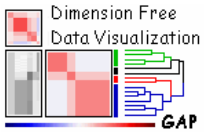
入門

一般相関プロット（GAP）
- 次元によらないデータ可視化 -

呉漢銘 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/

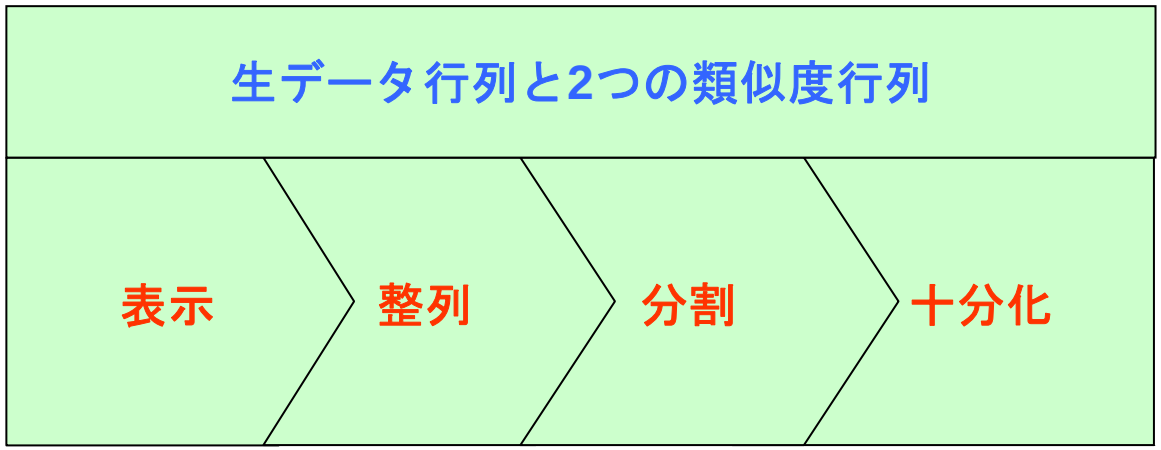


中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

2006/12/22

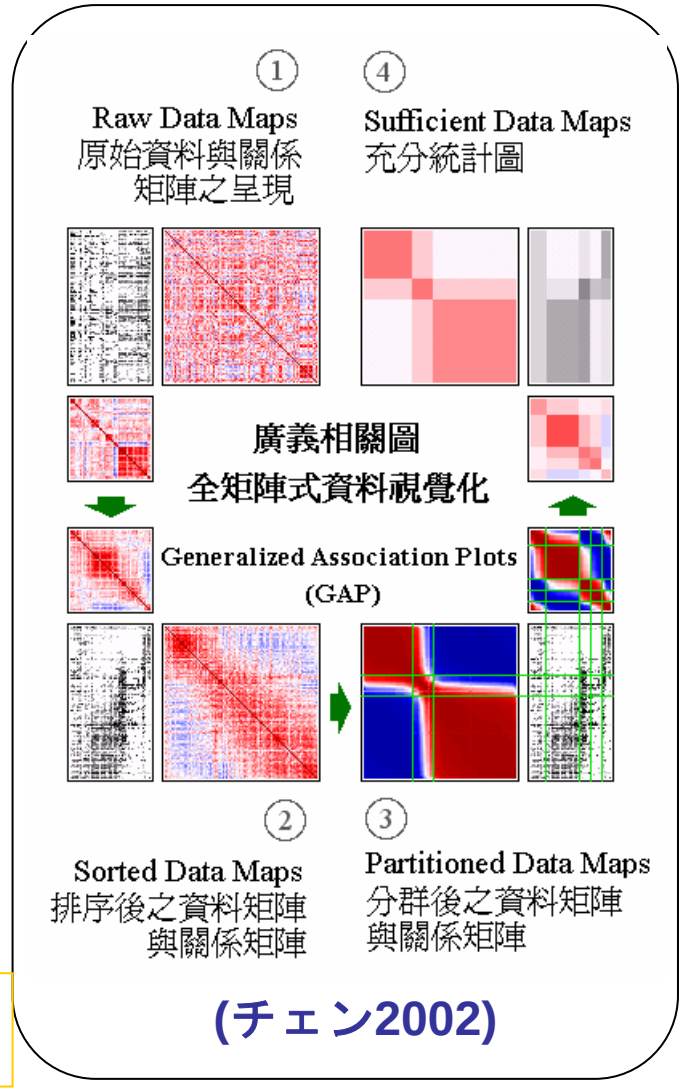
アウトライン

- 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

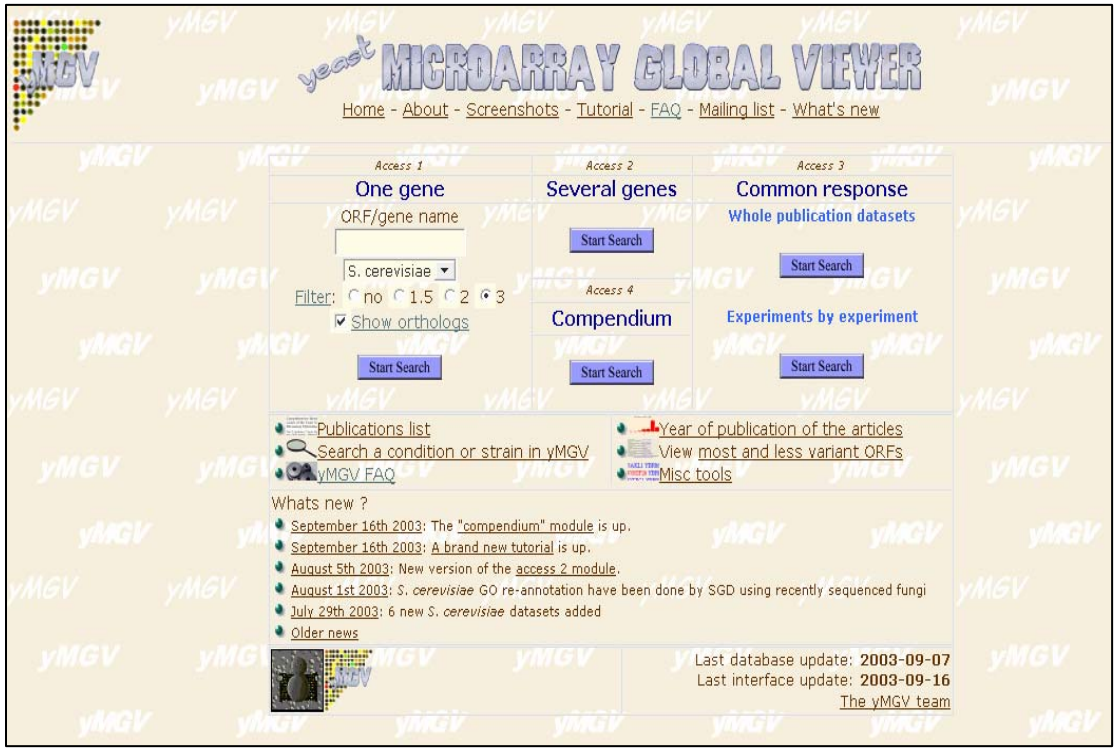
注意：　行列可視化(MV): 再整列行列、heatmap、色ヒストグラム、データイメージと行列可視化



