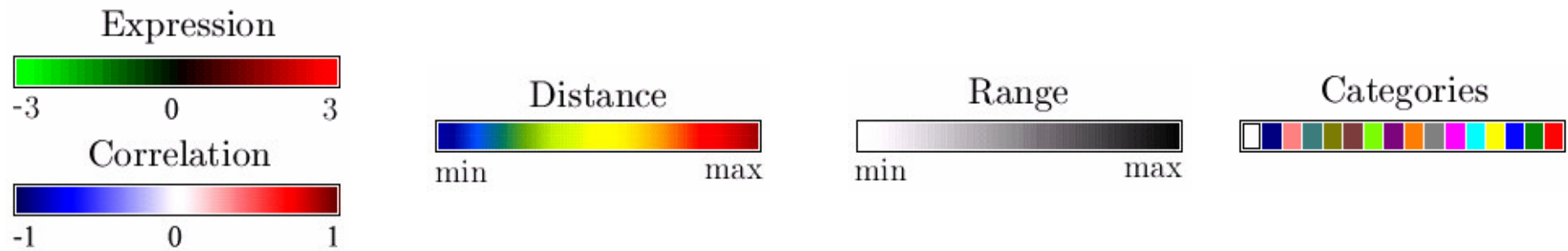
Raw Data Matrix



Color Spectra

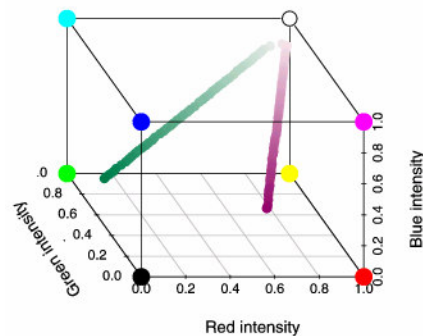
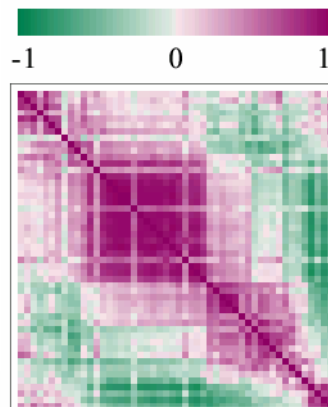


Correlation matrix map of 50 psychosis disorder variables

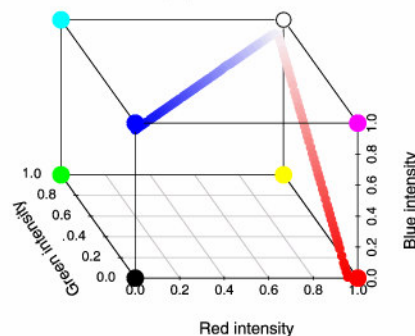
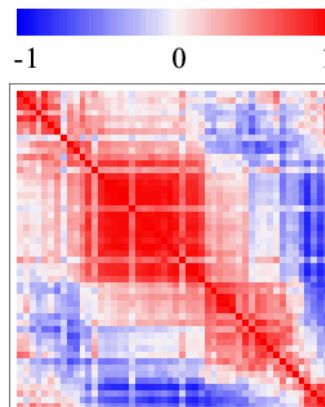


Correlation matrix map of 50 psychosis disorder variables

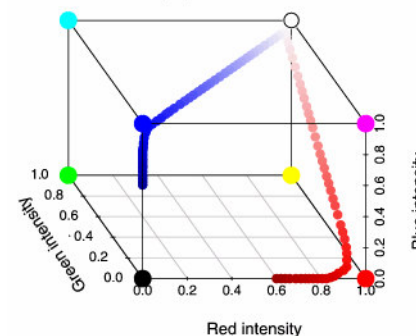
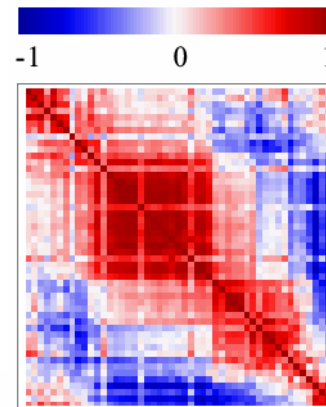
(a)



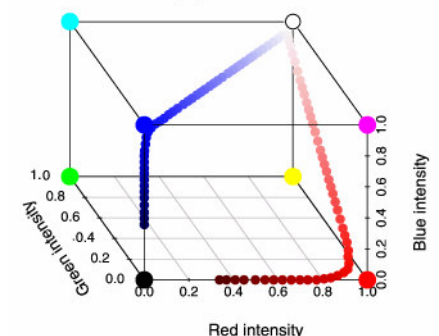
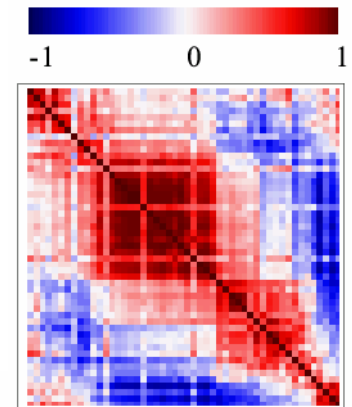
(b)



(c)



(d)

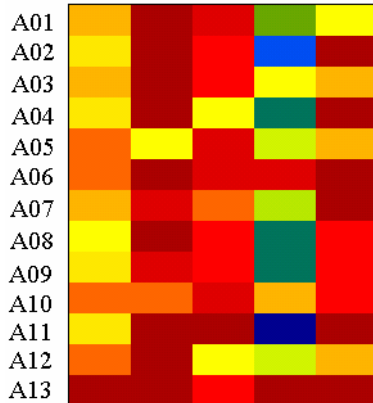


RGB

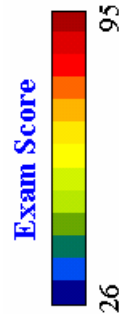
Display Conditions



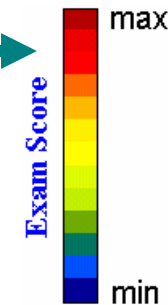
小考1 期中 小考2 期末 報告



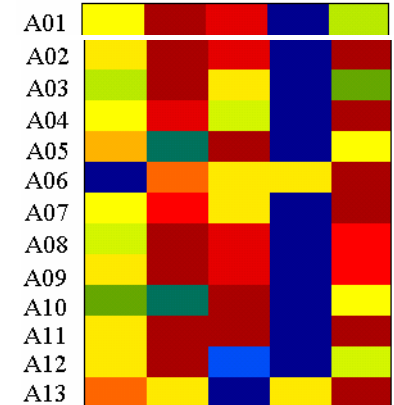
Range Matrix Condition



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



小考1 期中 小考2 期末 報告



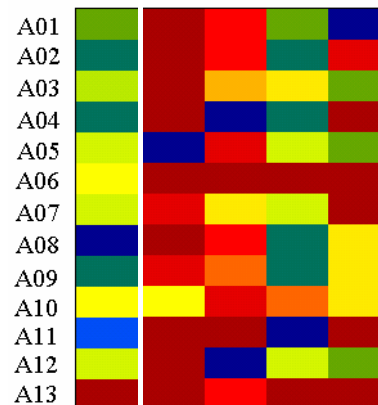
Range Row Condition

Exam Score



min max

小考1 期中 小考2 期末 報告



Range Column Condition

What about this one?

Exam Score

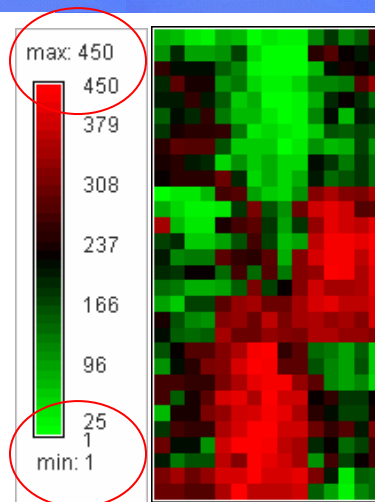
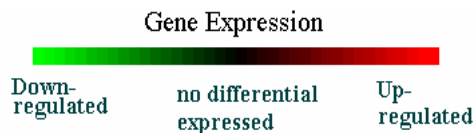


min max
0 100

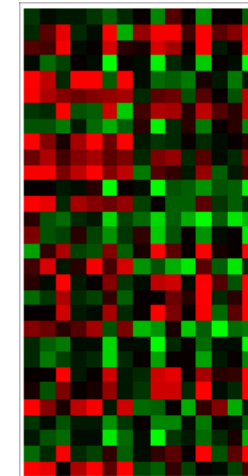
Display Conditions



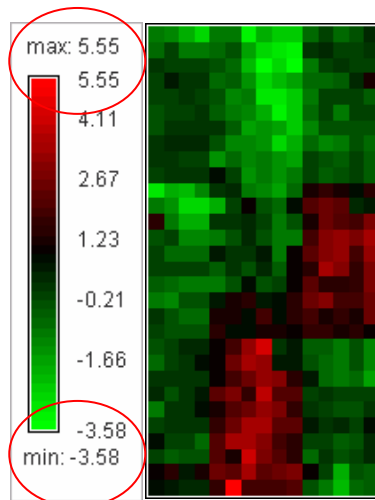
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.88	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	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	0.80	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	1.80	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	1.26



