

入門

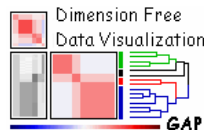
一般相関プロット(GAP) - 次元によらないデータ可視化 -

吳漢銘 Wu Han-Ming (Hank)

統計科学研究所、アカデミアシニカ、台北、台湾

hmwu@stat.sinica.edu.tw

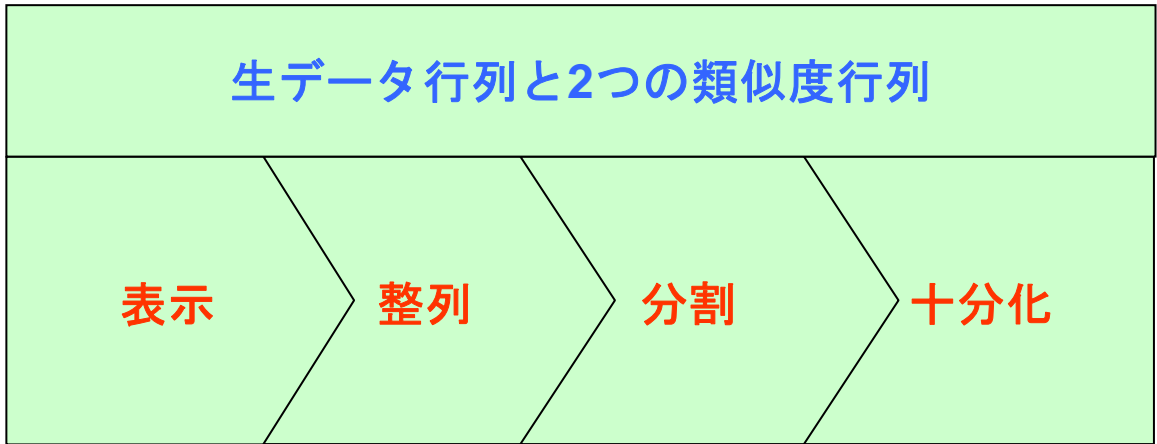
<http://www.sinica.edu.tw/~hmwu/>



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

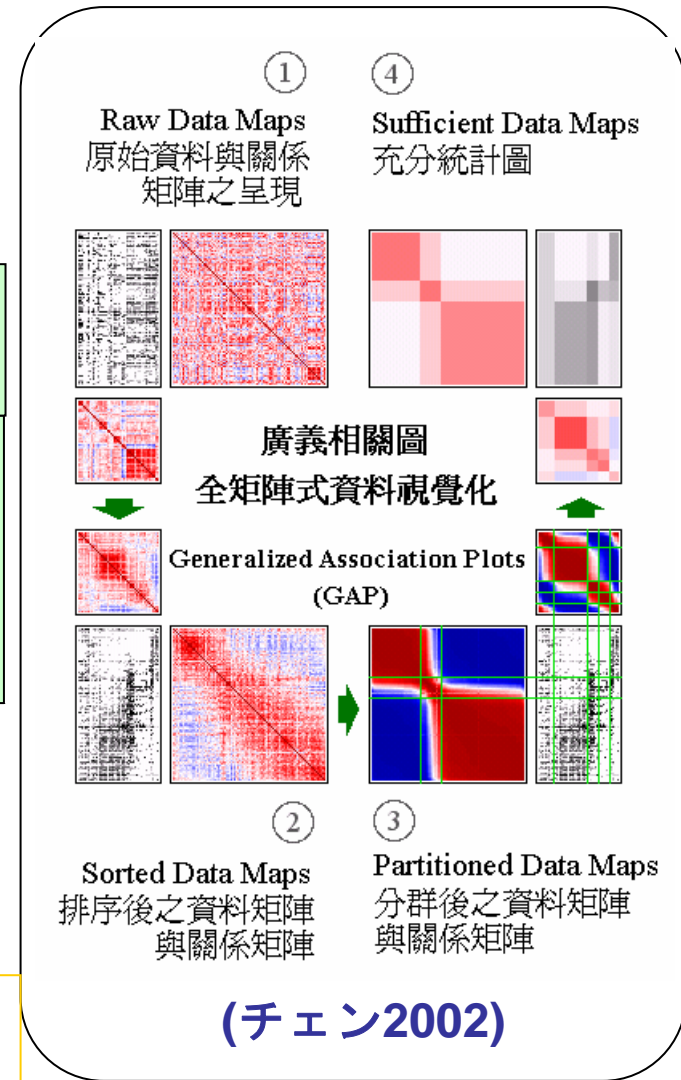
2006/12/22

- 2つのデモ用データセット
- 一般相関プロット(GAP)の4ステップ



- 一般性と柔軟性
- モジュール/ソフトウェア/結論

注意： 行列可視化(MV): 再整列行列、heatmap、色ヒストグラム、データイメージと行列可視化

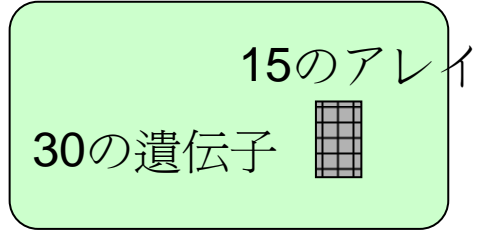
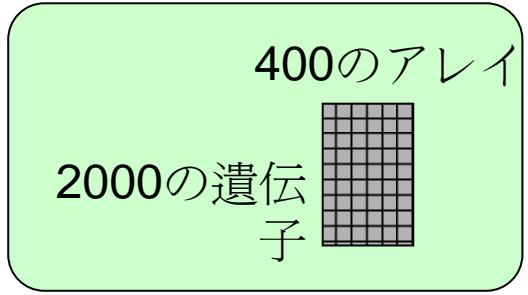
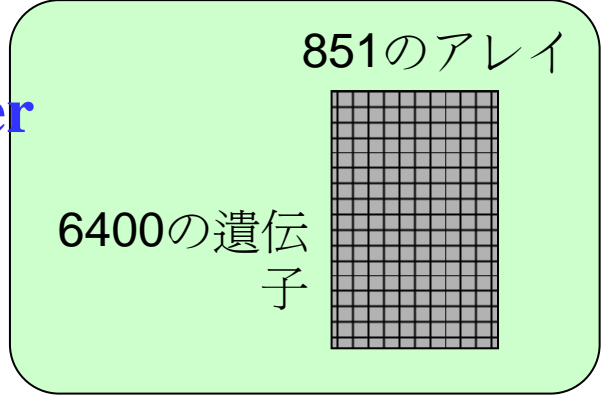


イースト細胞周期のマイクロアレイ遺伝子発現データ

yMGV: イーストMicroarray Global Viewer



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



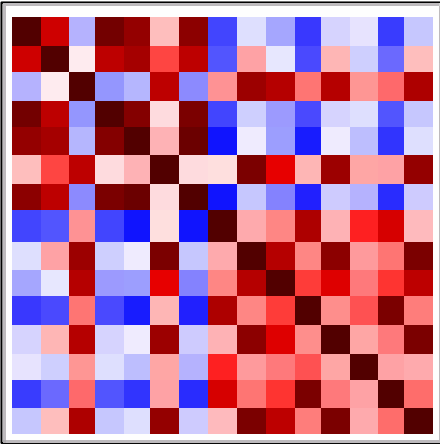
GAPの最初のステップ

生データ行列の表示

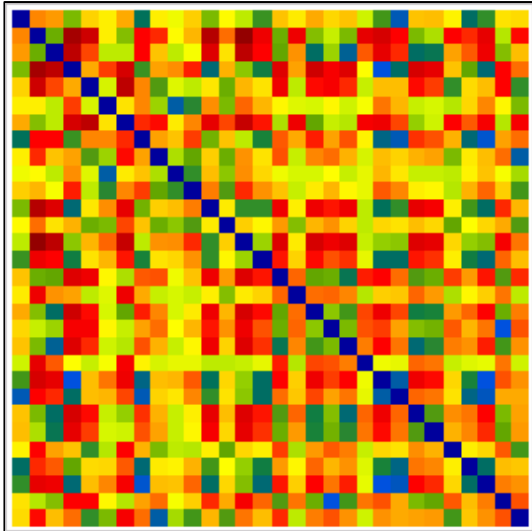
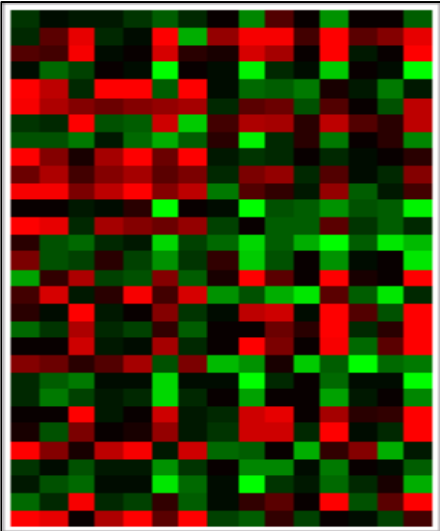
- データ変換
- 類似度の選択
- 色スペクトル
- 表示の条件

生データ行列の表示

↓
列の類似度行列



行の類似度行列



(1) 類似度の選択

ピアソン相関係数

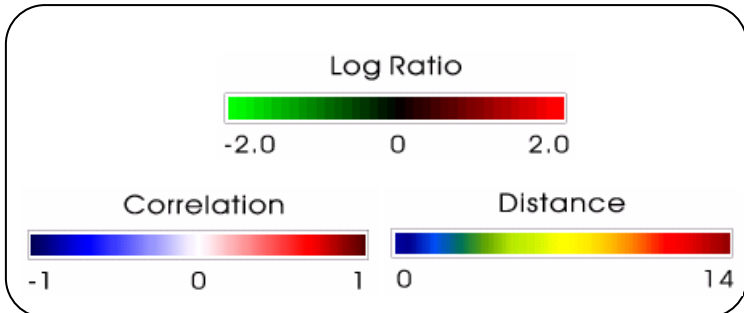
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

ユークリッド距離

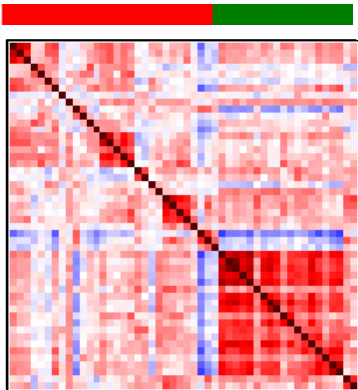
$$d_{xy} = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