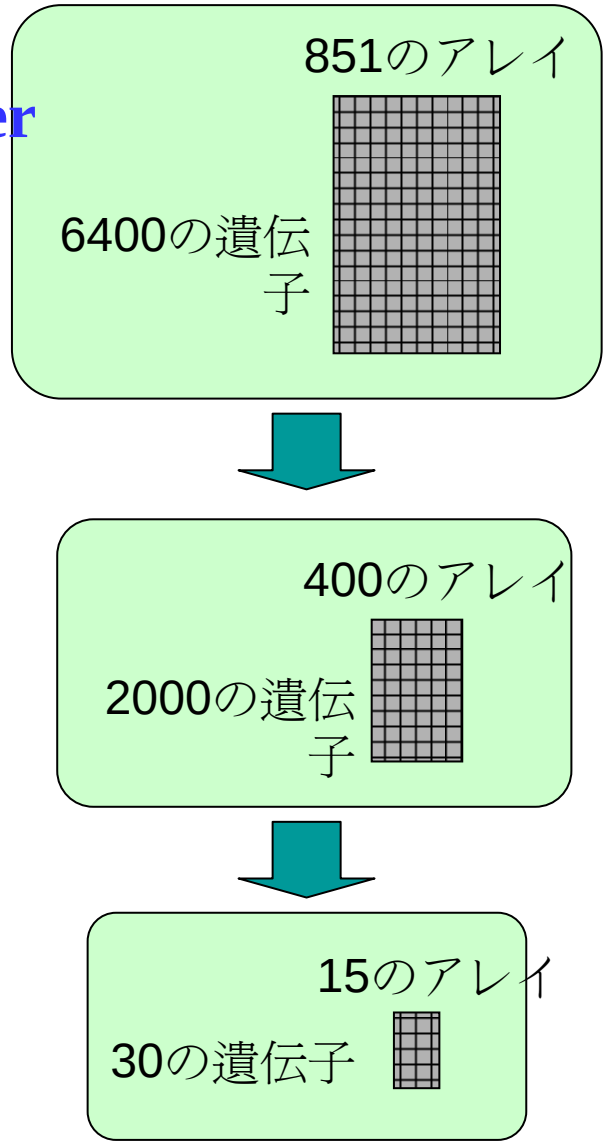
(チェン2002)

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/



精神疾患データ

The Andreasen's Positive and Negative Symptom Table

幻覚(AH1-6)

振舞い(BE1-4)

妄想(DL1-12)

思考錯乱(TH1-8)

陽性症状の査定(SAPS):
30項目,
4つのサブグループ

陰性病状の査定(SANS)
20の項目、
5つのサブグループ

表現(NA1-7)

スピーチ(NB1-4)

衛生(NC1-3)

活動(ND1-4)

不注意(NE1-2)