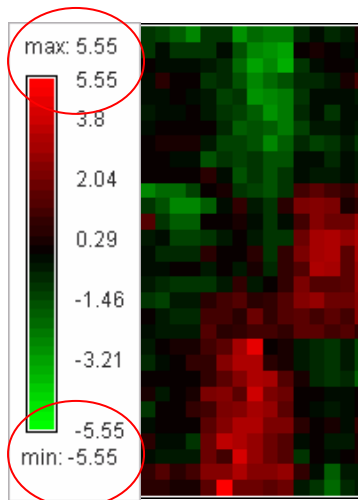
Rank Matrix Condition



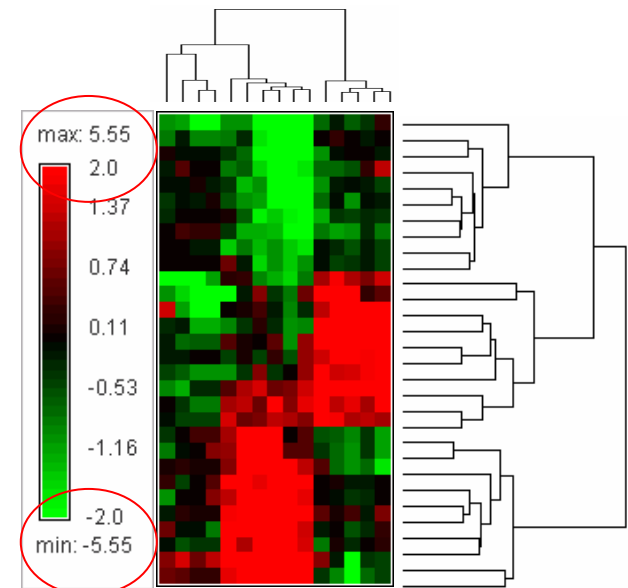
Without ordering



Range Matrix Condition



Center Matrix Condition



The 2nd Step of GAP

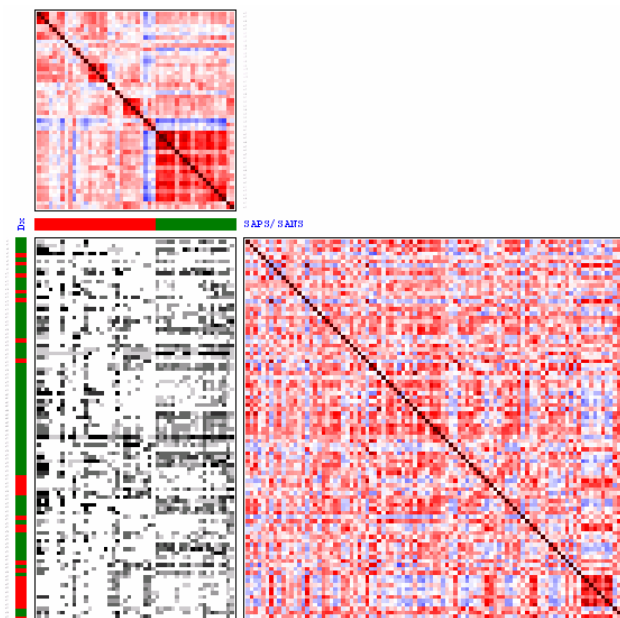
Seriation of Proximity Matrices and Raw Data Matrix

- **Relativity of a Statistical Graph**
- **Global Criterion**
 - **GAP Rank-Two Elliptical Seriation**
- **Local Criterion**
 - **Tree Seriation**
 - **Flipping of Tree Intermediate Nodes**

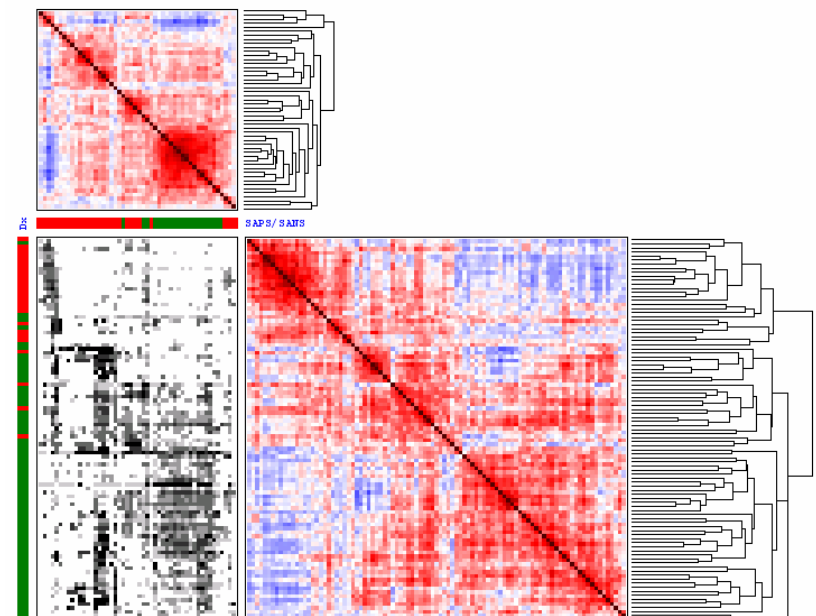
Relativity of a Statistical Graph



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

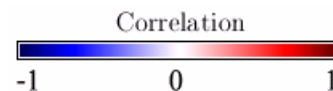


**Seriation
Methods**



(1) Rank Two Ellipse Ordering (Chen, 2002)

(2) Hierarchical Clustering Tree
(Average-Linkage)

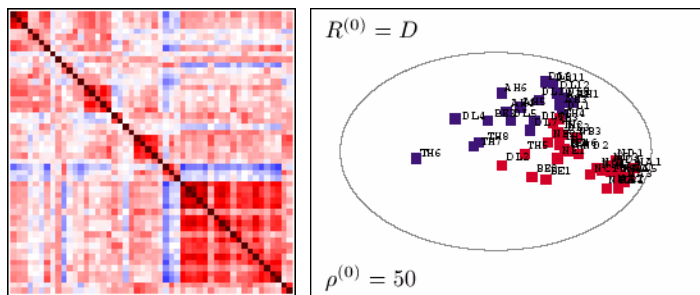


GAP Rank-Two Elliptical Seriation

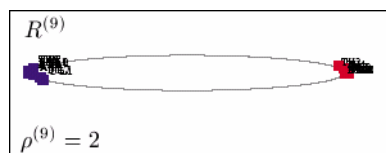
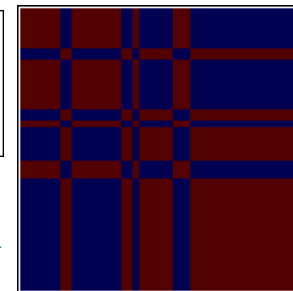
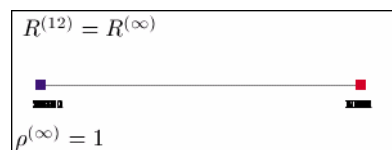
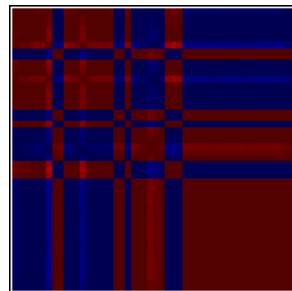
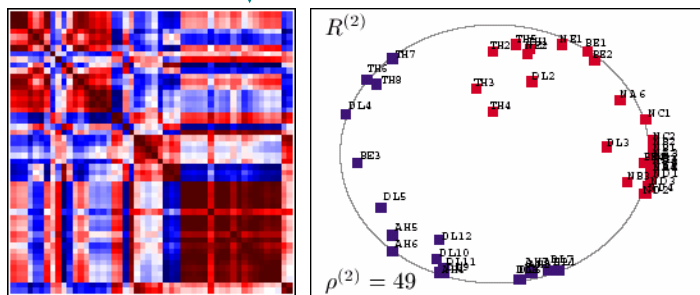
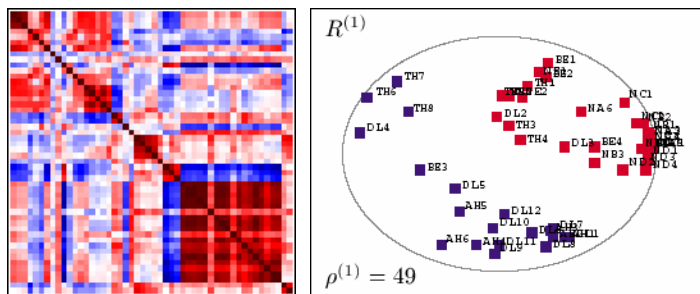
13/28

Seriation Algorithms with Converging Correlation Matrices

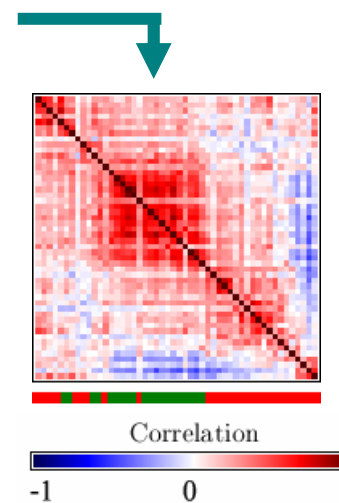
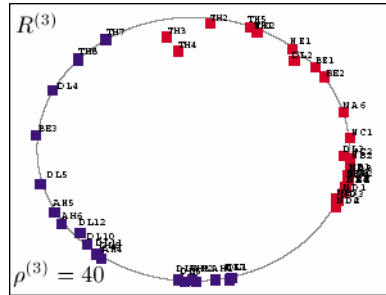
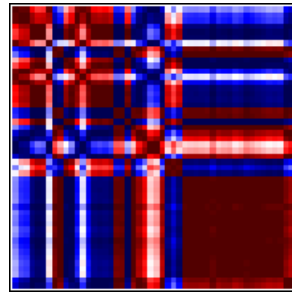
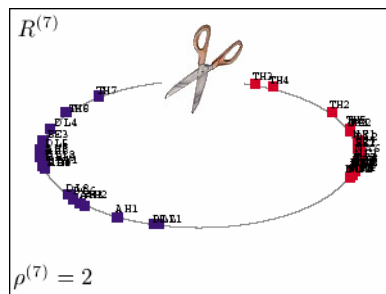
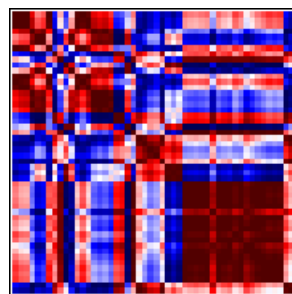
Correlation Matrix (without ordering)



First two Eigenvectors



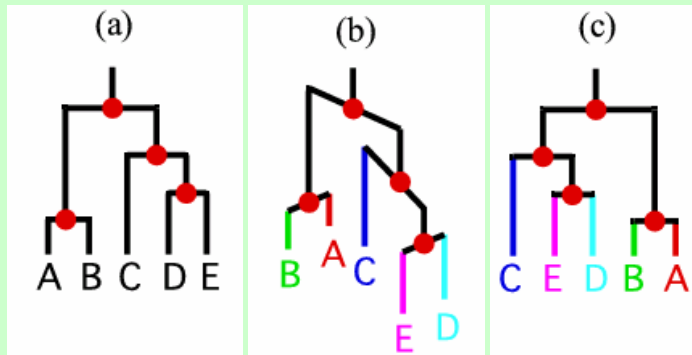
The p objects fall on an ellipse and have unique relative position on the ellipse (Chen 2002).