他の類似性/相違性測度

(2) 色のスペクトル



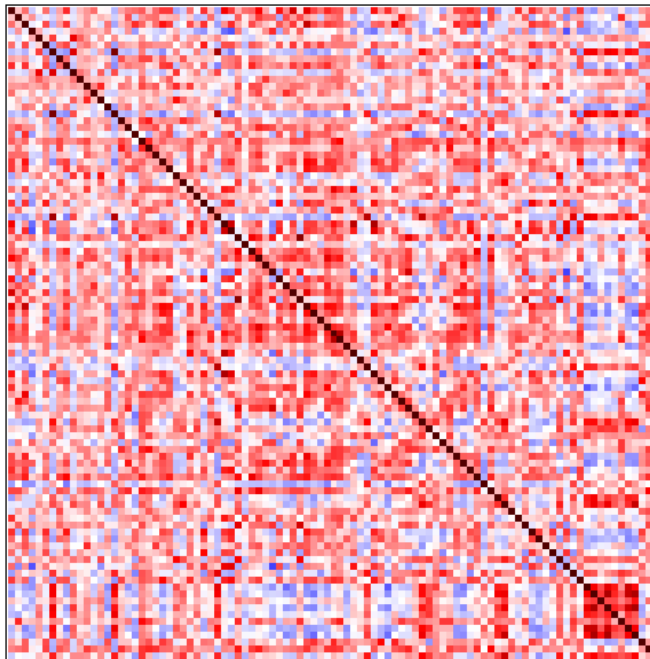
生データ行列の表示



変数の相関行列



観測値の相関行列

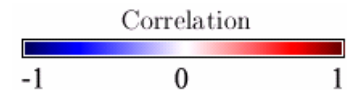


生データ行列

(1) 類似度の選択

ピアソン相関係数

(2) 色のスペクトル



Symptoms

SAPS

SANS

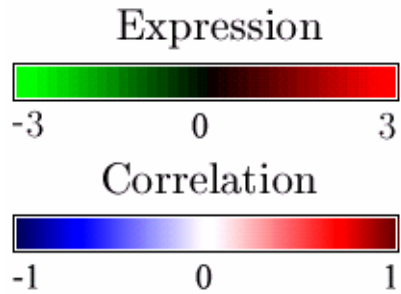
Patients

Schizophrenic

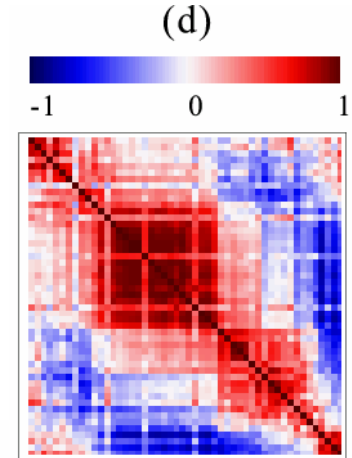
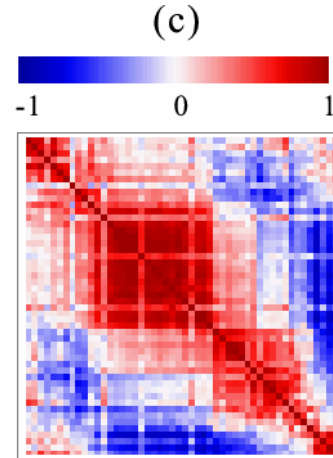
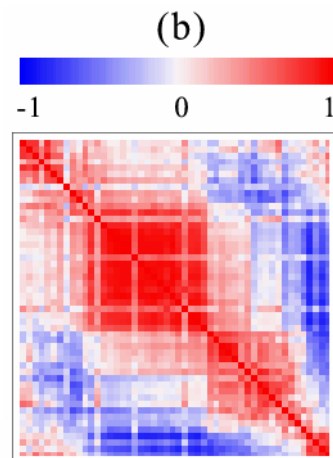
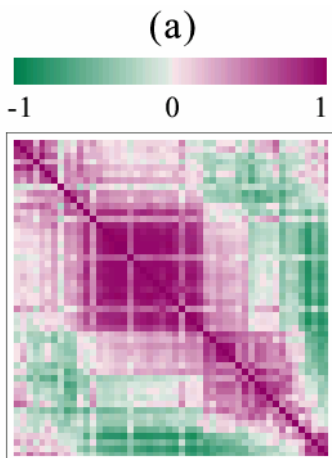
Bipolar disorders

色のスペクトル

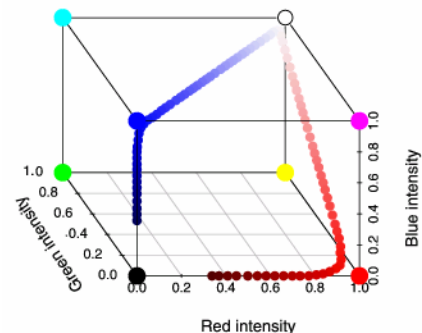
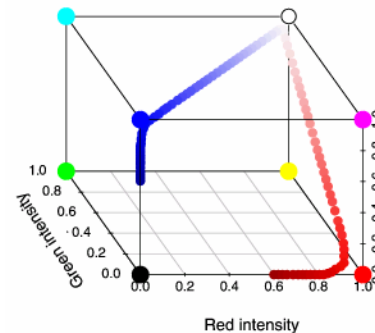
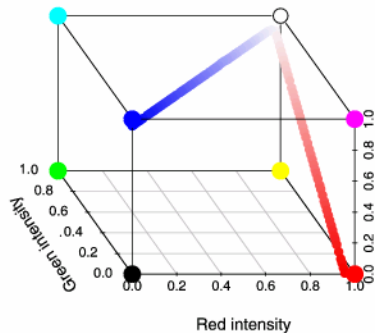
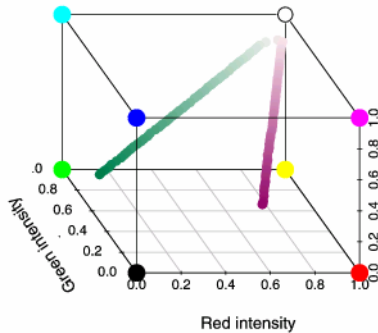
色空間の幾何学と色彩学に関する研究の進展



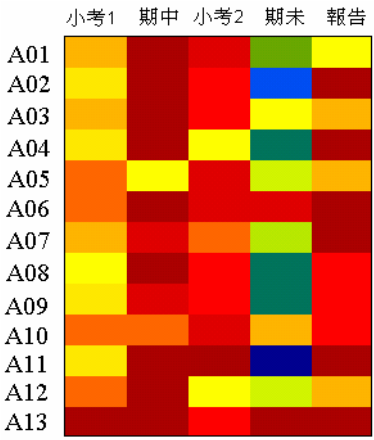
精神疾患の50変数の相関行列図



RGB

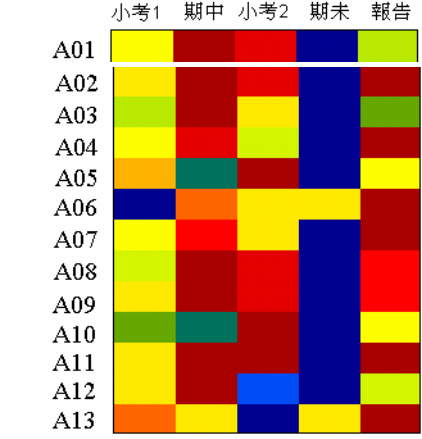
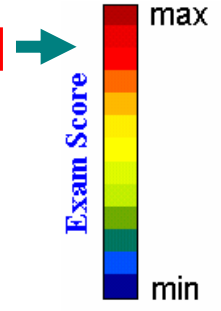


表示の条件

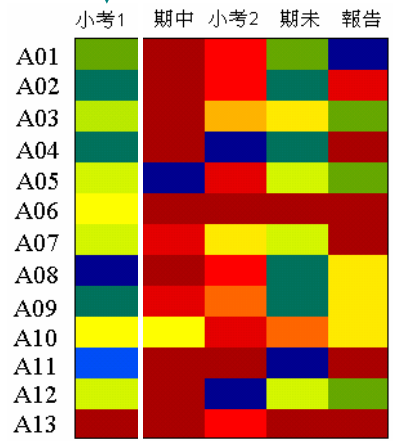


行列範囲の条件

	A	B	C	D	E	F
1	学号	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

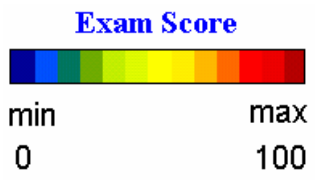


行範囲の条件



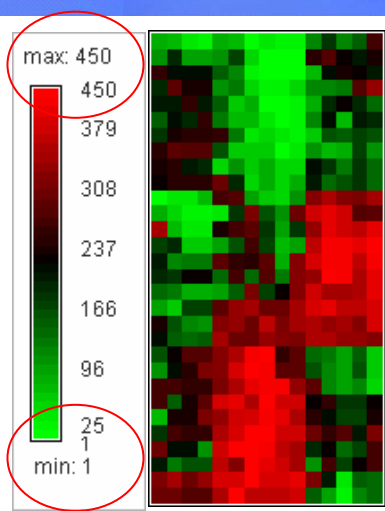
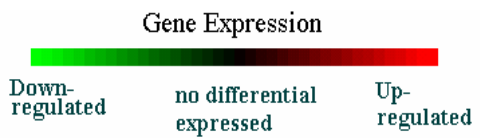
列範囲の条件

これはどうか？

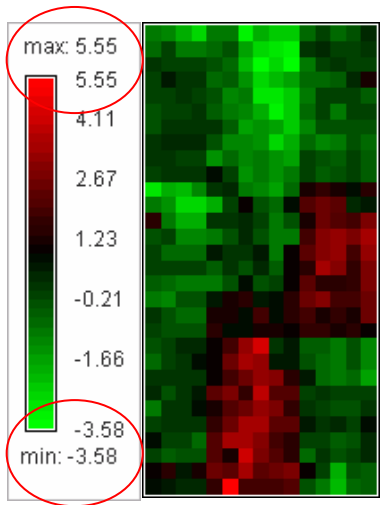
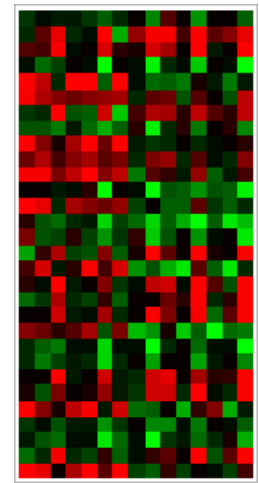


表示の条件

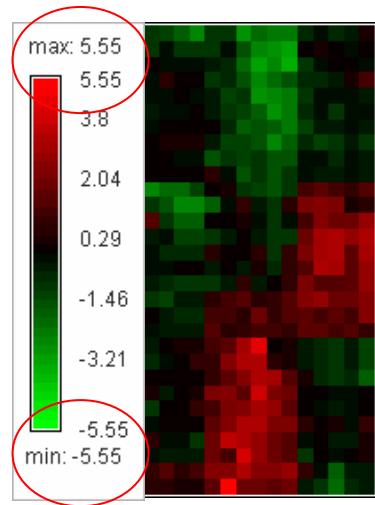
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.88	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.95	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	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	0.78	-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	0.34	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	0.01	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