すべての病状が
6ポイント(0-5)
で記録される

生データ行列

95個の対象

69 統合失調症

26 バイポーラ異状

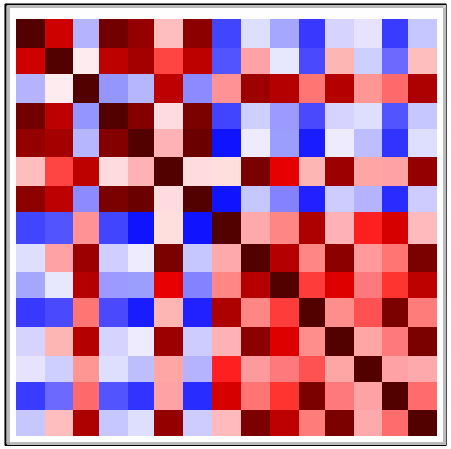
GAPの最初のステップ

生データ行列の表示

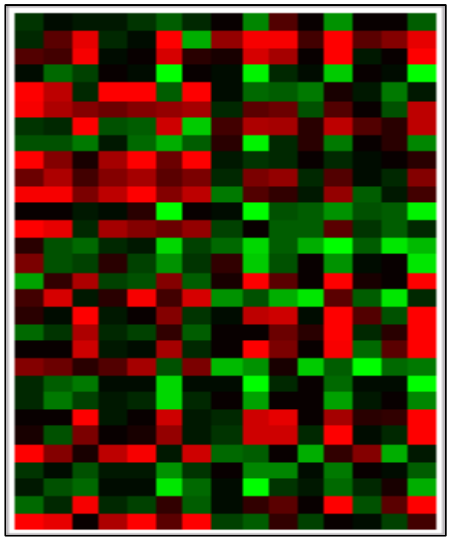
- データ変換
- 類似度の選択
- 色スペクトル
- 表示の条件

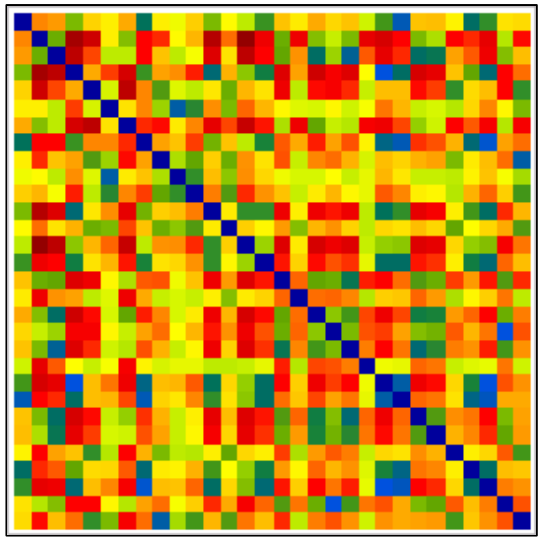
生データ行列の表示

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

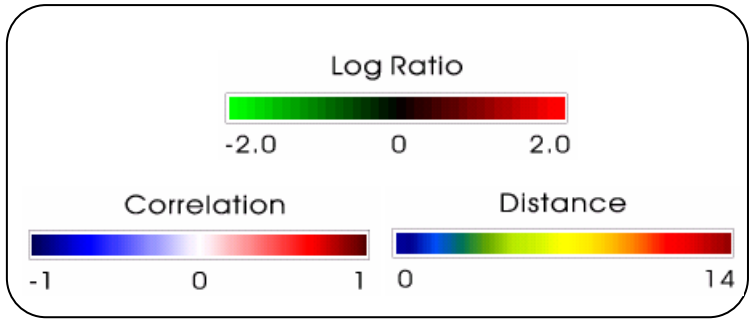
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

ユークリッド距離

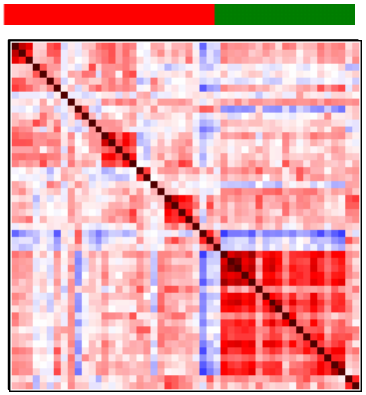
$$d_{xy} = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