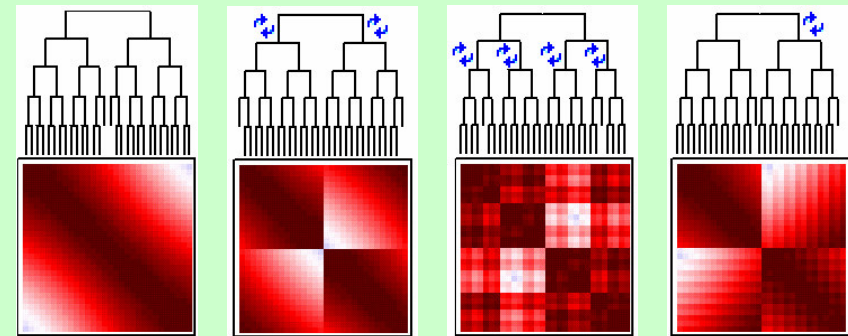
Hierarchical Clustering Tree with a Dendrogram



Tree seriation

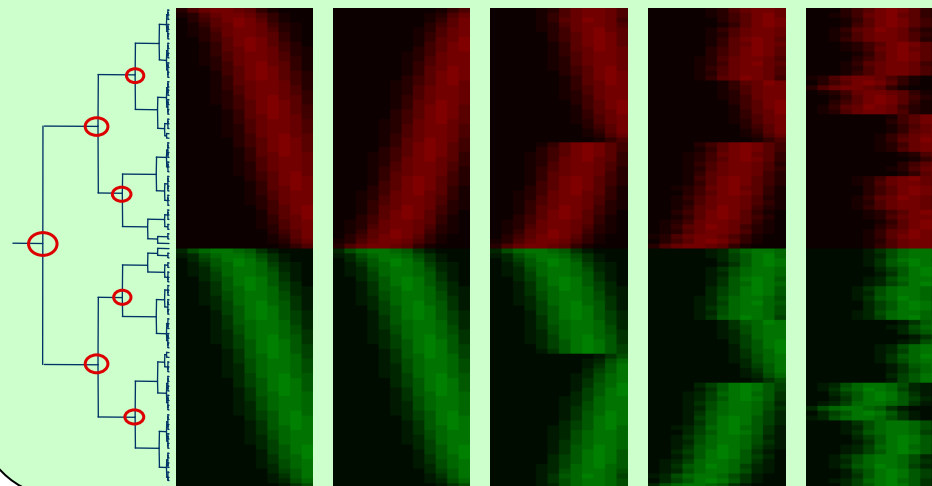


Tree seriation for proximity matrices



Tree seriation for raw data matrices

ideal model 1 flip 3 flips 5 flips many flips



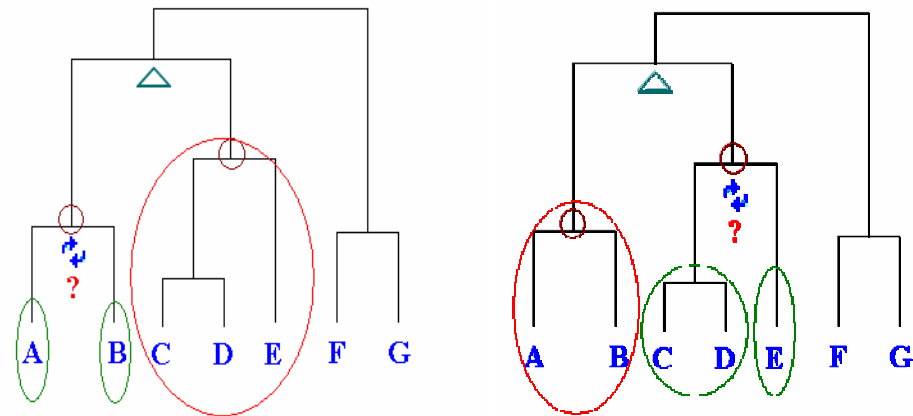
**Different Seriations
Generated from Identical
Tree Structure**

Internal Tree Flips

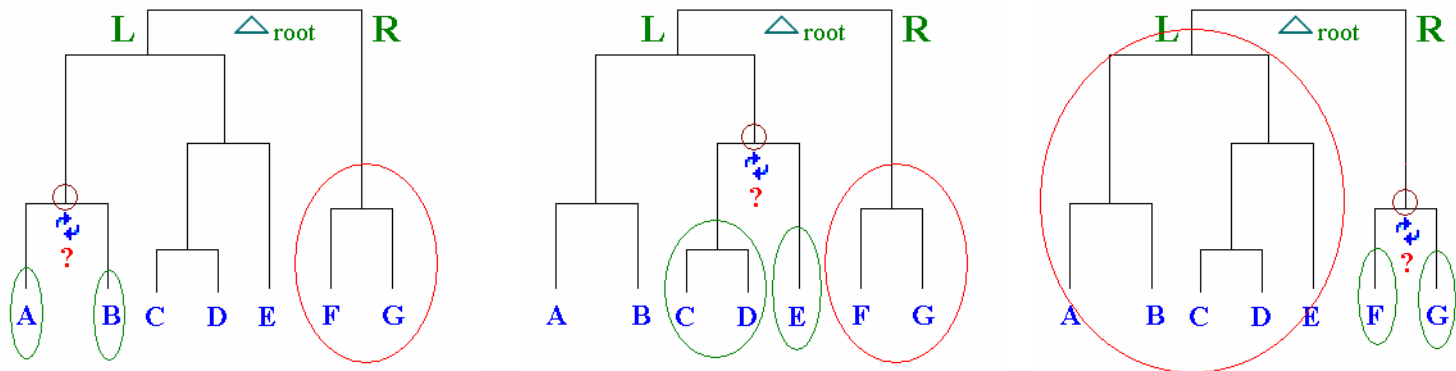
15/28

Uncle Approach

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

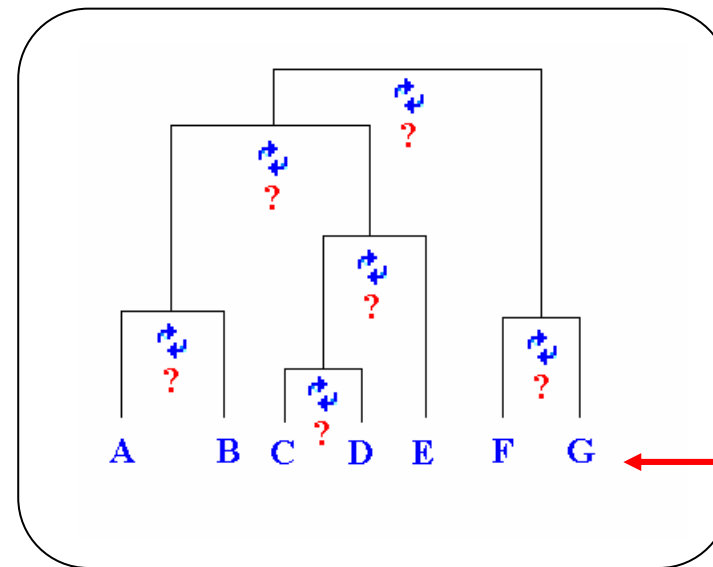


GrandPa Approach



External Tree Flips

16/28



External Ordering

D E A F B C G

As close 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) ...