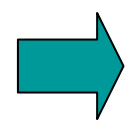
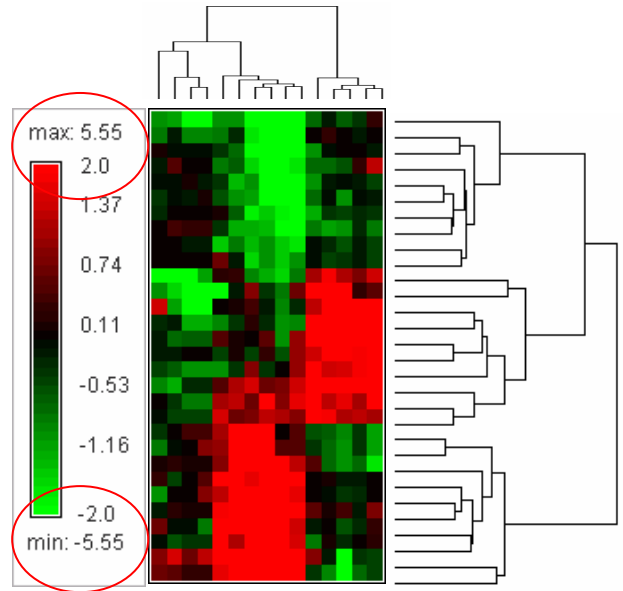
行列ランクの条件



行列範囲の条件



中央値の条件



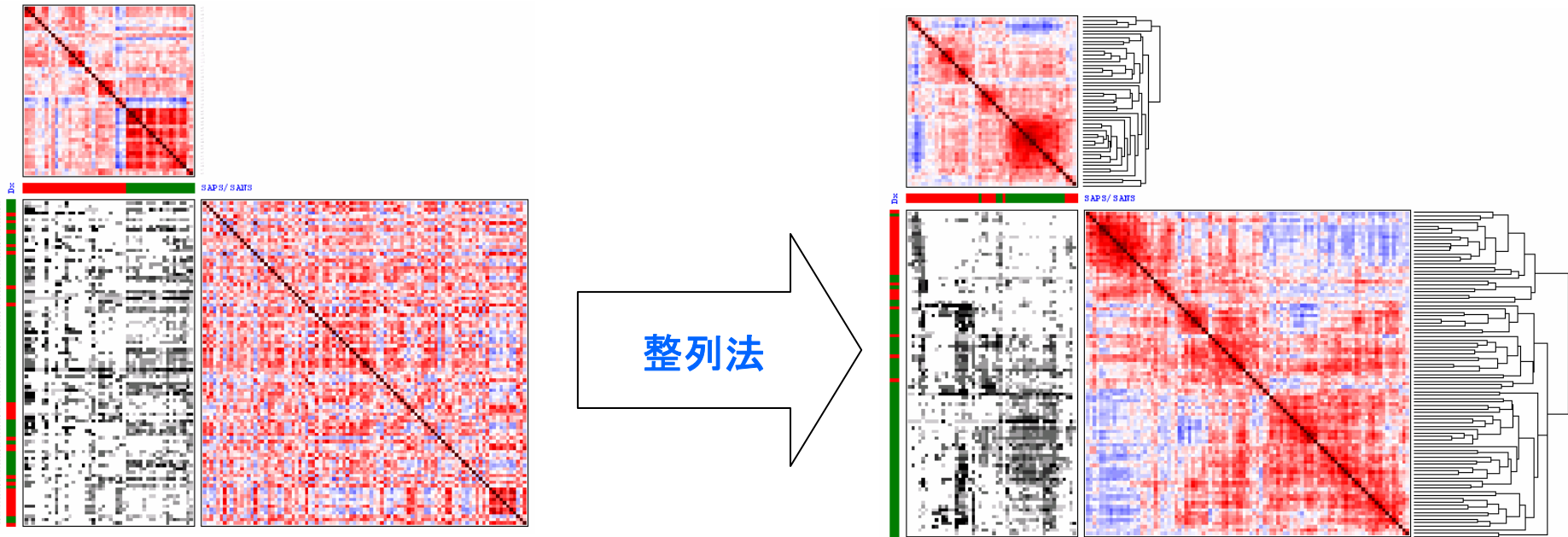
GAPの第2ステップ

類似度行列と生データ行列の整列

- 統計グラフの関連性
- 大域的評価基準
 - GAPランク2楕円整列
- 局所的評価基準
 - 木による整列
 - 木の間ノードの反転

統計グラフの相対性

より近い(遠方の)位置に同様の(異なった)物を配置する



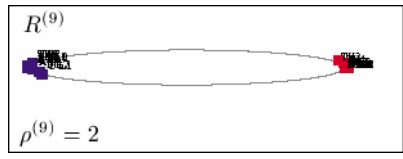
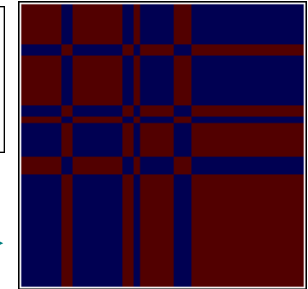
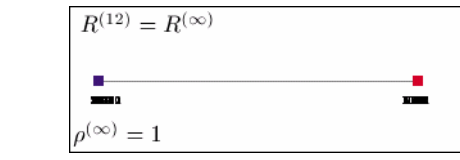
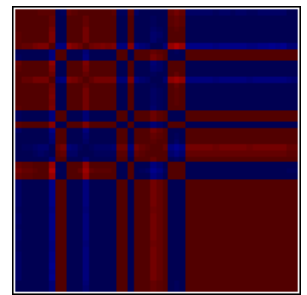
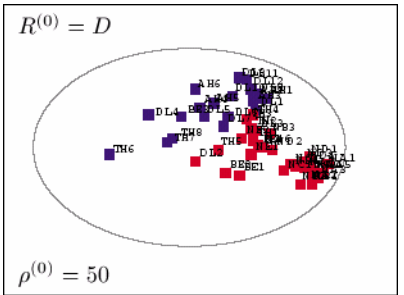
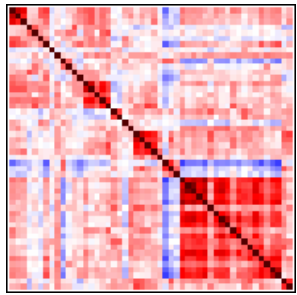
- (1) ランク2楕円整列(チェン、2002)
- (2) 階層的クラスター木(群平均距離)



GAPランク2楕円整列

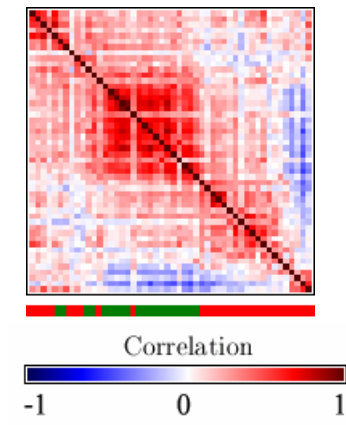
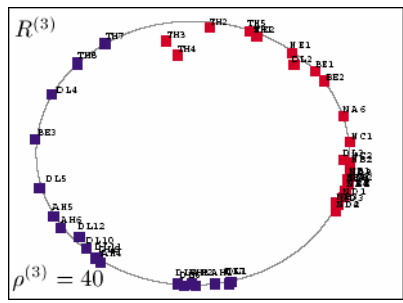
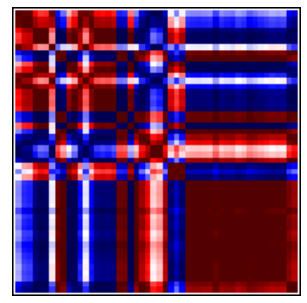
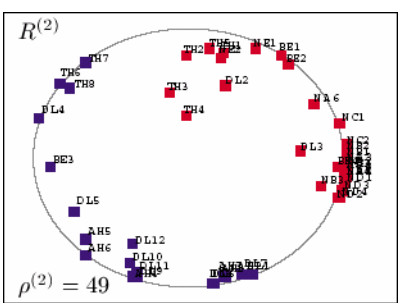
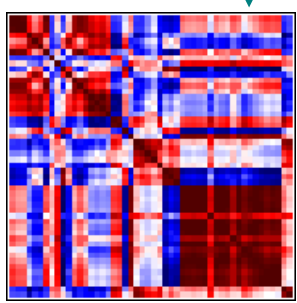
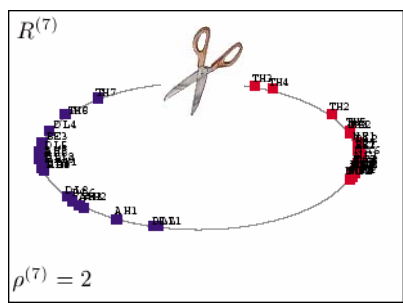
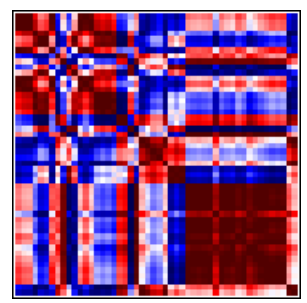
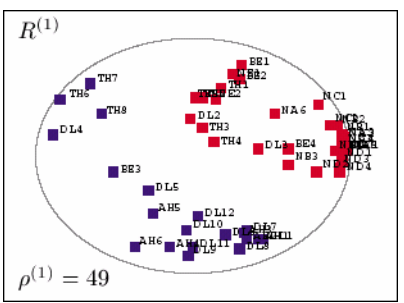
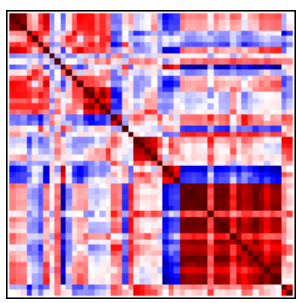
相関行列を収束させる整列アルゴリズム

相関行列(整列なし)



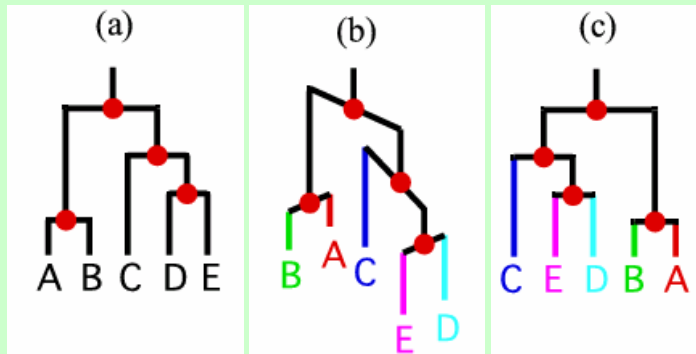
最初の2固有ベクトルへの射影

P個の観測値は楕円上に配置され、特定の相対的位置をとる(チェン2002)