他の類似性/相違性測度

(2) 色のスペクトル

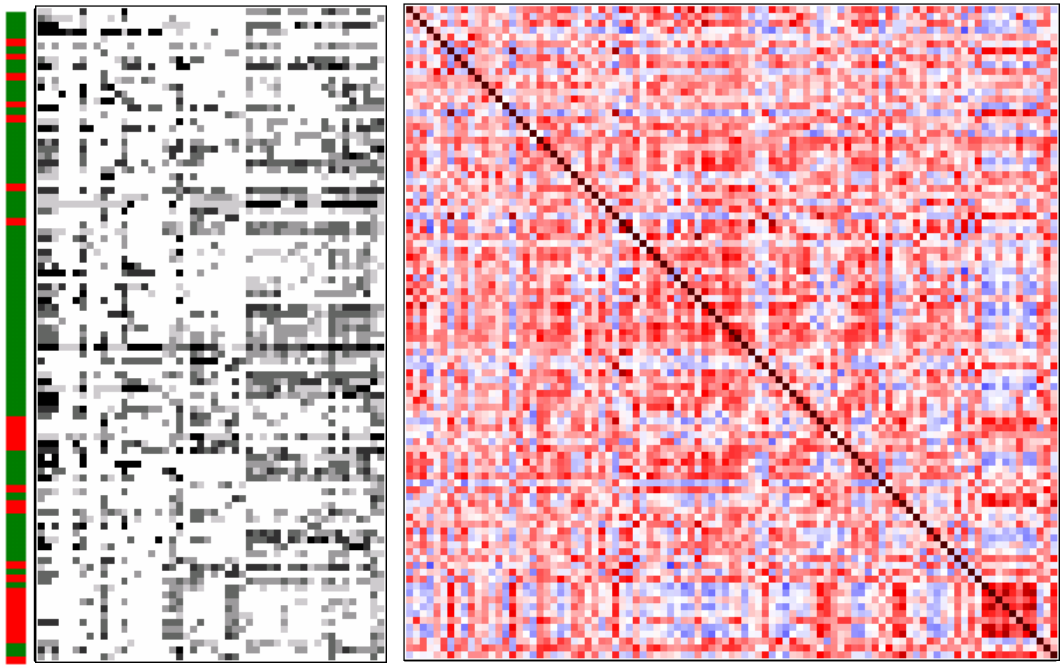


生データ行列の表示



変数の相関行列

観測値の相関行列

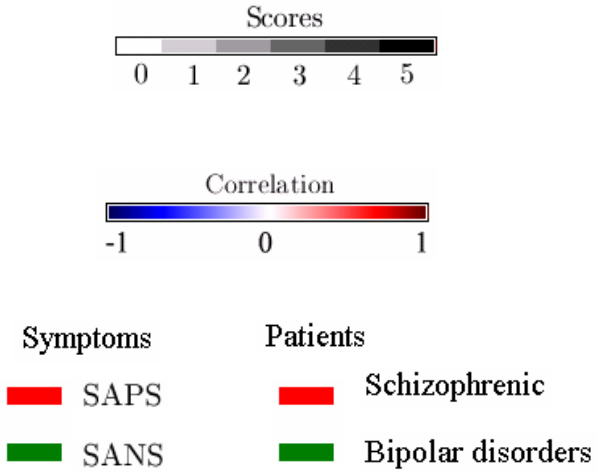


生データ行列

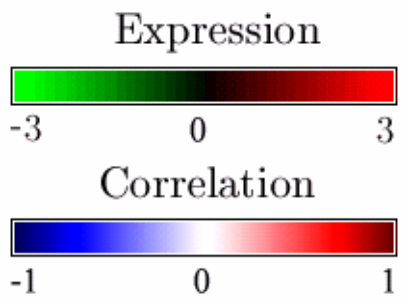
(1) 類似度の選択

ピアソン相関係数

(2) 色のスペクトル

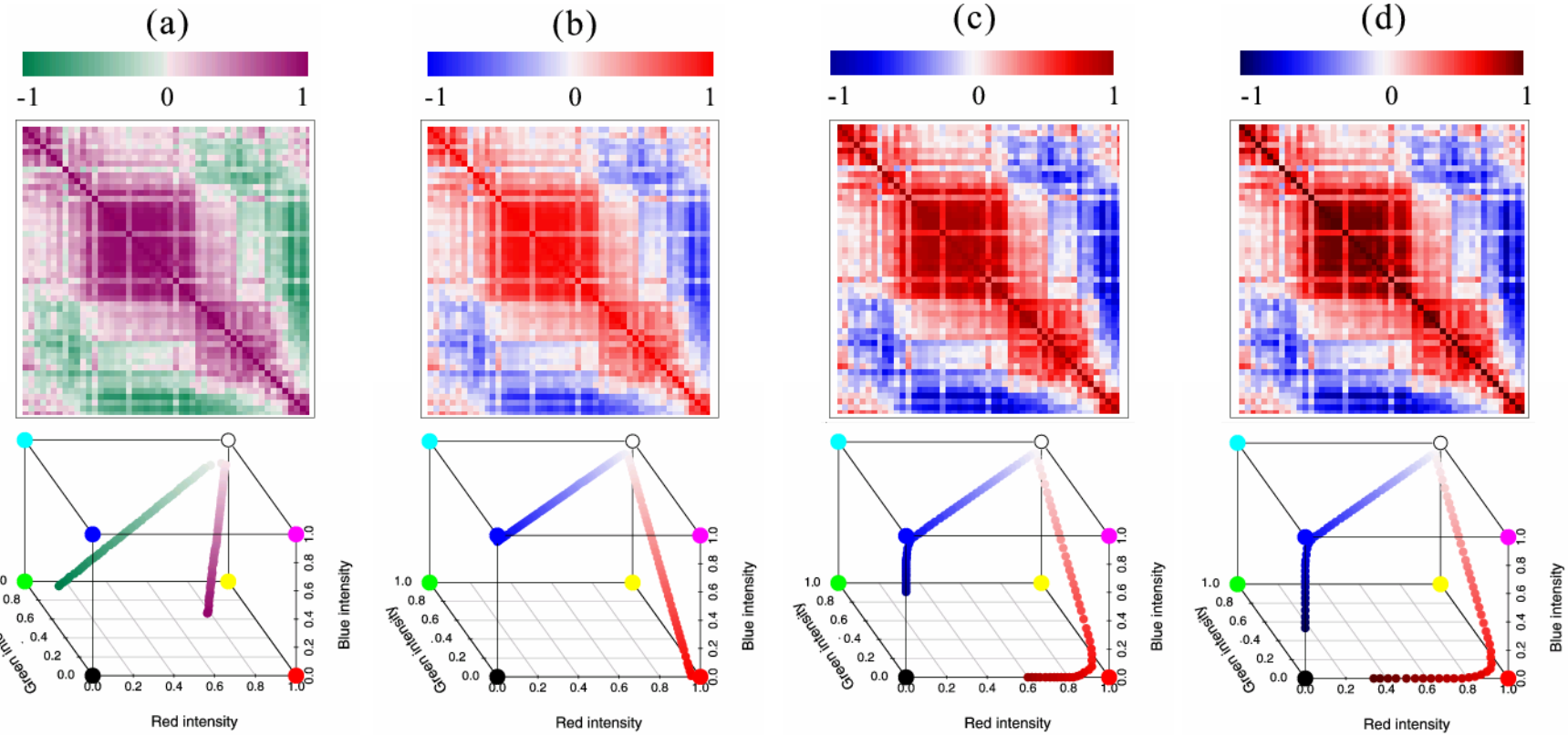


色のスペクトル

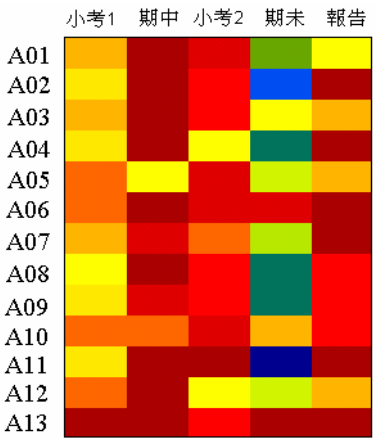


精神疾患の50変数の相関行列図

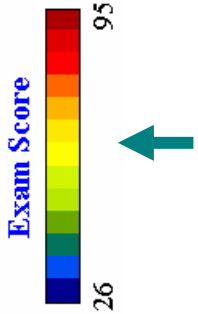
RGB



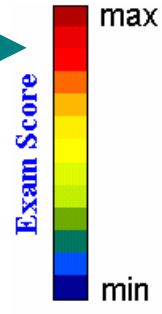
表示の条件

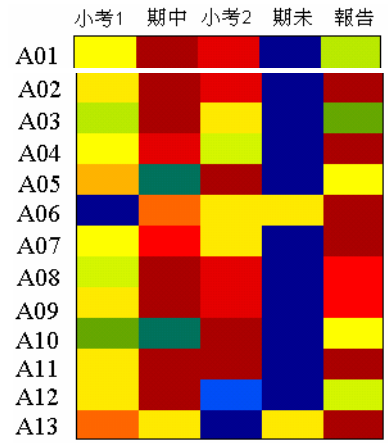


行列範囲の条件

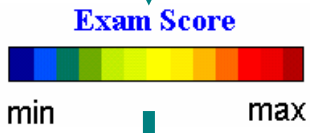


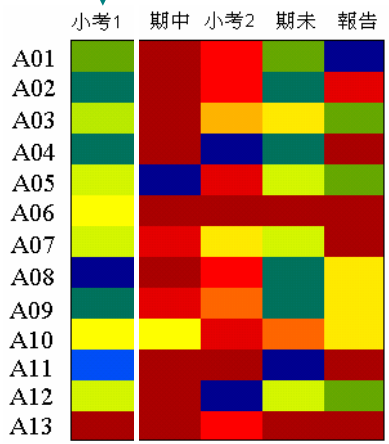
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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