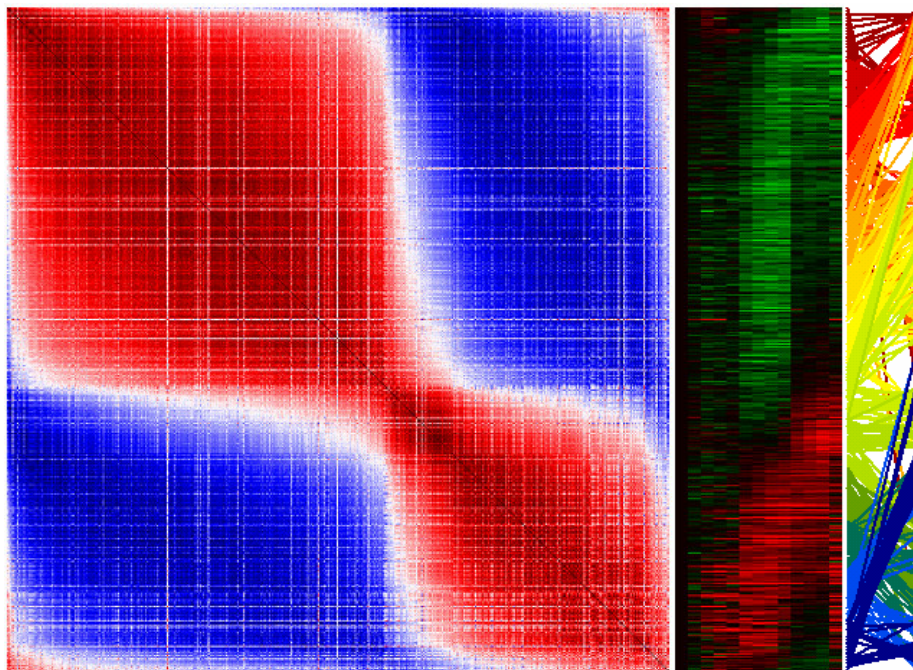
Global vs. Local Seriation



Data: 517 genes by 13 arrays

GAP Rank-two elliptical seriation

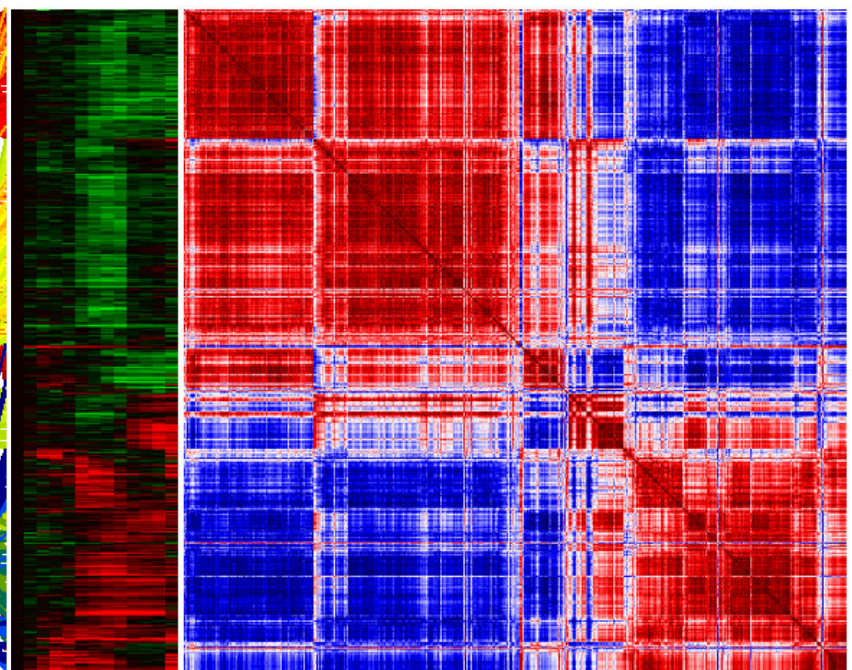
time
→



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

Michael Eisen (1998) tree seriation

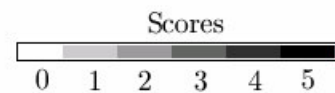
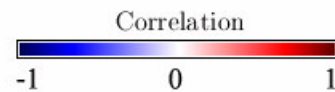
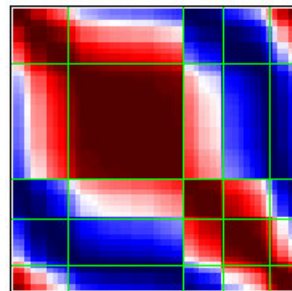
time
→



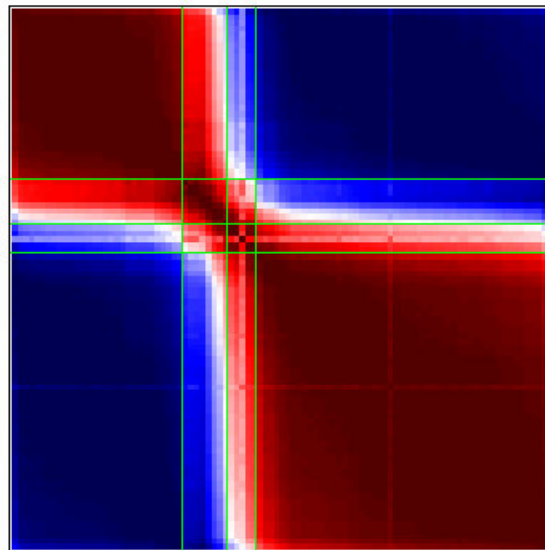
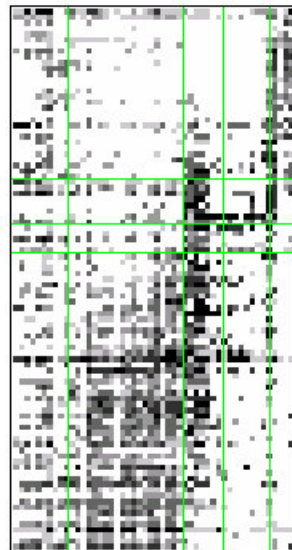
-1 0 1

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

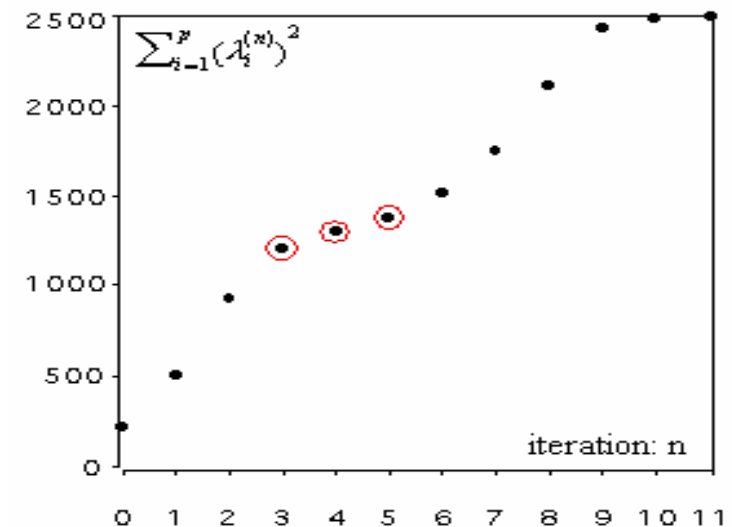
Partitions of Permuted Matrix Maps



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



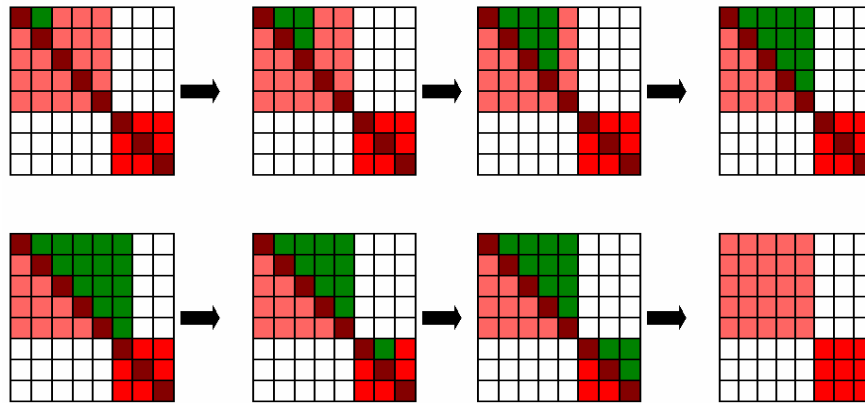
Sum squared eigenvalues
(sum squared correlations)