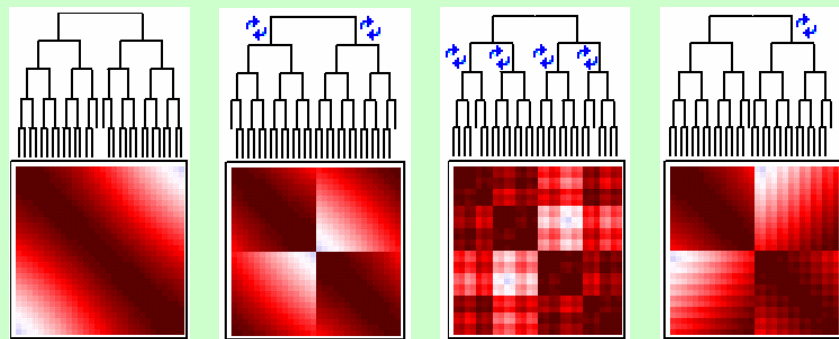


デンドログラムと階層的クラスタ木

木整列

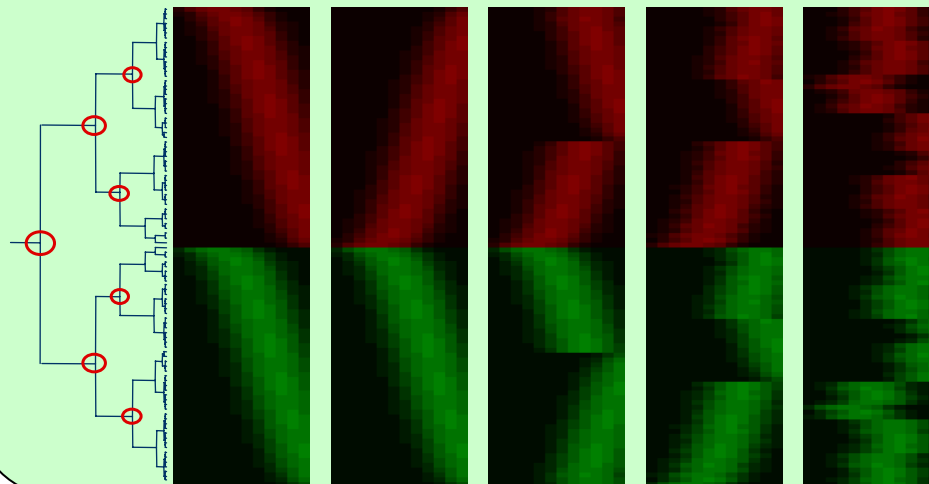


類似度行列の木整列



生データ行列の木整列

理想モデル 1反転 3反転 5反転 多数の反転

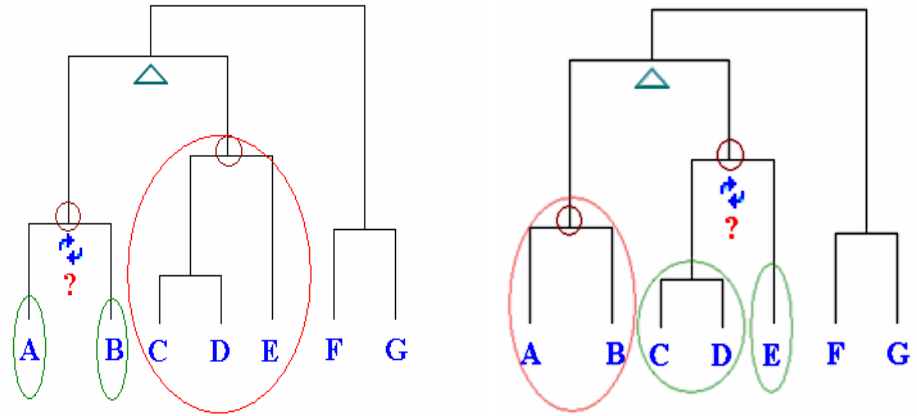


同一の木構造による異なる整列

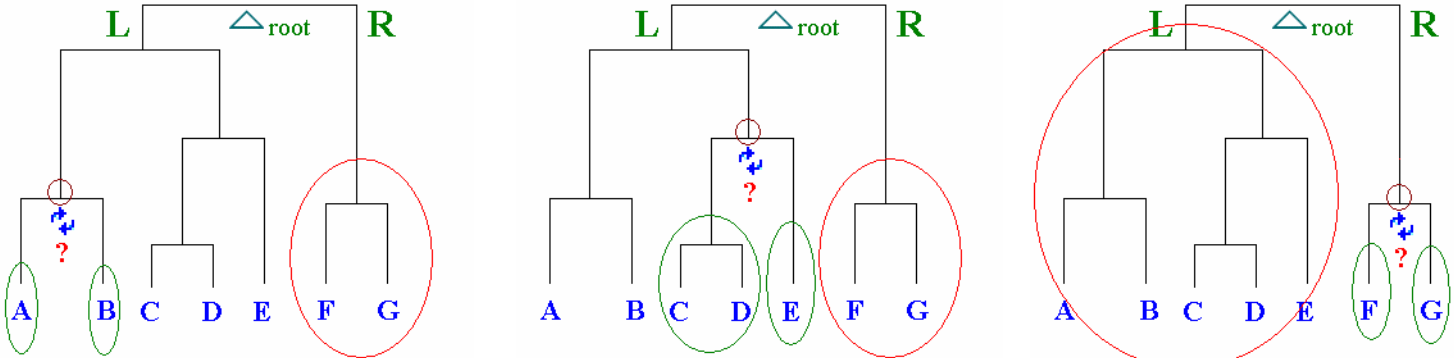
内部木反転

Uncle Approach

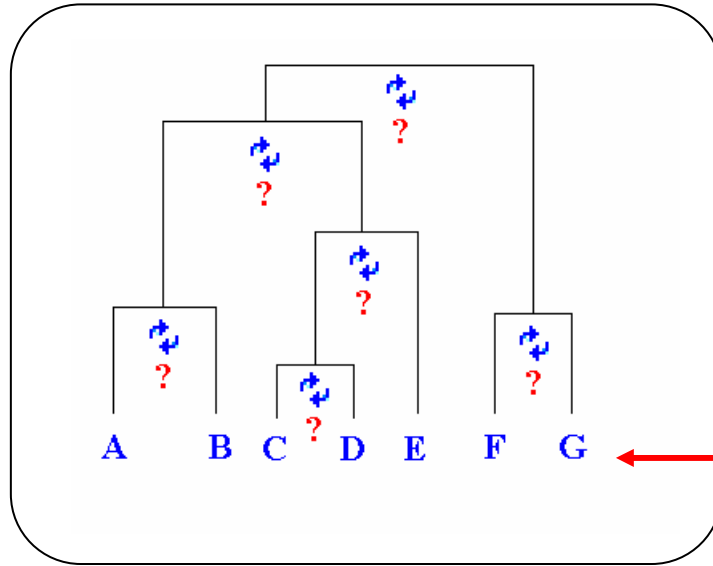
if $d(A, \{C, D, E\}) < d(B, \{C, D, E\})$
then flip



GrandPa Approach



外部木反転



External Ordering

D E A F B C G

できるだけ近い

外部整列の方法

- (1) 平均表現レベルによる(Cluster Software, Eisen et al 1998)
- (2) 一次元SOMの結果を使用する
- (3) ...

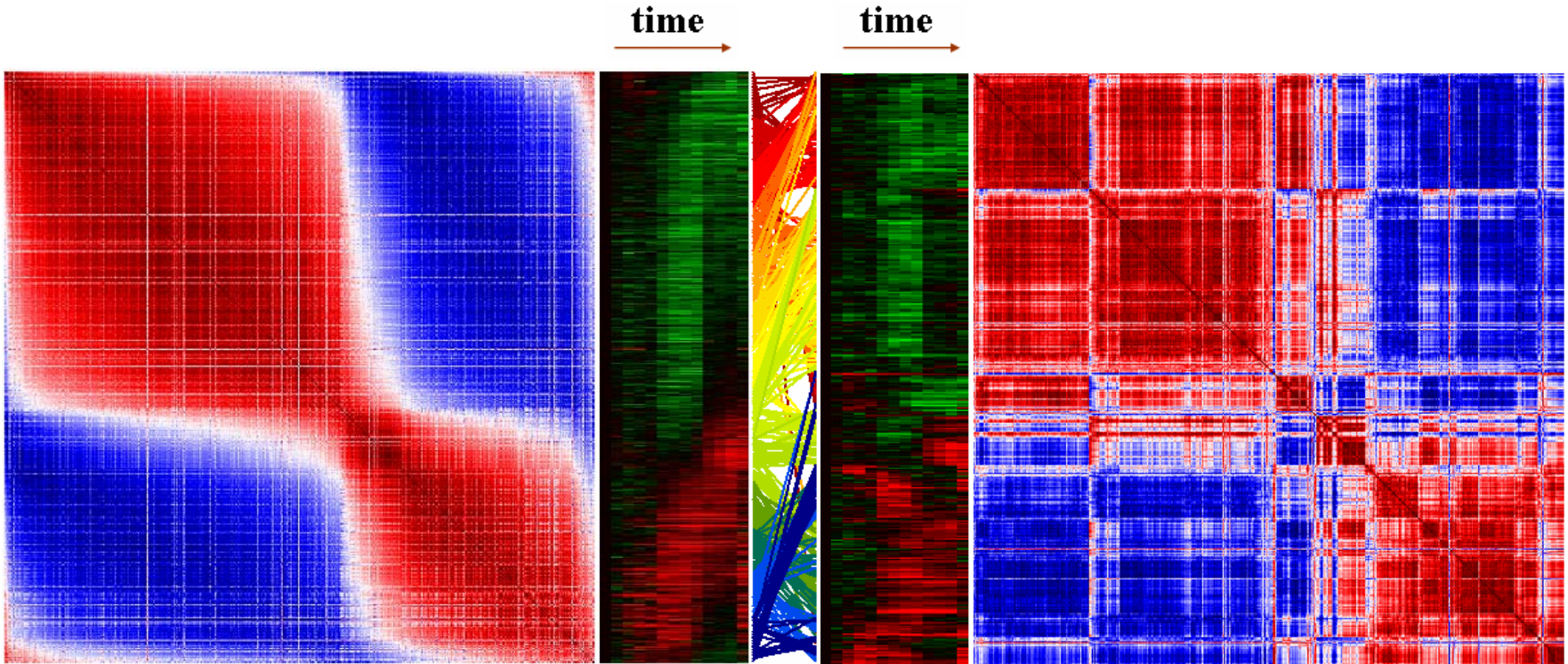
大域的 vs. 局所的整列



データ: 517遺伝子、13アレイ

GAPランク2楕円整列

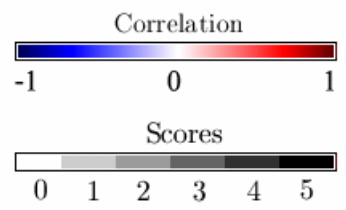
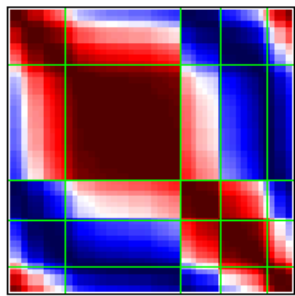
Michael Eisen(1998)木整列