行範囲の条件

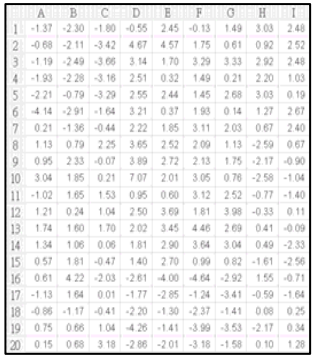


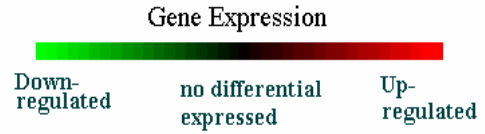


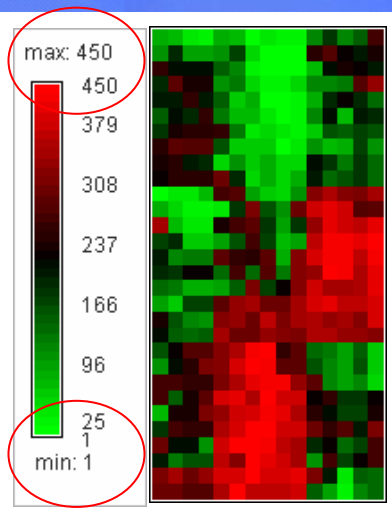
列範囲の条件

これはどうか?