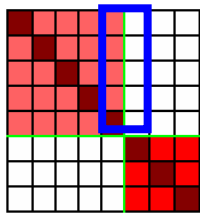
Partitions of Permuted Matrix Maps



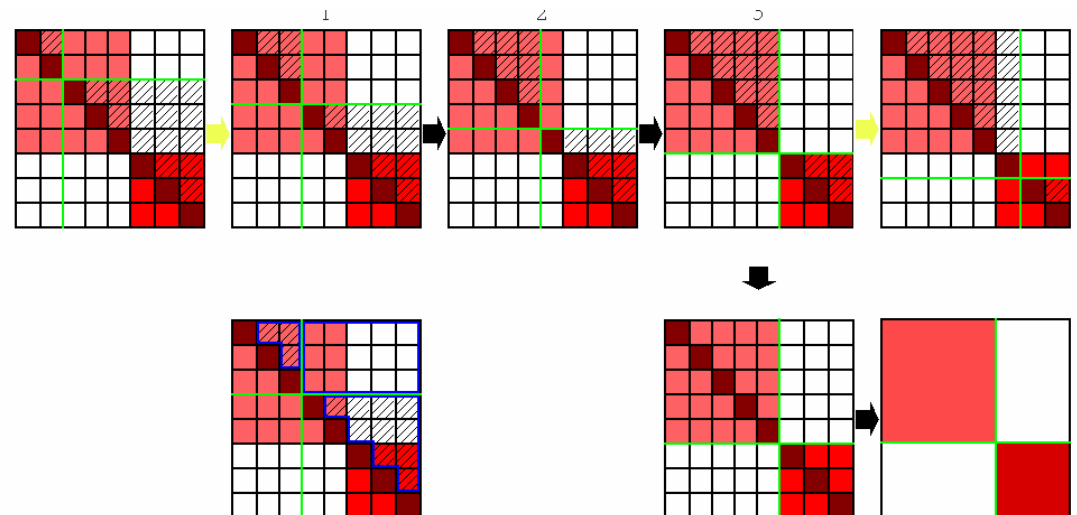
One-Way block Searching



Two-Sample Problem



Within-Sum-of-Square Approach



The 4th Step of GAP

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

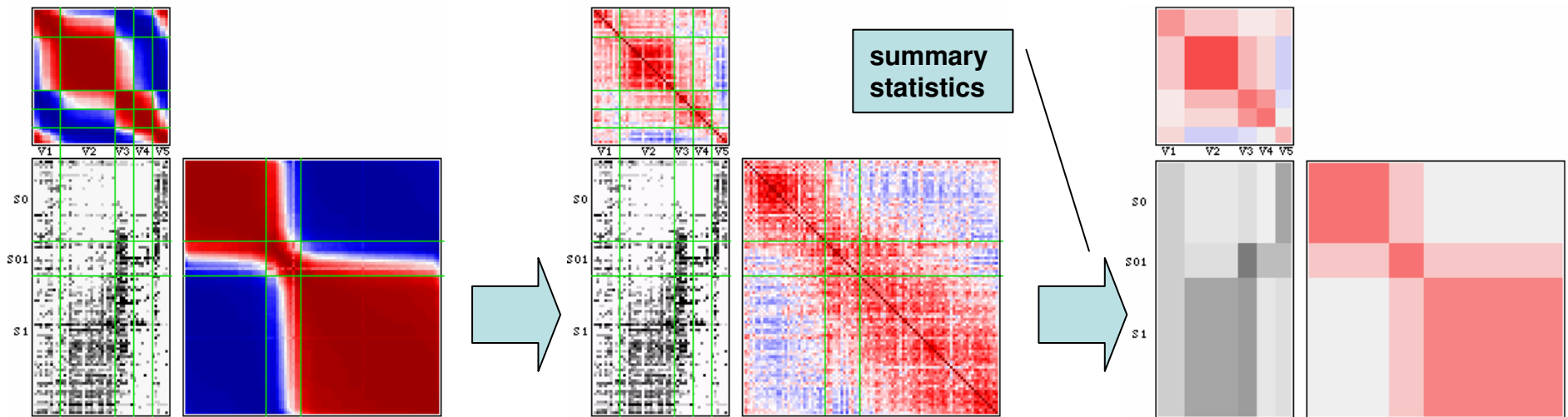
Sufficient Statistics

	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	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

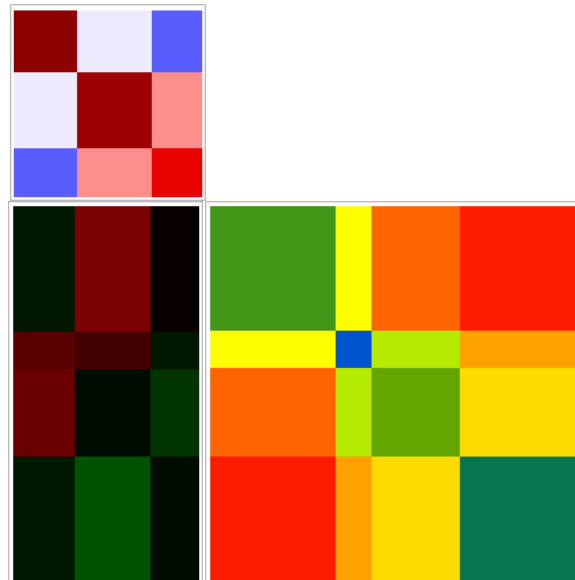
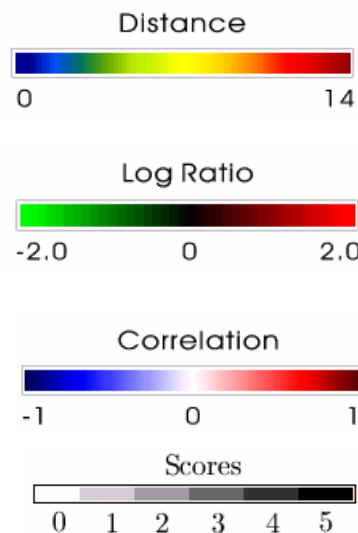
Sufficient Graph

21/28



Summarizing Display

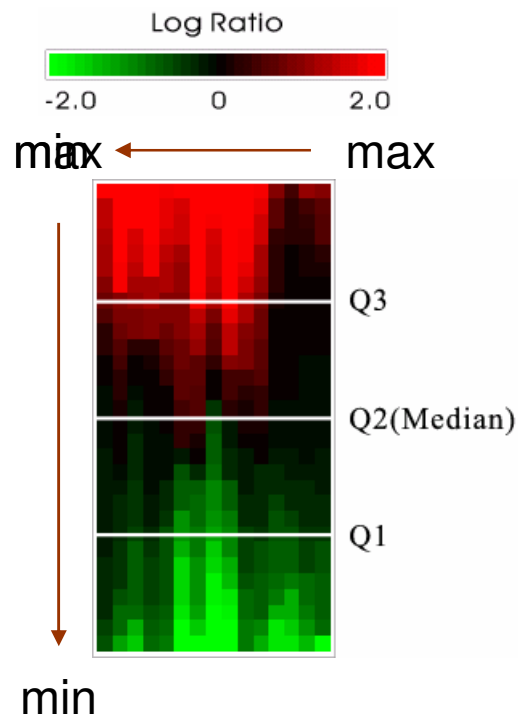
- (1) subject-subject
- (2) variable-variable
- (3) subject-variable



Generalization and Flexibility

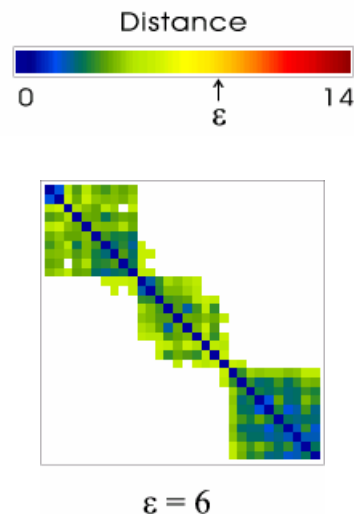
22/28

Sediment Display



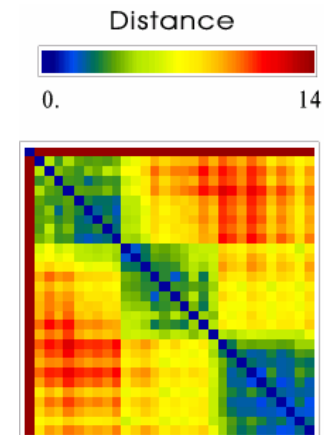
Similar information to that given by a boxplot when the color strips at the quartile positions are extracted.

Sectional Display

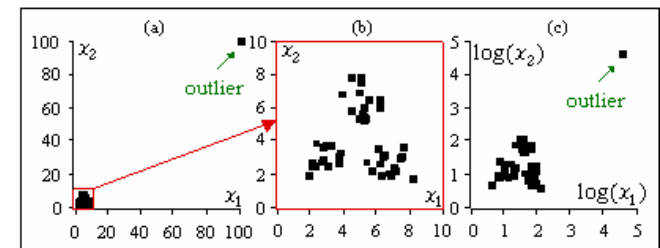


Display only those numerical values that satisfy certain conditions.