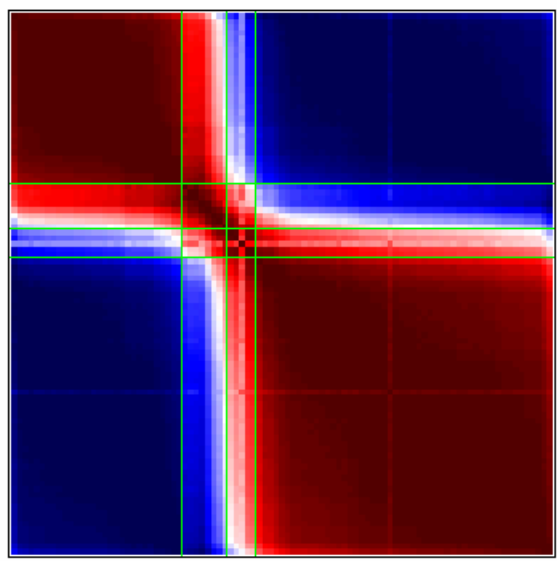
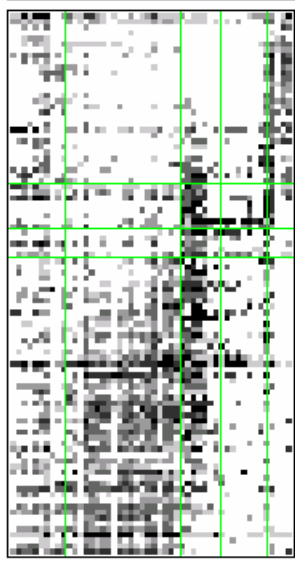
>8>6 >4 >2 1:1 >2 >4 >6>8

-1 0 1

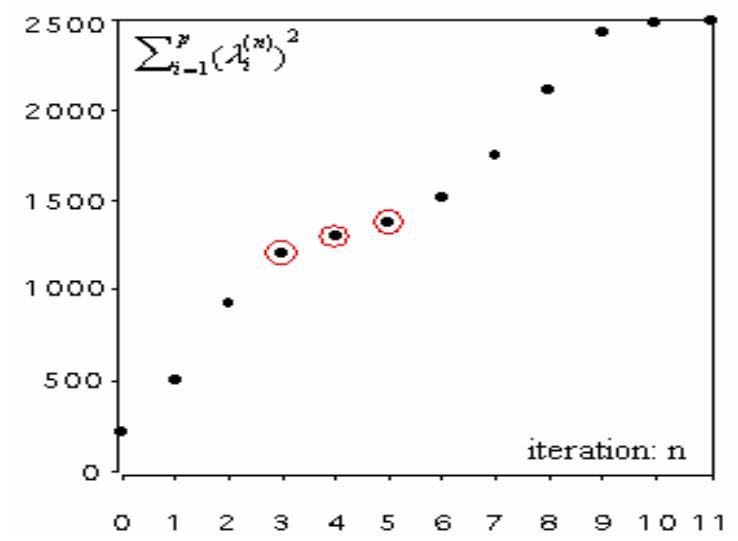
並べ替えられた行列地図の分割



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

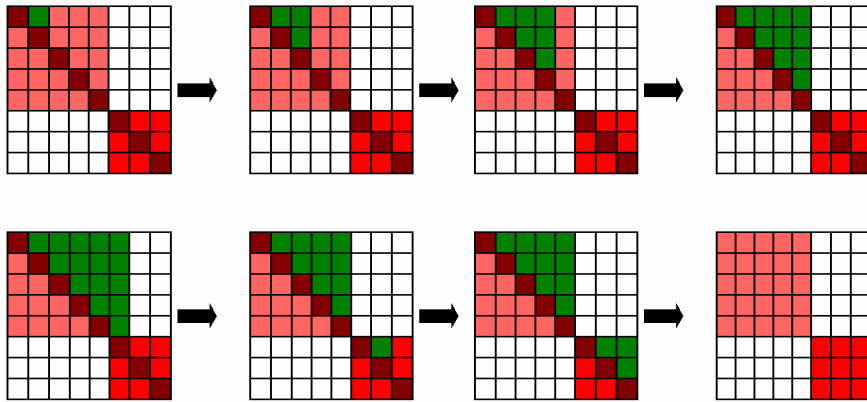


固有値の二乗の合計(相関の二乗の合計)

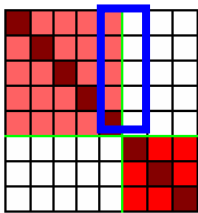


並べ替えられた行列地図の分割

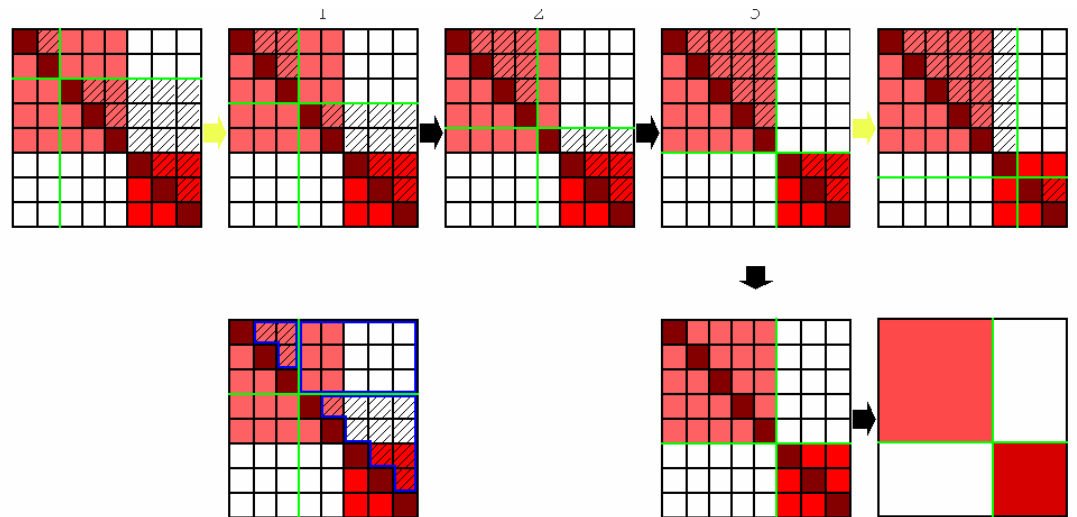
1方向ブロック探索



2サンプル問題



内部二乗合計アプローチ





GAPの第4ステップ

十分グラフ

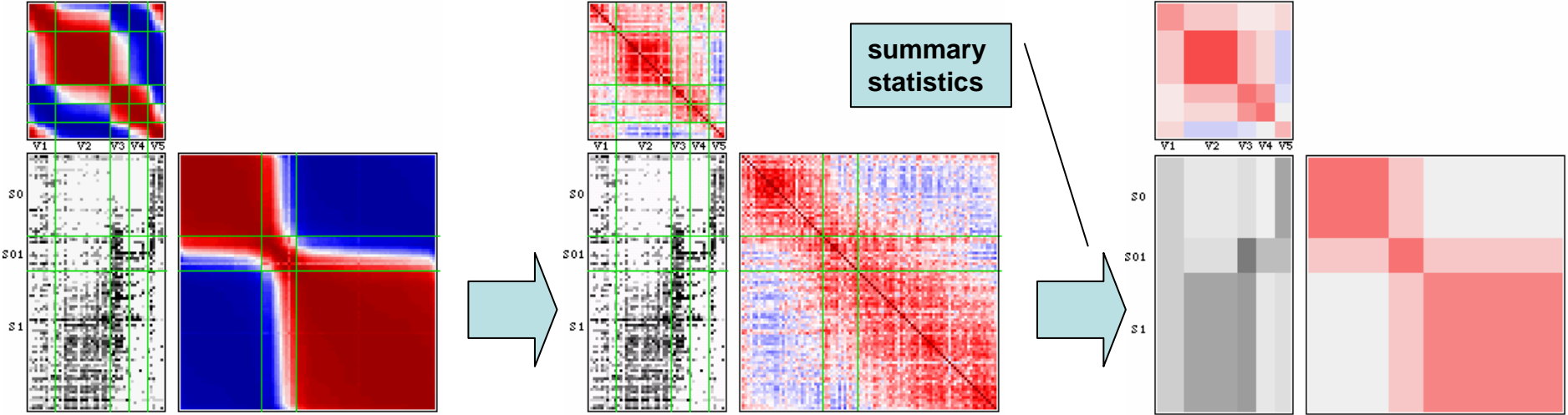
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

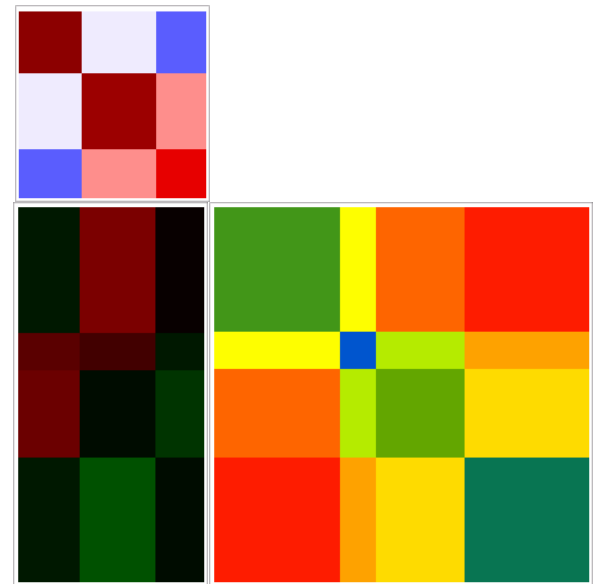
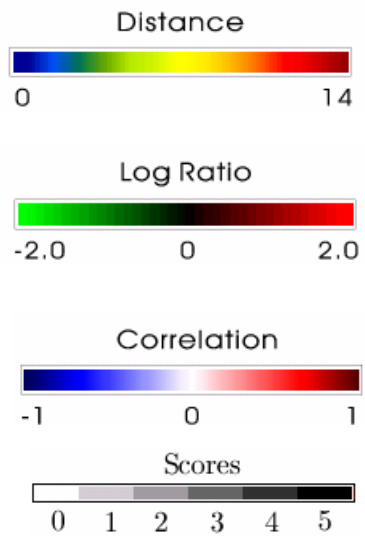
70	低平均	65.67	81.83	73.67	53.67	72
	高平均	77.83	90.67	86.67	53.67	94.17