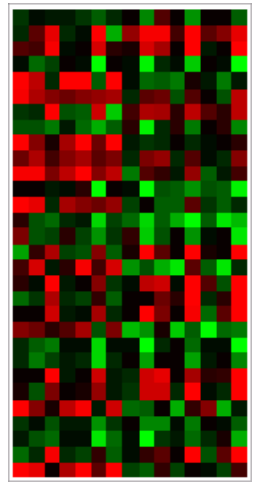
表示の条件



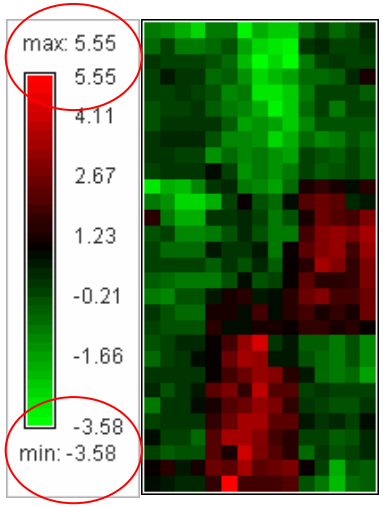




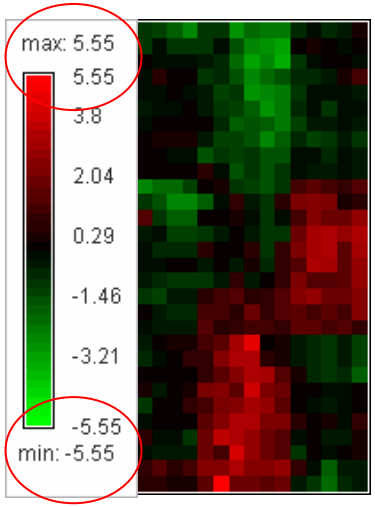
行列ランクの条件



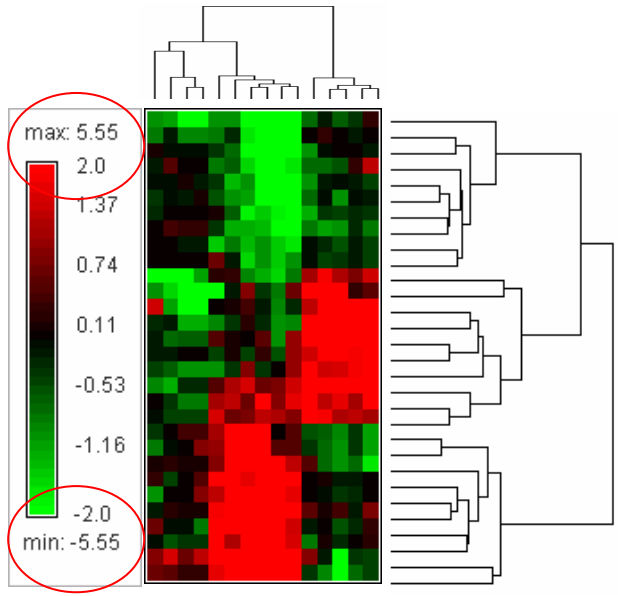
整列しない



行列範囲の条件



中央値の条件

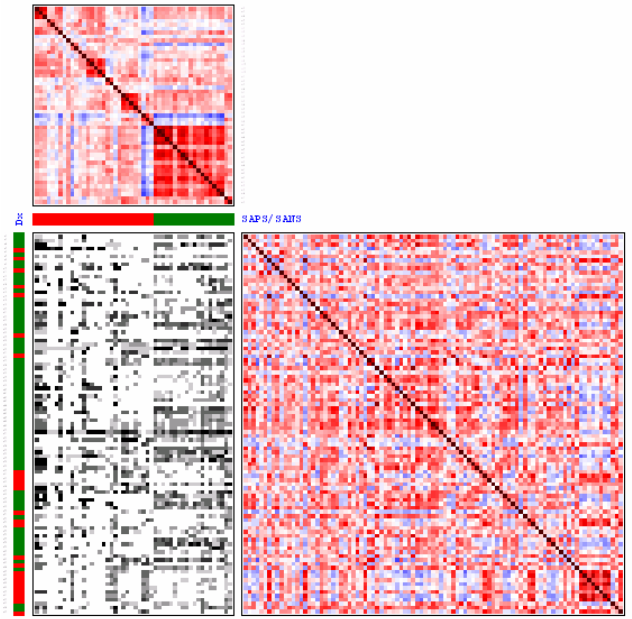


類似度行列と生データ行列の整列

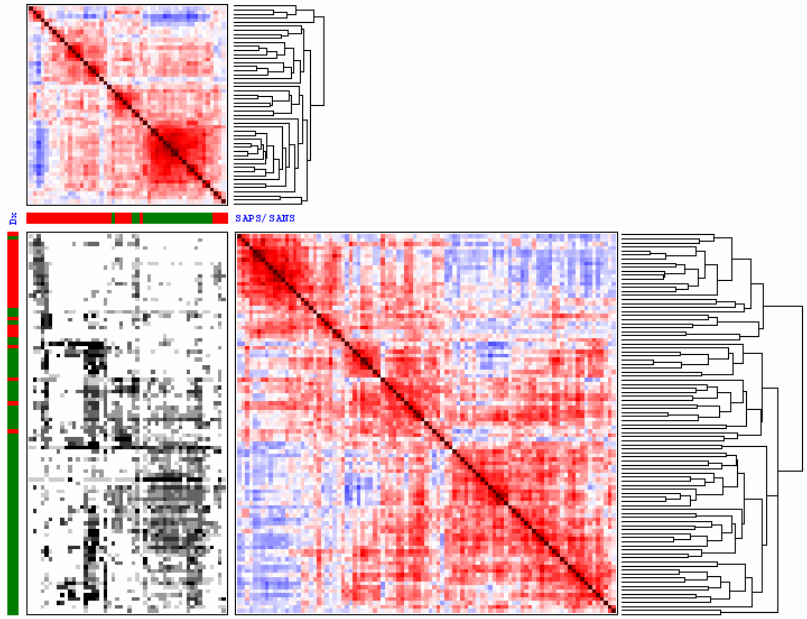
- 統計グラフの関連性
- 大域的評価基準
  - GAPランク2楕円整列
- 局所的評価基準
  - 木による整列
  - 木の中間ノードの反転

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

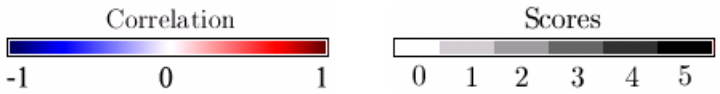


整列法



(1) ランク2楕円整列(チェン、2002)

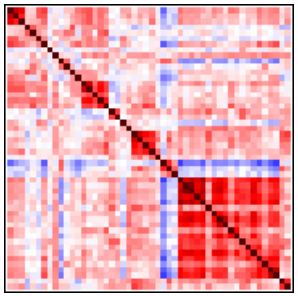
(2) 階層的クラスター木(群平均距離)

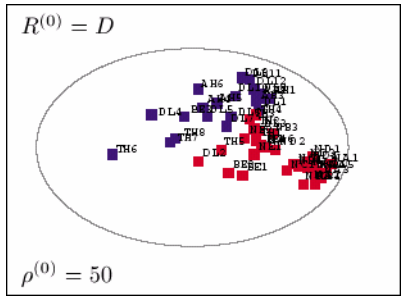


GAPランク2楕円整列

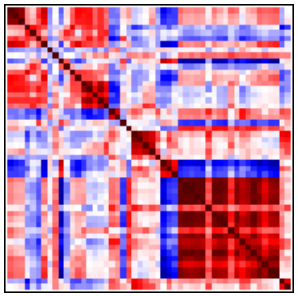
相関行列を収束させる整列アルゴリズム

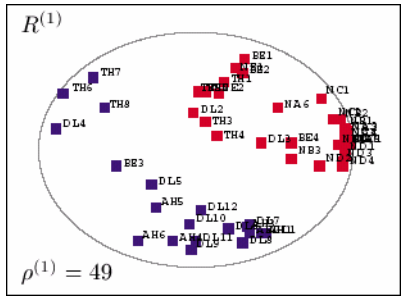
相関行列(整列なし)