Restricted Display

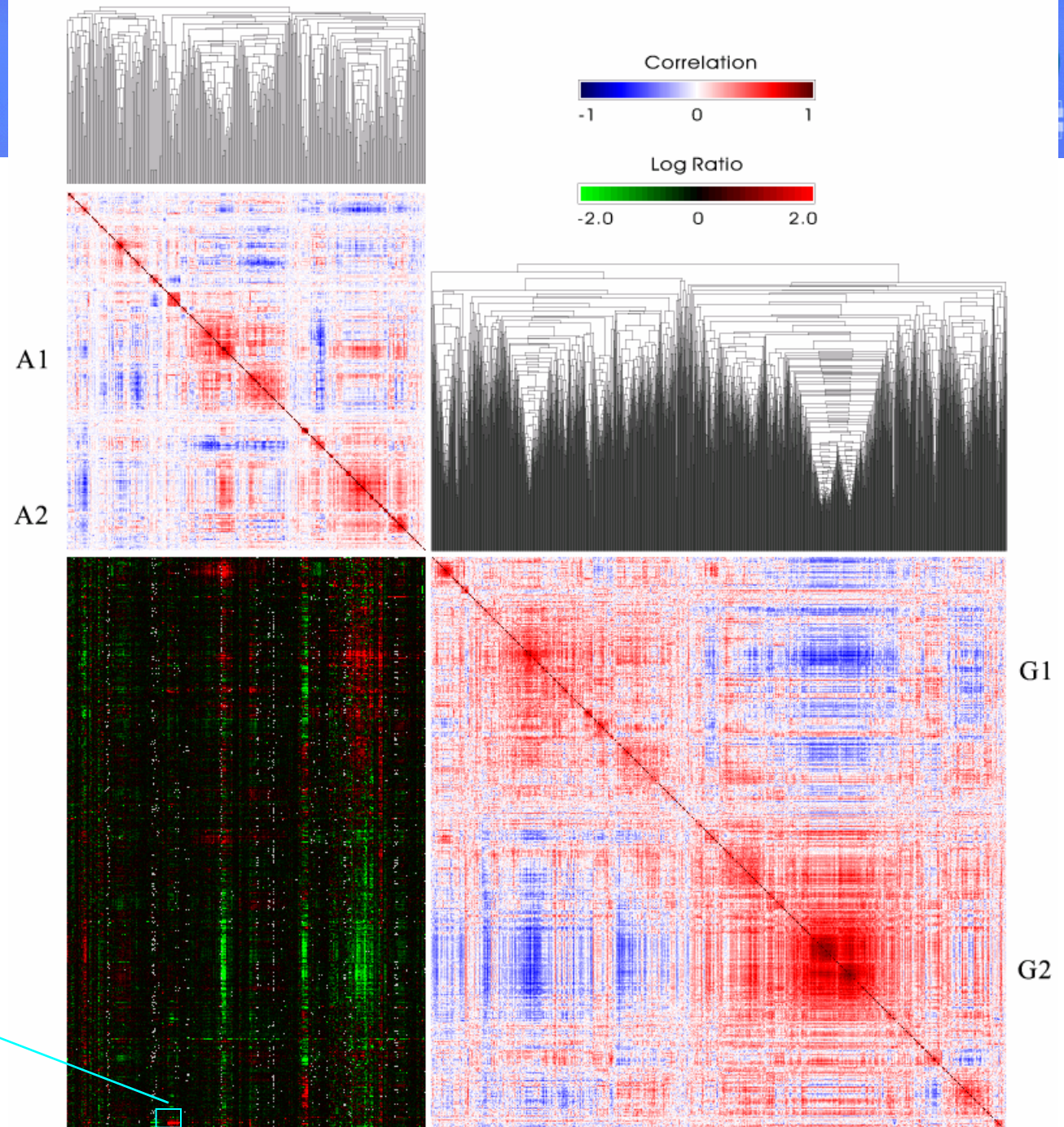


Resolution of a Statistical Graph



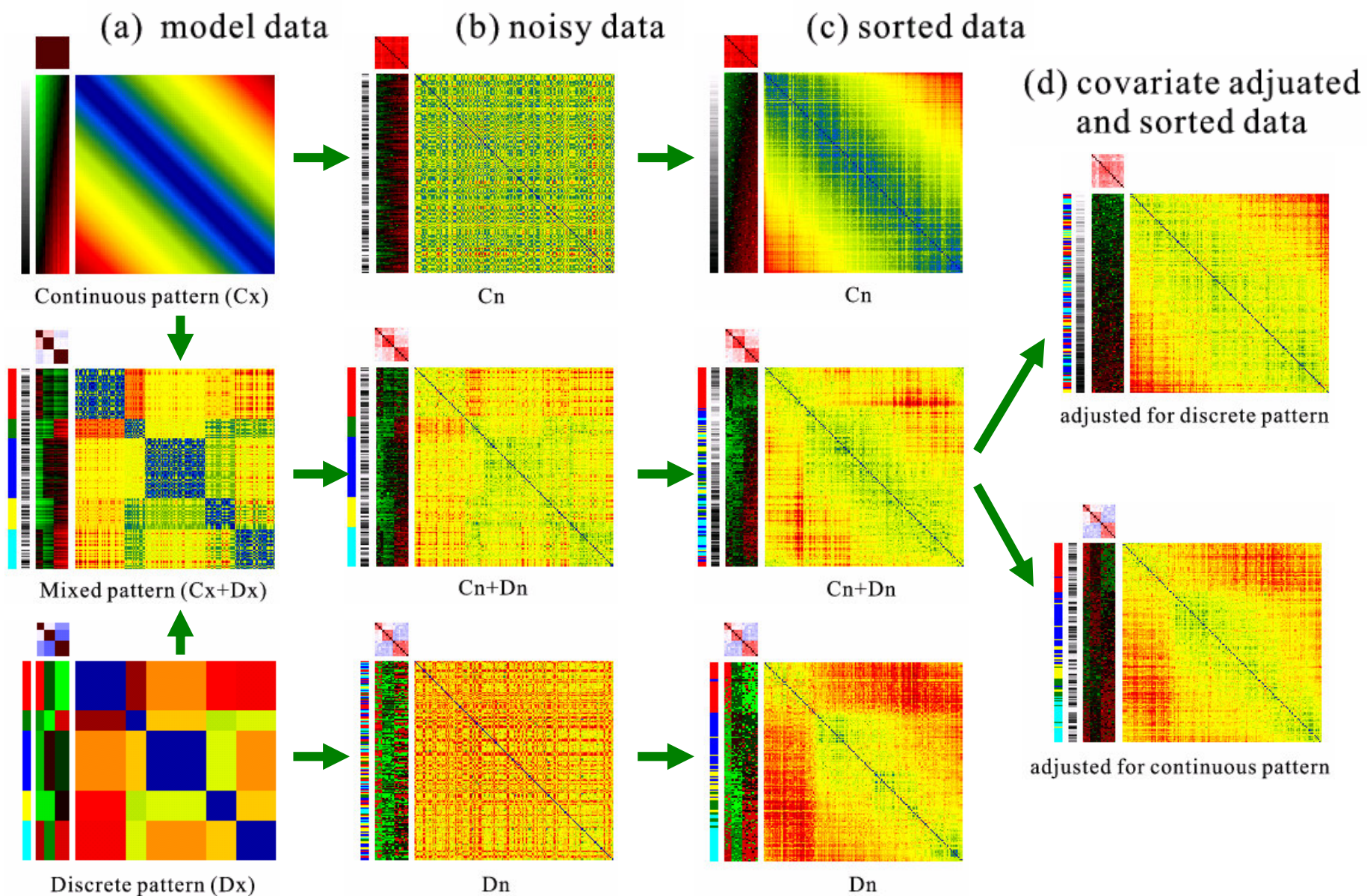
An Example

- 2000 genes by 400 arrays with relatively fewer missing values.
- Pearson's correlation coefficient.
- Average linkage clustering trees.
- The basic gene clustering structure and array (experiments) grouping patterns can be identified using these tree sorted matrix maps.



Module: MV for Covariate Adjustment

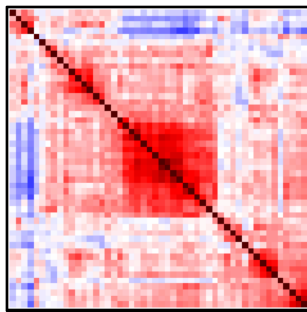
24/28



Module: Interactive Diagnostic System for Hierarchical Clustering Tree with Matrix Visualization

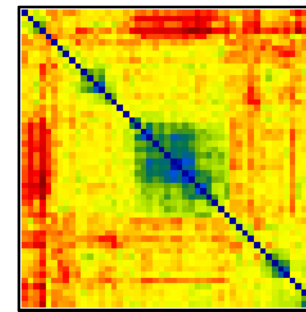
25/28

(1) Input **Proximity** Matrix



(e.g., Pearson's Correlation)

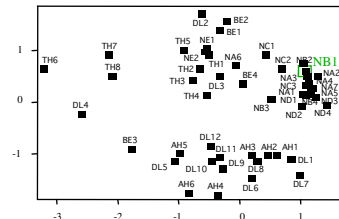
(2) Transformed **Disparity** Matrix



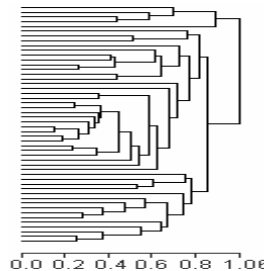
(e.g., Distance)

Statistical Modeling

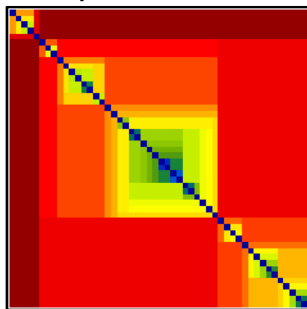
Multidimensional Scaling (MDS)



Hierarchical Clustering Tree (HCT)

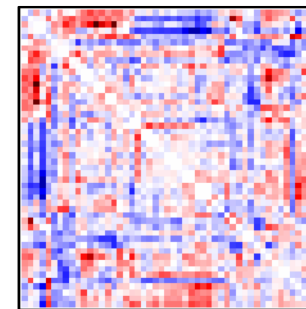


(e.g., Cophenetic Matrix)

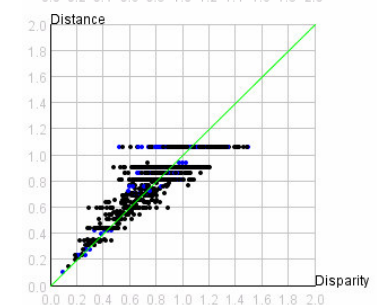
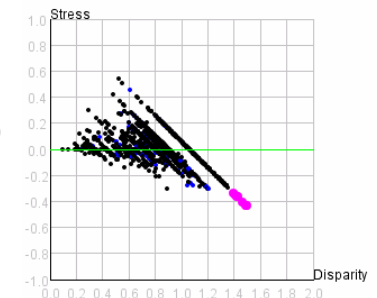
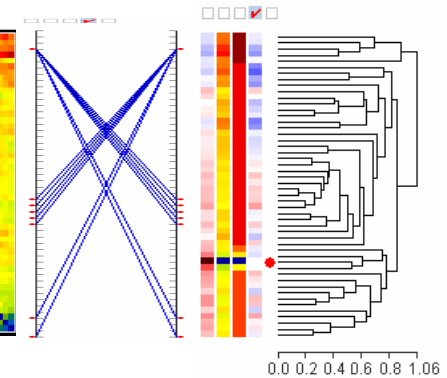


(3) Output **Distance** Matrix

(e.g., Residual Matrix)