十分グラフ



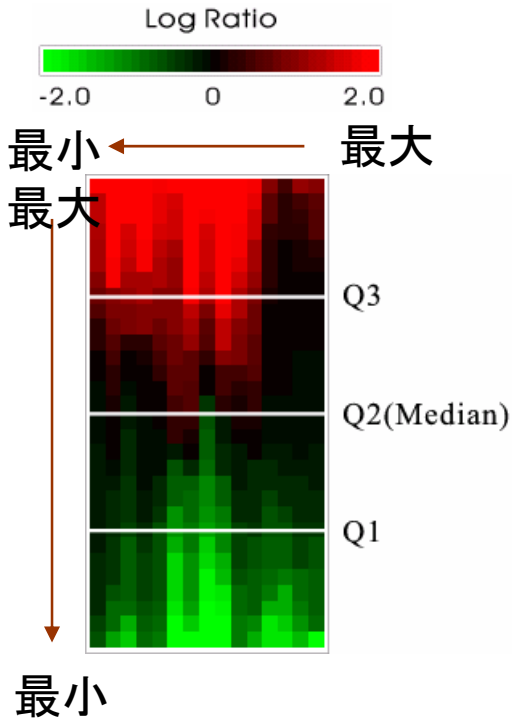
要約表示

- (1) 観測値-観測値
- (2) 変数-変数
- (3) 観測値-変数



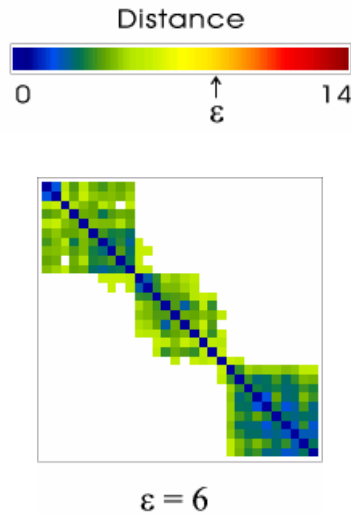
一般化と柔軟性

沈殿表示



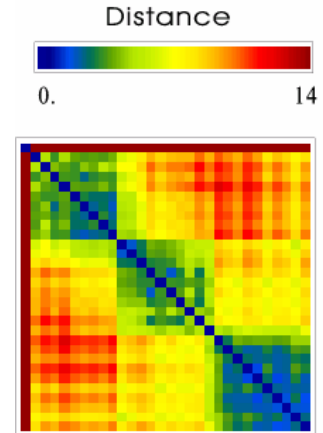
四分位数が計算されたボックスプロットによって与えられたものと同様の情報

部分表示

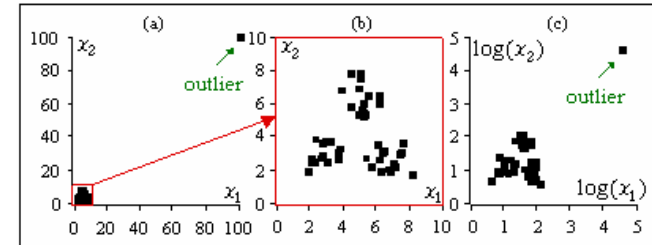


ある条件を満たす数値だけを表示

制限表示

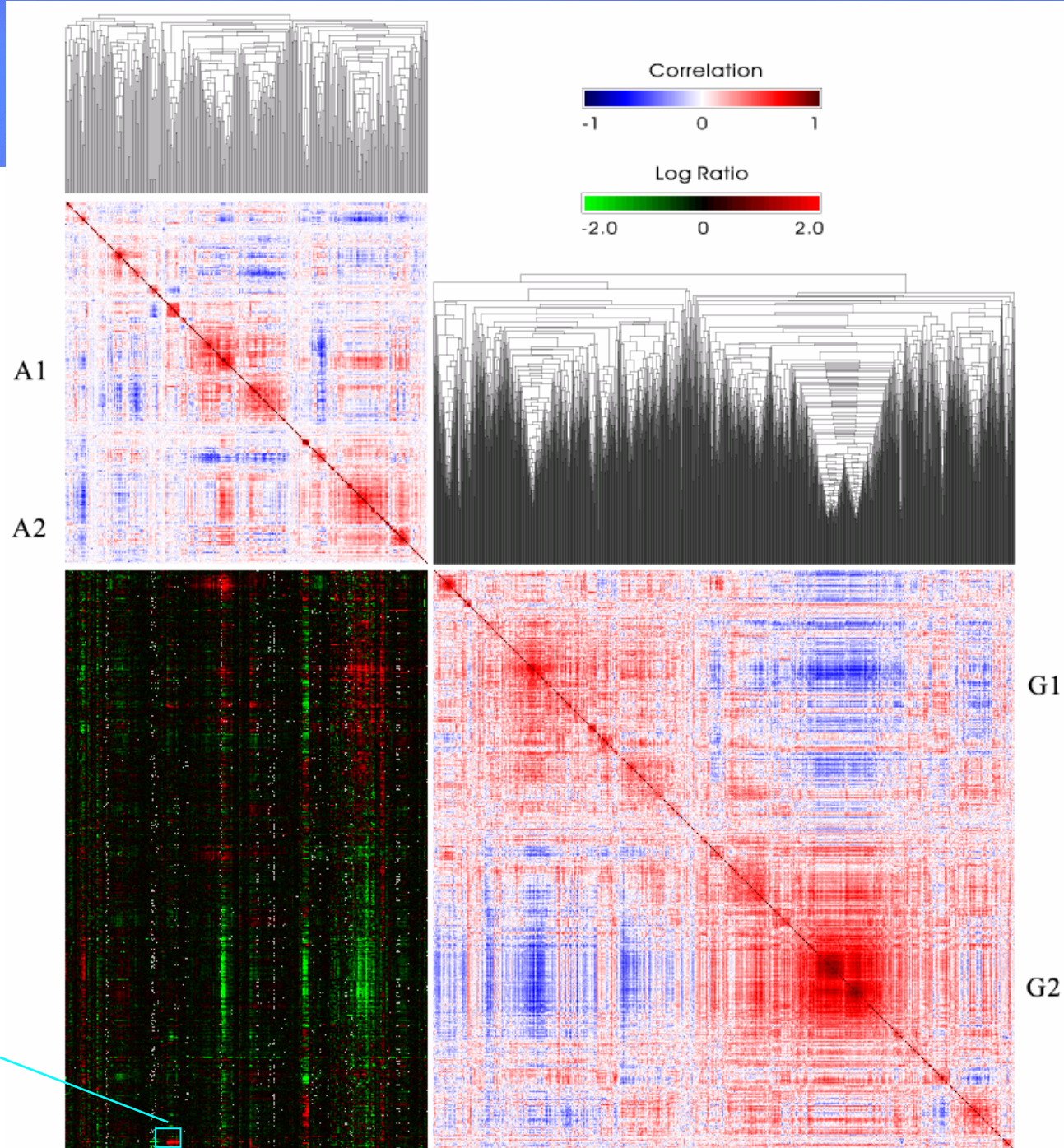


統計グラフの分解能

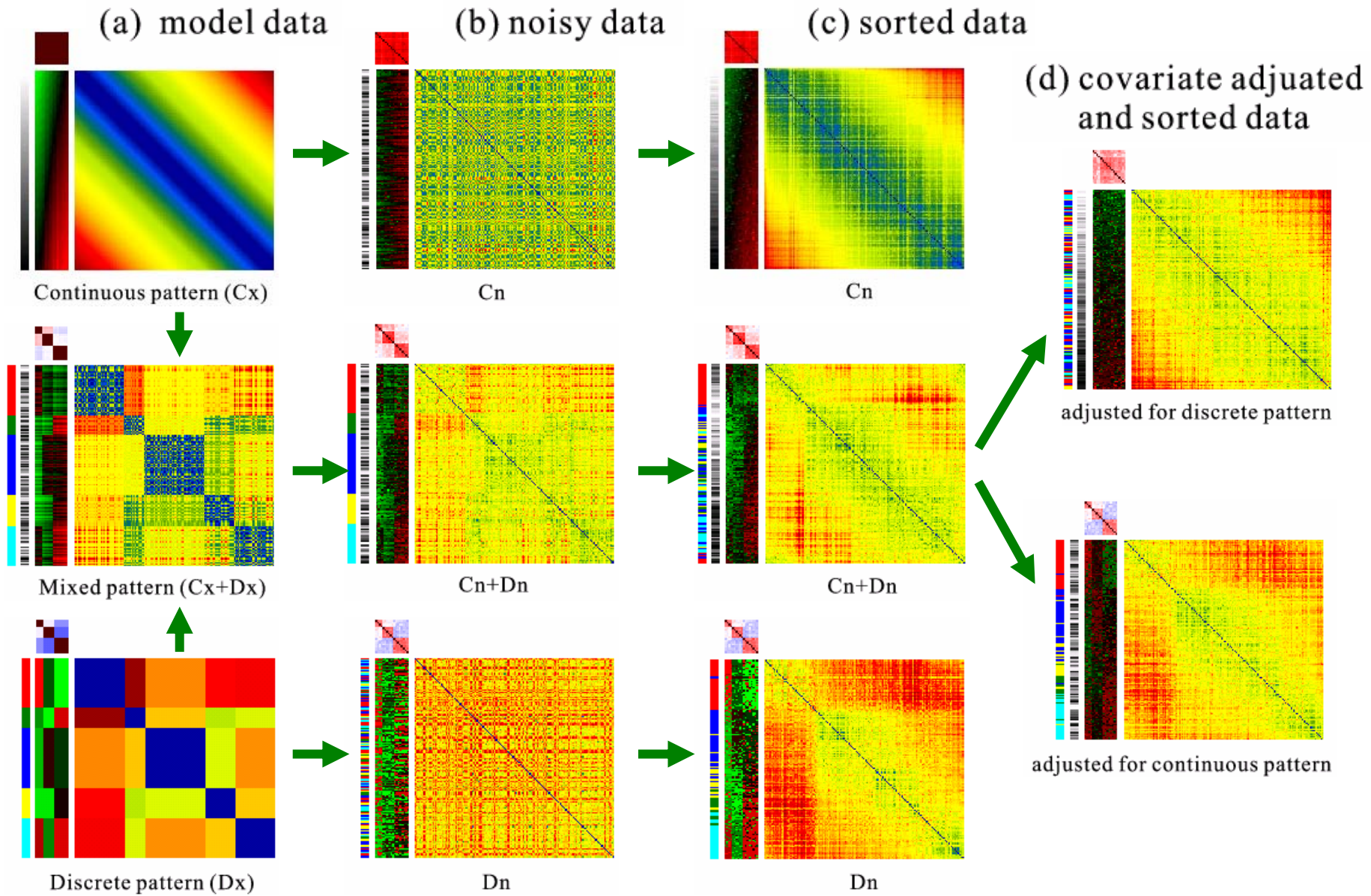


例

- 比較的少ない欠測値がある400アレイの2000遺伝子
- ピアソンの相関係数。
- 群平均クラスター木。
- これらの木分類された行列地図を使用することで、基本的な遺伝子クラスタ構造とアレイの（実験的な）分類パターンが特定できる