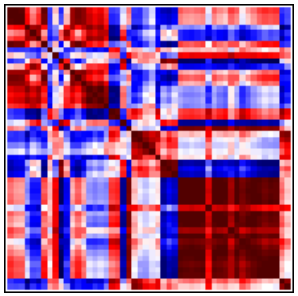


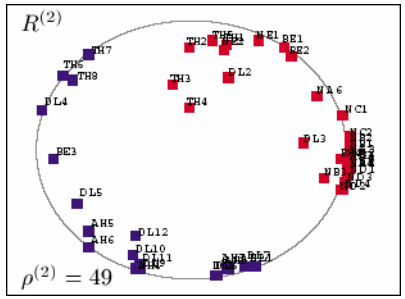


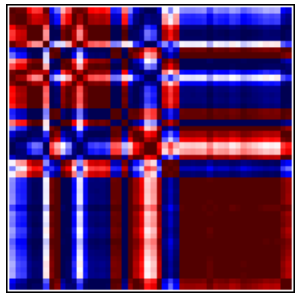
最初の2固有ベクトルへの射影

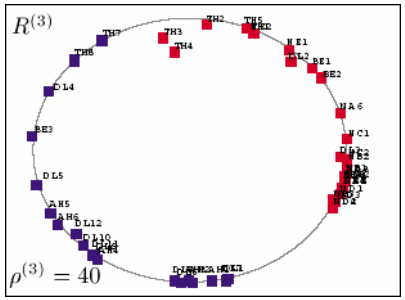


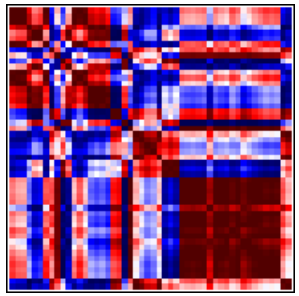


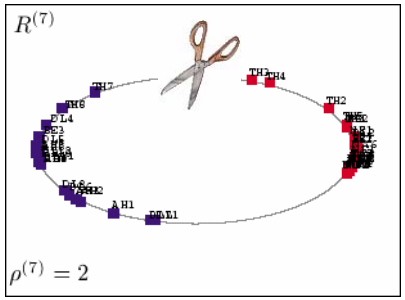




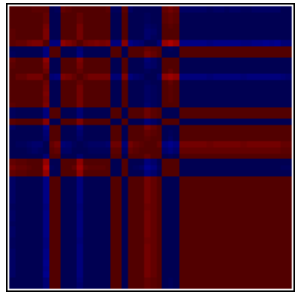


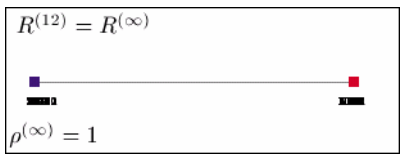


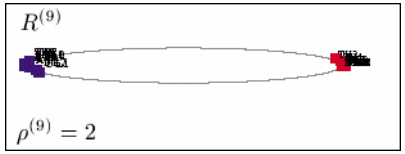


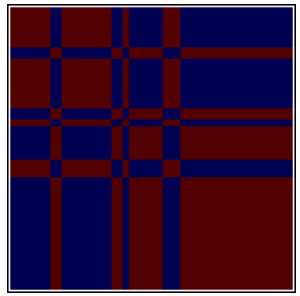


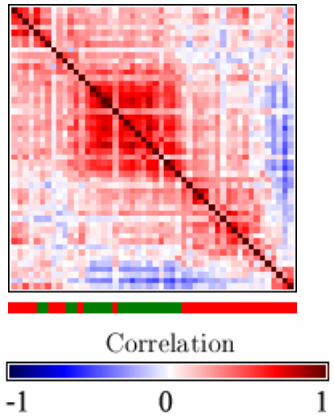
P個の観測値は楕円上に配置され、特定の相対的位置をとる(チェン2002)