(4) **Stress** Matrix



GAP Software version 0.2

26/28

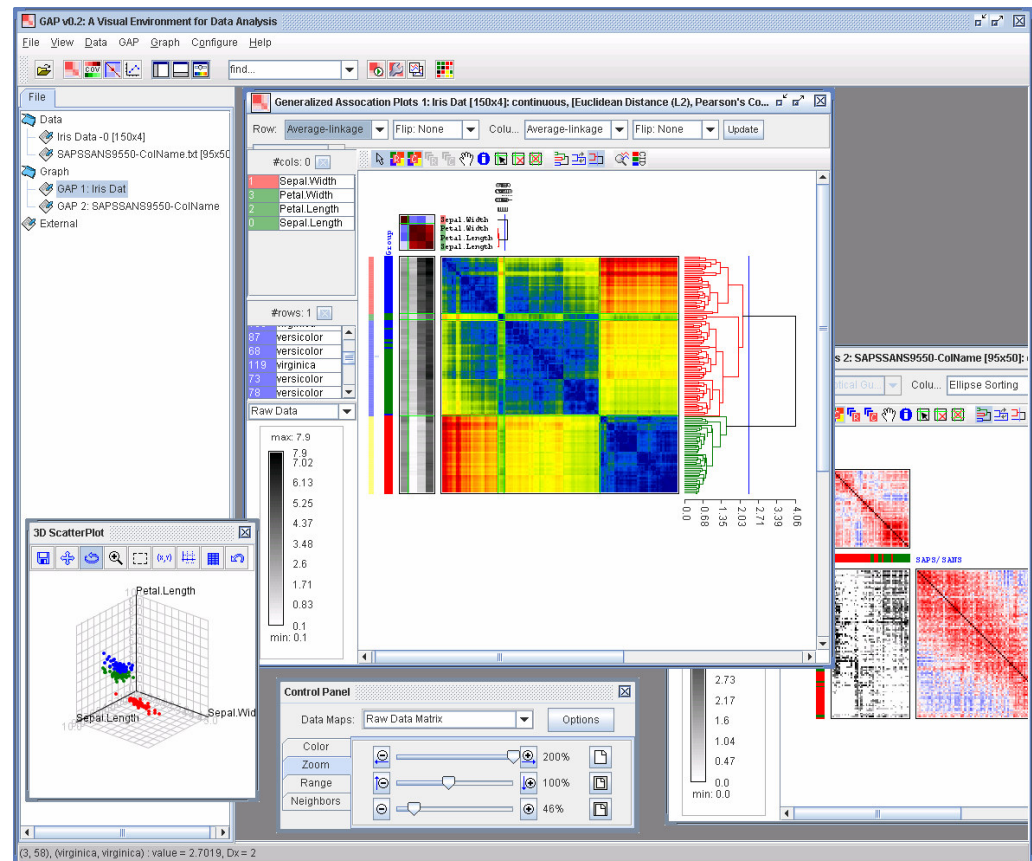
Generalized Association Plots

- 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)



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

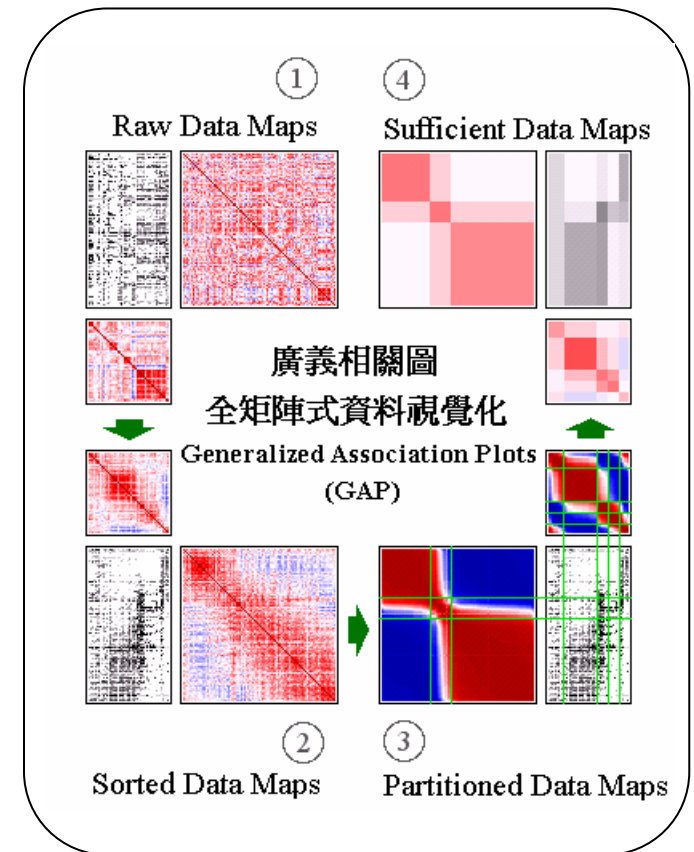
Conclusion

27/28

MV is the **color order-based** representation of data matrices.

A MV displays provide five levels of information:

1. raw scores for every sample/variable combination;
2. an individual sample score vector across all variables, and an individual variable vector across all samples;
3. an association score for every sample-sample and variable-variable relationship;
4. a grouping structure for variables and a clustering effect for sample;
5. an interaction pattern of sample-clusters on variable-groups.



- A preliminary step in modern exploratory data analysis.
- A continuing and active topic of research and application.
- New generation of exploratory data analysis (EDA) tool.

Web Site

28/28

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/

Dimension Free Data Visualization
GAP

Lab for Information Visualization
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 | **GAP Forum** | Links | About Us

Chun-houh Chen
陳君厚
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

News/Conference [past events and more]
2006: 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

① Raw Data Maps

Generalized Association Plots (GAP)

for Dimension Free Data Visualization

② Sorted Data Maps

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

Han-Ming Wu (Hank) - 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://www.sinica.edu.tw/~hmwu/

Welcome To Hank's Homepage!

Home | Experience | Research | Publication | Course | Talks | Software | Links | Updated 2006/10/24

Han-Ming Wu (Hank) 吳漢銘
Postdoctoral Fellow
Institute of Statistical Science, Academia Sinica
128 Academia Rd. Sec.2, Nankang
Taipei, Taiwan 11529
Tel: +886-2-27835611 ext: 309
E-mail: hmwu@stat.sinica.edu.tw
HomePage: <http://www.sinica.edu.tw/~hmwu/>

Education

- Ph.D. (9/1997 - 10/2003), Institute of Statistics
National Chiao Tung University, Taiwan, R.O.C.
- M.S. (9/1995 - 9/1997), Institute of Mathematical Statistics
National Chung Cheng University, Taiwan, R.O.C.
- B.S. (9/1991 - 9/1995), Department of Mathematics
Tam Kang University, Taiwan, R.O.C.

Research Interests

- Bioinformatics: [Statistical Microarray Data Analysis](#)
- Information Visualization: [Matrix Visualization](#)
- Dimension Reduction: [Sliced Inverse Regression](#)
- Statistical Computing Using Java and R
- Statistical Learning: Kernel Machines, Manifold Learning.
- Statistical Applications: Image Segmentation

Conference/Workshop

- Dec 22, 2006
GAP Tutorial
[The Institute of Statistical Mathematics, Tokyo, Japan](#)
- July 29 - August 2, 2007
Joint Statistical Meetings
Salt Lake City, Utah
- More...

myLinks

- My Wedding
- My Photo
- My Painting
- [中研院統計所 助理網](#)
- [Handbook of Computational Statistics Volume III: Data Visualization](#)

吳漢銘, 中央研究院 統計科學研究所, 11529 台北市南港區研究院路2段126號, Tel: +886-2-27835611 ext. 309

Lab for Information Visualization

cchen@stat.sinica.edu.tw
<http://gap.stat.sinica.edu.tw>



Appendix

Related Works of MV

30/28

Concept:

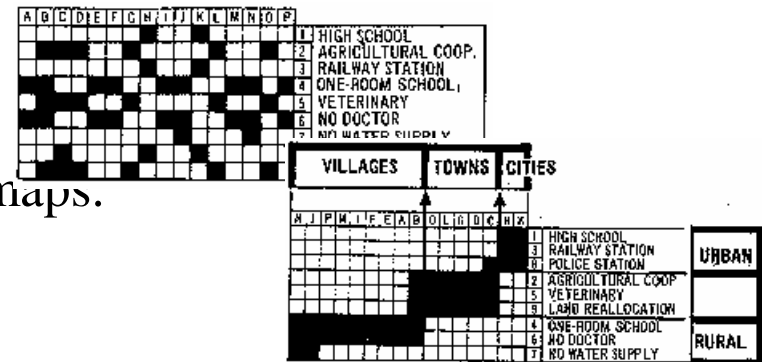
- Bertin (1967): reorderable matrix.
- Carmichael and Sneath (1969): taxometric maps.

Clustering of data arrays:

- Hartigan (1972): direct clustering of a data matrix.
- Tibshirani (1999): block clustering.
- Lenstra (1974): traveling-salesman problem.
- Slagle *et al.* (1975): shortest spanning path.

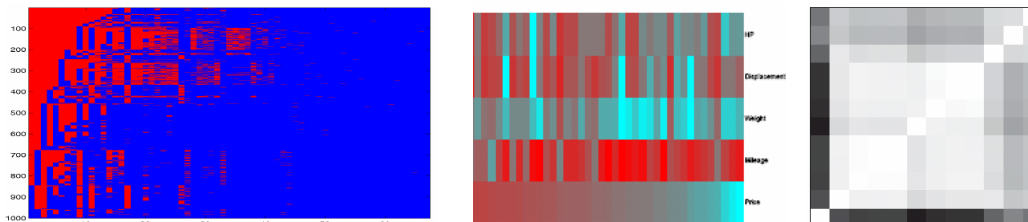
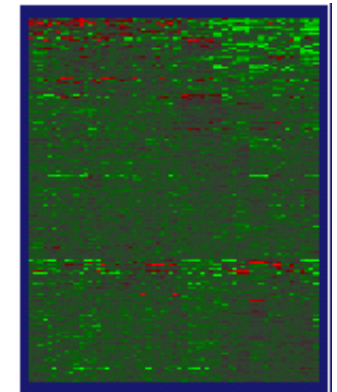
Colour Representation:

- Wegman (1990): colour histogram.
- Minnotte and West (1998): data image.
- Marchette and Solka (2003): outlier detection.