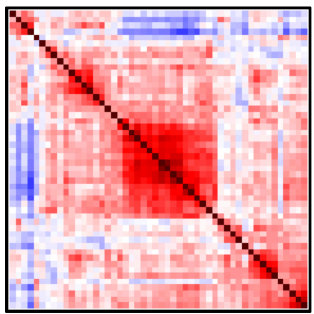


モジュール: 共変量調整のためのMV



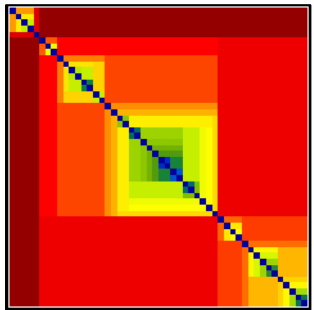
モジュール: 行列可視化による階層的クラスタリングのための対話的診断システム

(1) 類似度行列を入力



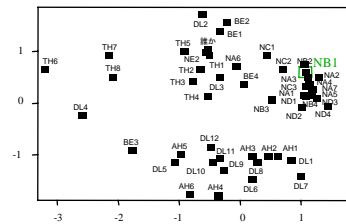
(例: ピアソンの相関係数)

(例: Cophenetic行列)

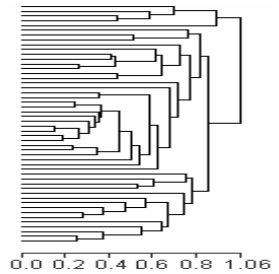


(3) 距離行列を出力

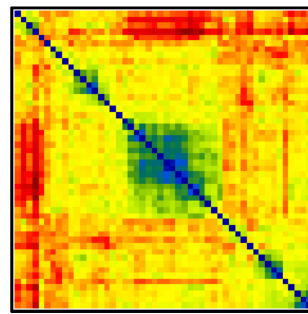
統計的モデリング
多次元の尺度構成法(MDS)
階層的クラスタリング木(HCT)



階層的クラスタリング木(HCT)

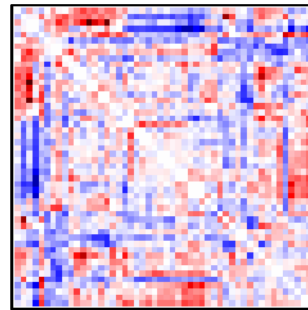


(2) 変換された非類似度行列

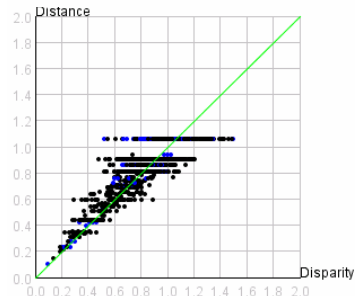
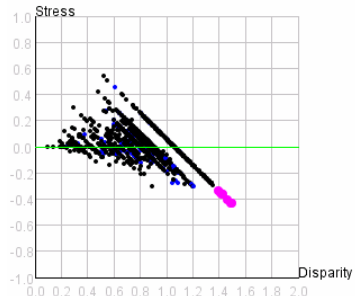
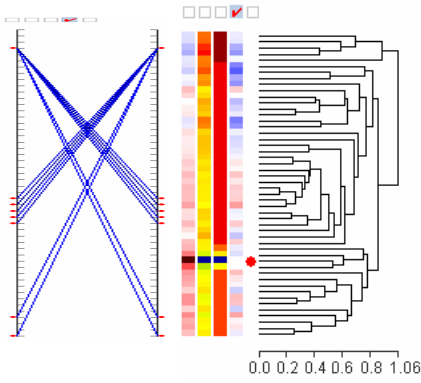


(例: 距離)

(例えば、残差行列)



(4) ストレス行列



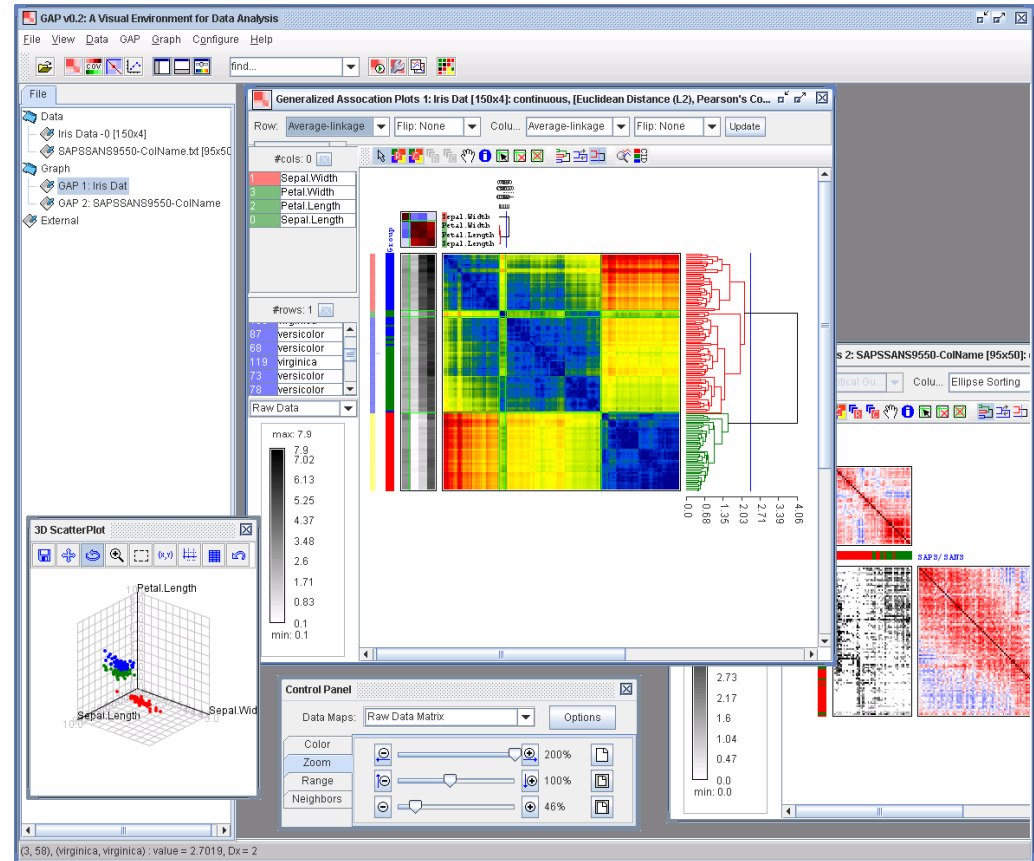
一般相関プロット

- データタイプの入力する:
連続または2進
- 様々な整列アルゴリズム
とクラスタリング分析
- 様々な表示の条件
- モジュール: 共変量で調
整されているGAP、非線
形相関、欠測値補填



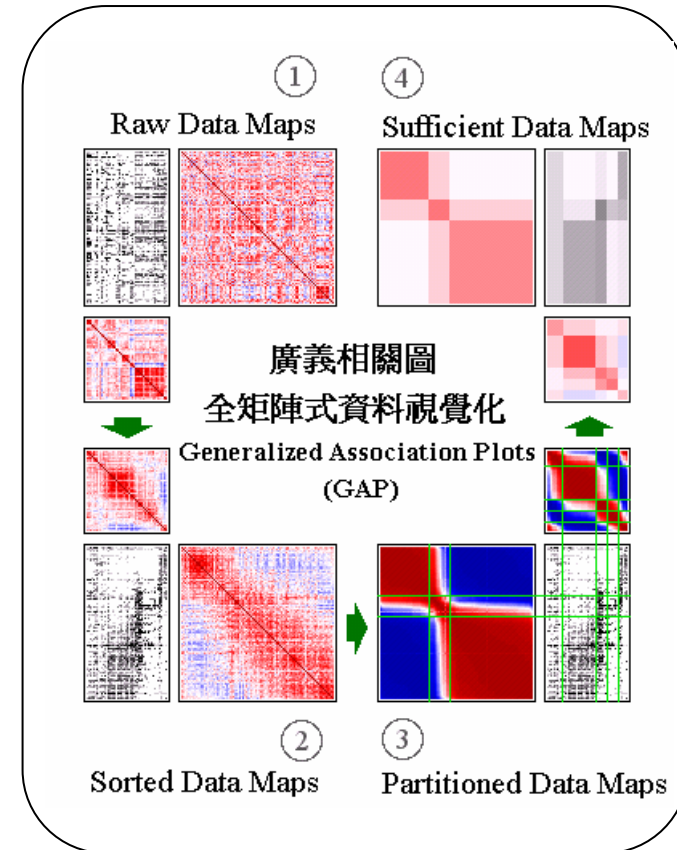
統計的プロット

- 2D 散布図、(回転可能
な)3D散布図



MVはデータ行列の色による整列表現である
MV表示は5つのレベルの情報を提供する:

1. あらゆるサンプル/変数の組み合わせの生スコア
2. すべての変数に対する個々のサンプルのスコアベクトル、およびすべてのサンプルに対する個々の変数のベクトル
3. あらゆるサンプル-サンプルと変数-変数の関係のための相関スコア
4. 変数の組分け構造とサンプルのクラスタリング効果
5. 変数グループにおけるサンプル-クラスタの相互パターン



- 現代的探索的データ解析における予備ステップ
- 研究と応用のための継続的でアクティブな話題
- 新世代の探索的データ解析(EDA)ツール

ウェブサイト



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 **Lab for Information Visualization** 中央研究院 統計科學研究所
資訊視覺化研究室 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

情報可視化研究室

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

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...