4. UN VOTES IN 1969-1970*

State	1	2	3	4	5	6	7	8	9	10	11	12	13	14
USSR	1	1	1	1	3	1	2	3	1	3	2	2	1	3
BGR	1	1	1	1	3	1	1	3	1	3	2	2	1	3
YUG	1	3	3	3	3	1	1	3	1	2	3	1	1	2
SYR	1	2	2	2	3	1	1	3	1	2	3	1	1	3
UAR	1	3	3	3	3	1	1	3	2	2	3	1	1	3
KEN	1	3	3	3	3	1	1	3	2	5	3	1	1	3
TAN	1	2	2	2	3	1	1	3	2	5	3	1	1	3
SEN	1	3	3	3	1	2	2	2	2	1	3	1	1	2
DAH	1	3	3	3	1	3	1	3	3	1	3	1	2	2
USA	1	3	3	3	1	3	3	1	3	1	1	3	3	1
URK	1	3	3	3	1	1	3	2	3	1	1	3	3	1
FRA	1	3	3	3	3	1	2	3	3	1	1	3	2	2
SWE	1	3	3	3	3	1	2	3	3	1	1	3	3	1
NOR	1	3	3	3	3	1	1	3	2	3	1	1	3	3
ALA	1	3	3	3	1	3	1	2	3	1	1	3	3	1
NZ	1	3	3	3	1	3	1	1	3	1	1	3	3	1
MEX	1	2	2	2	1	3	3	1	3	1	1	2	1	1
VEN	1	2	2	2	1	3	3	1	3	1	2	1	1	1
BRA	1	2	2	2	1	3	3	1	3	1	1	3	1	1

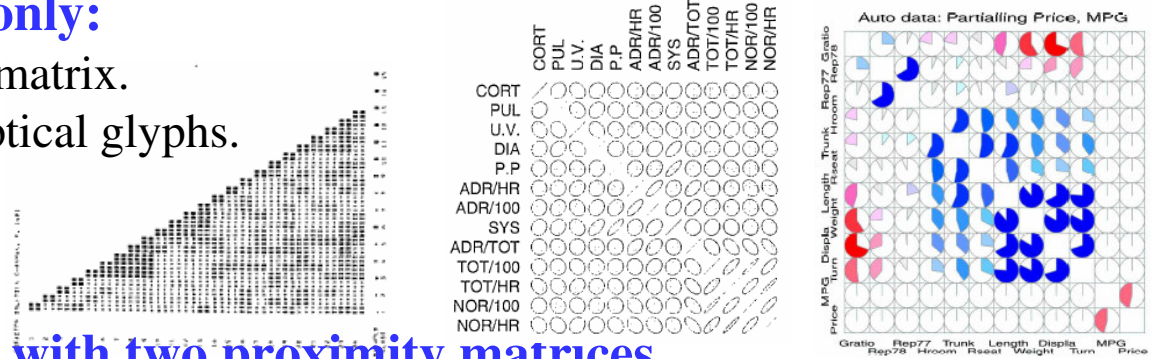


Related Works (conti.)

31/28

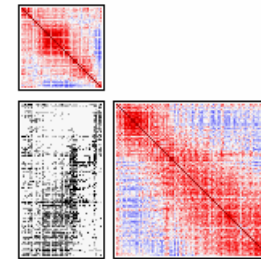
Exploring proximity matrices only:

- Ling (1973): shaded correlation matrix.
- Murdoch and Chow (1996): elliptical glyphs.
- Friendly (2002): corrgrams.



Integration of raw data matrix with two proximity matrices

- Chen (1996, 1999, and 2002): generalized association plots (GAP).



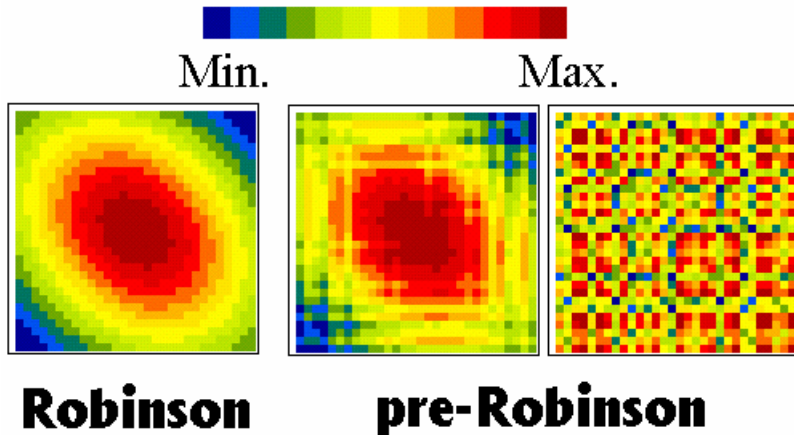
Reordering of variables and samples

- Chen (2002): concept of relativity of a statistical graph.
- Friendly and Kwan (2003): effect ordering of data displays.
- Hurley (2004): placing interesting displays in prominent positions.

Matrix Visualization (MV): reorderable matrix, the heatmap, color histogram, data image and matrix visualization.

Criteria for a *good* Permutation

32/28



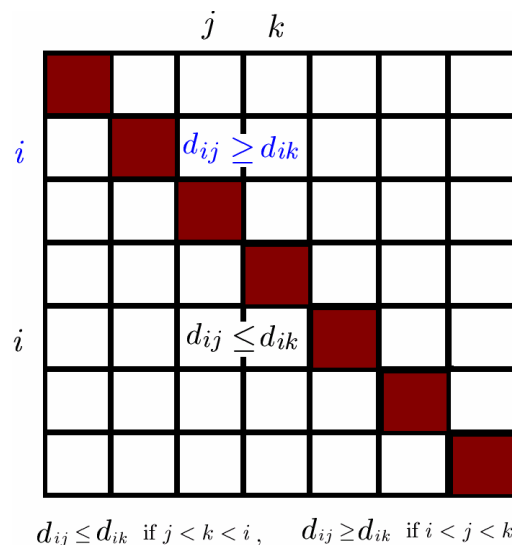
Global criterion: Anti-Robinson Measurements

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

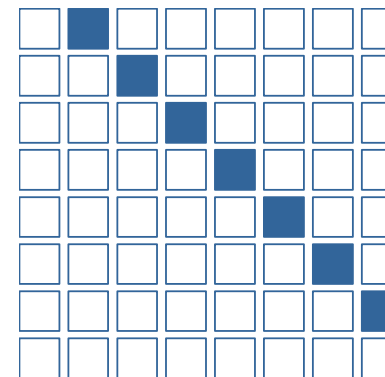
$$AR(s) = \sum_{i=1}^p \left[\sum_{j < k < i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i < j < k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}| \right],$$

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

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



Local criterion: Minimal Span Loss Function



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

Hierarchical Clustering Tree

33/28

(Kaufman and Rousseeuw, 1990)

Example: Average-Linkage

distance matrix

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



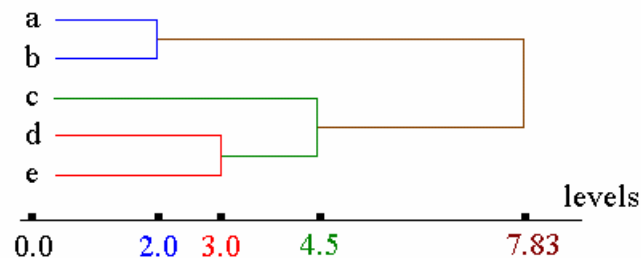
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0



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



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



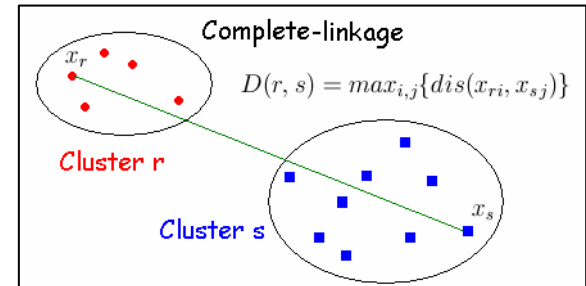
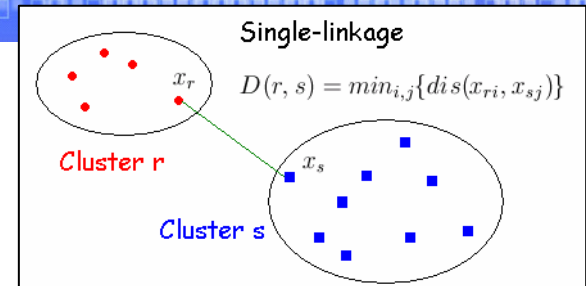
$$D(\{a, b\}, \{c\}) = \frac{1}{2}[D(a, c) + D(b, c)]$$

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

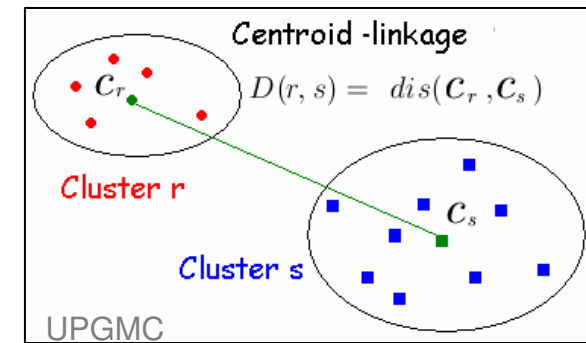
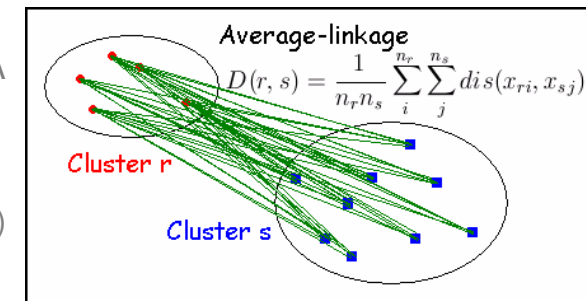
$$D(\{a, b\}, \{d, e\})$$

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

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



UPGMA
(Unweighted
Pair-Groups
Method
Average)



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