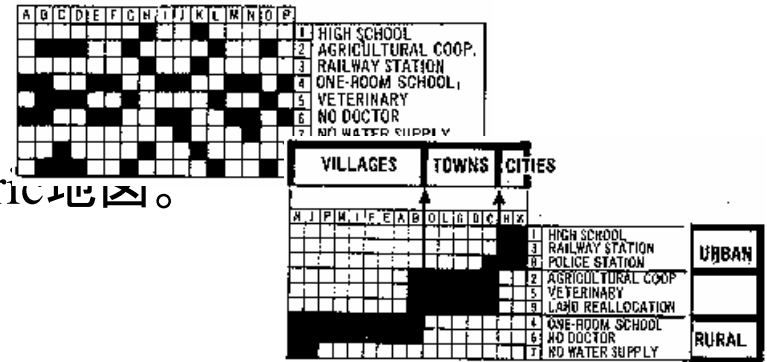
myLinks

- 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二段128號, Tel. +886-2-27835611 ext: 309

概念:

- ベルタン(1967): 整列可能行列
- カーマイケルとスニース(1969): taxometric地図。



データアレイのクラスタリング:

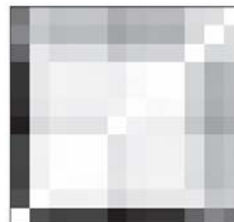
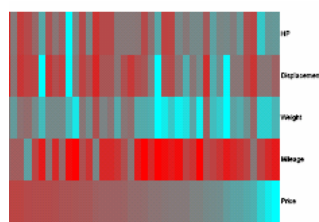
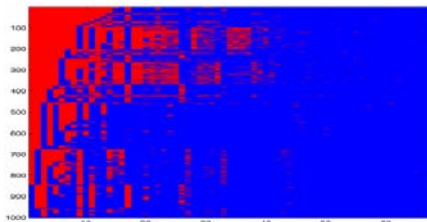
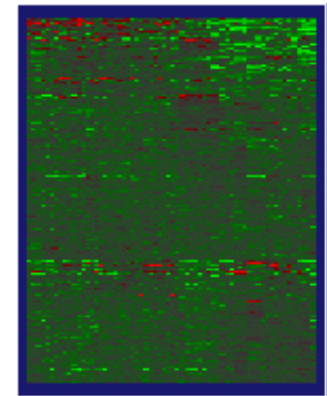
- Hartigan(1972): データ行列の直接クラスタリング
- Tibshirani(1999): ブロッククラスタリング
- Lenstra(1974): 巡回セールスマン問題
- スレイグル他 (1975): 最短経路

色表現:

- ウェッグマン(1990): カラーヒストグラム
- Minnotte and West(1998): データイメージ
- Marchette and Solka(2003): 外れ値検出

4. UN VOTES IN 1969-1970*

State	EASE	HUNG	CHINA	KORSA	SO AF	PAPUA								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
USR	1	1	1	1	2	3	1	3	2	2	1	3		
BGA	1	1	1	1	1	3	2	2	1	3				
YUG	1	3	3	3	1	1	3	1	2	3	1	1	2	
SYR	1	2	2	3	1	1	3	1	2	3	1	1	3	
UAR	1	3	3	3	1	1	3	2	2	3	1	1	3	
KEN	1	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	3	1	3	1	1	2	
DAH	1	3	3	3	1	3	1	3	1	3	1	2	2	
UBA	1	3	3	3	1	3	1	3	1	1	3	3	1	
UNK	1	3	3	3	1	3	2	3	1	1	3	3	1	
FRA	1	3	3	3	5	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	4	1	2	3	3	1	1	3	3	1
ALA	1	3	3	3	1	3	1	3	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	1	2	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	1	1	1



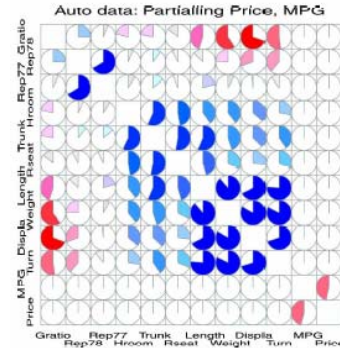
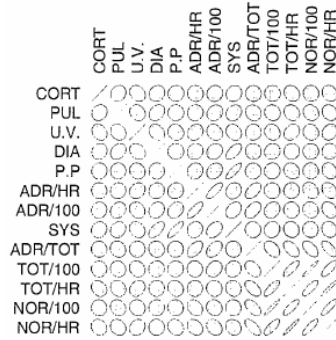
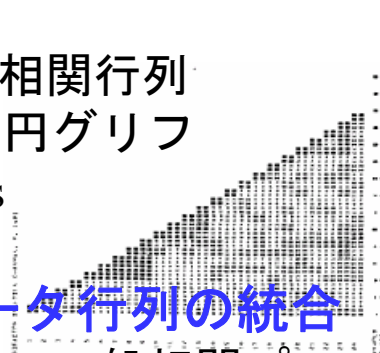
関連(続き)

類似度行列だけを探る:

- Ling(1973): 陰影をつけられた相関行列
- マードックとチャウ(1996): 楕円グリフ
- フレンドリー(2002): corrgrams

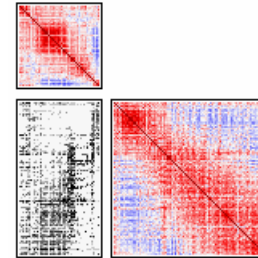
2つの類似度行列による生データ行列の統合

- チェン(1996、1999および2002): 一般相関プロット(GAP)



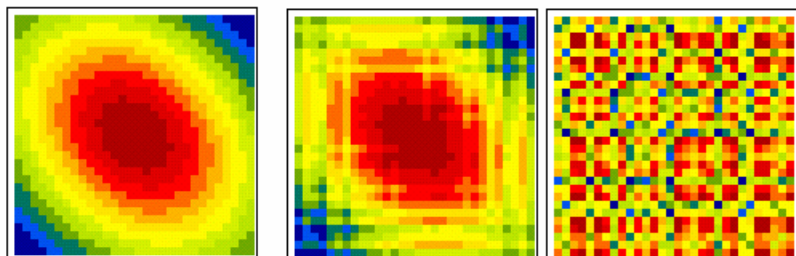
変数とサンプルの整列

- チェン(2002): 統計グラフの関連性の概念
- フレンドリーとクワン(2003): データ表示の効果整列
- ハーリー(2004): おもしろい表示を際立った位置に置く



行列可視化(MV): 再生列行列、heatmap、色ヒストグラム、データイメージおよび行列可視化

良い並び替えの評価基準



Robinson

pre-Robinson

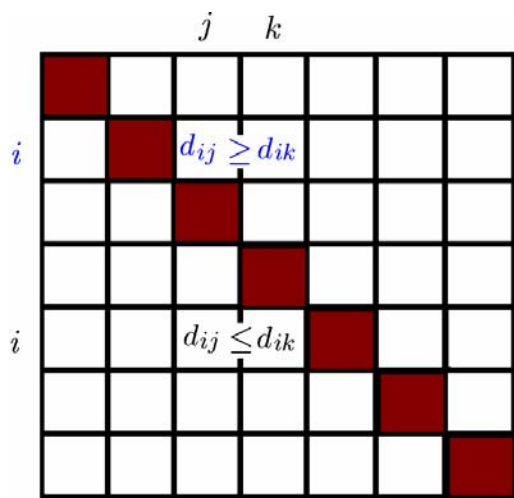
グローバルな評価基準: 非ロビンソン測度

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

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

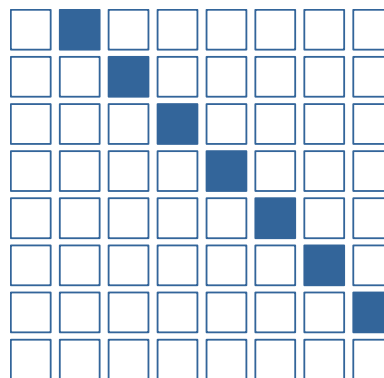
$$AR(w) = \sum_{i=1}^p \left[\sum_{j < k < i} I(d_{ij} < d_{ik}) |j - k| |d_{ij} - d_{ik}| + \sum_{i < j < k} I(d_{ij} > d_{ik}) |j - k| |d_{ij} - d_{ik}| \right].$$

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



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

局所的评价基準: 最小長さ損失関数



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

階層的なクラスタリング木

(クープマンとRousseeuw、1990)

例: 平均リンテージ

距離行列

	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



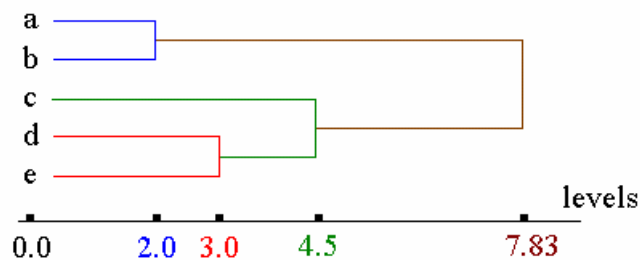
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0



	{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



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

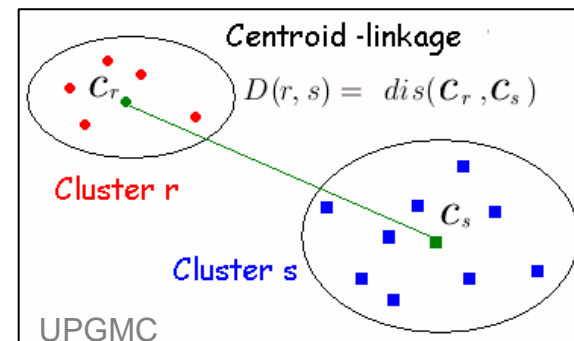
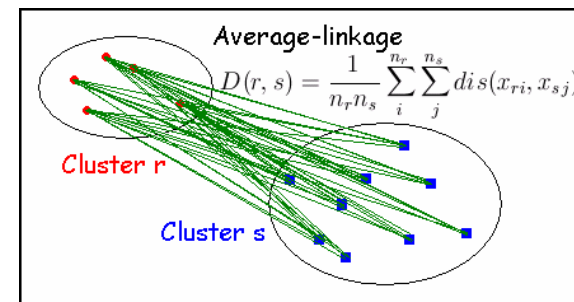
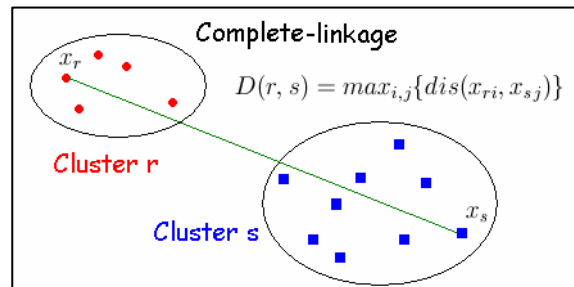
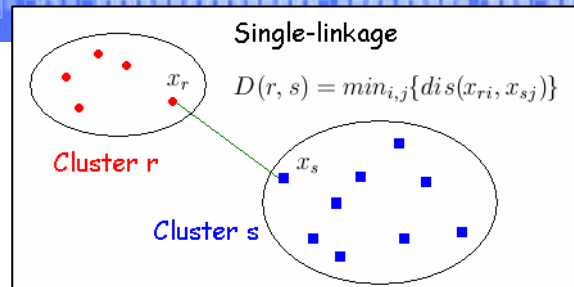
$$= \frac{1}{2}(6 + 5) = 5.5$$

UPGMA(重みのない
対グループ法平均)

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

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

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



- 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.