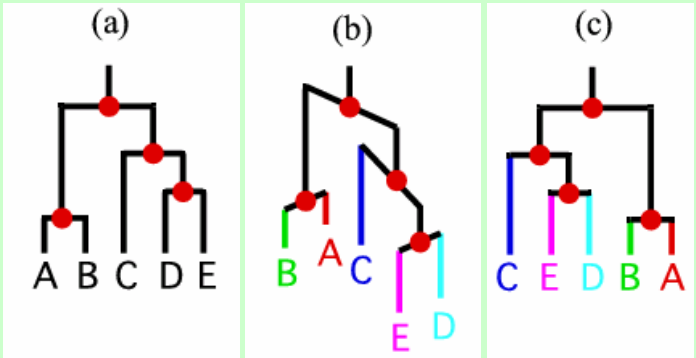




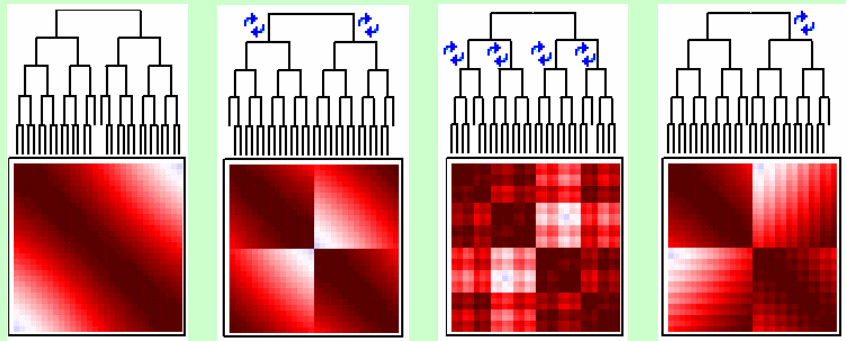


デンドログラムと階層的クラスター木

木整列

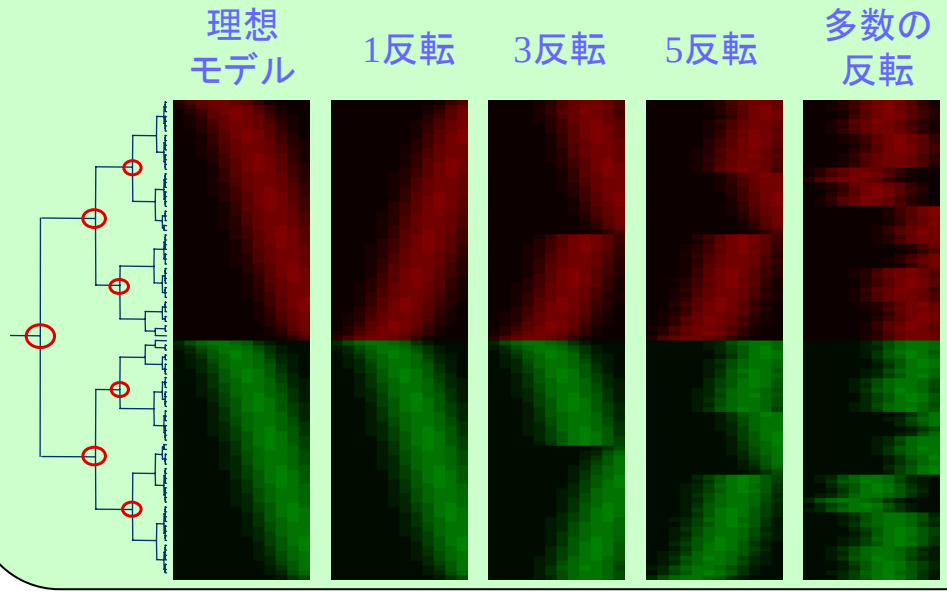


類似度行列の木整列



同一の木構造による異なる整列

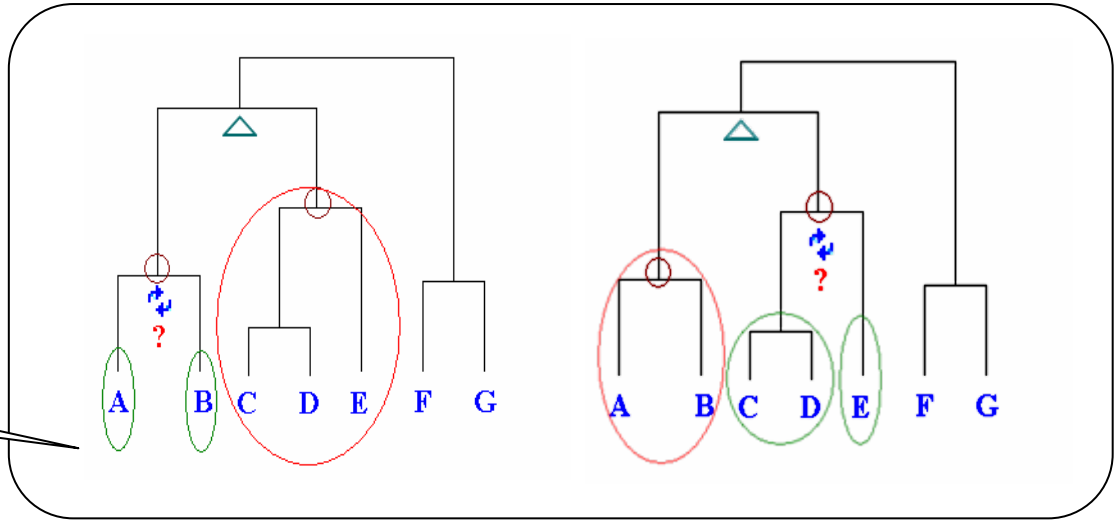
生データ行列の木整列



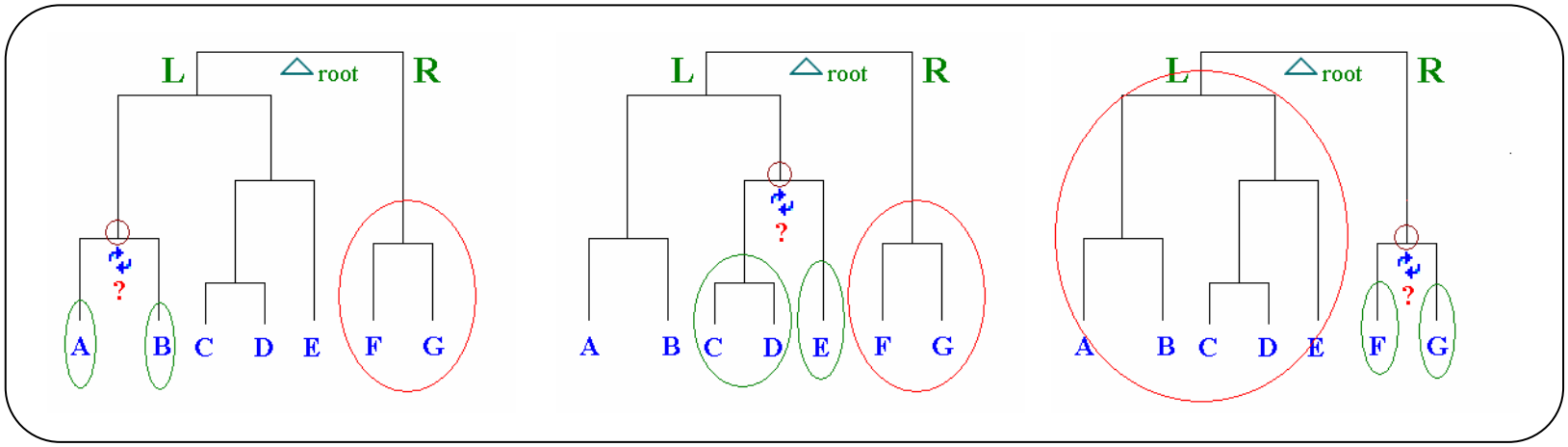
内部木反転

Uncle Approach

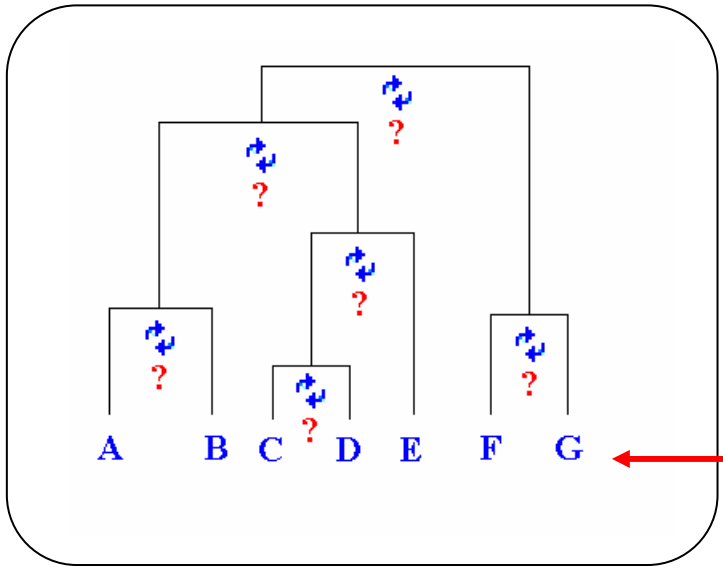
if d(A, {C,D, E}) < d(B, {C,D, E})
then flip



GrandPa Approach



外部木反転



できるだけ近い

External Ordering D E A F B C G

外部整列の方法

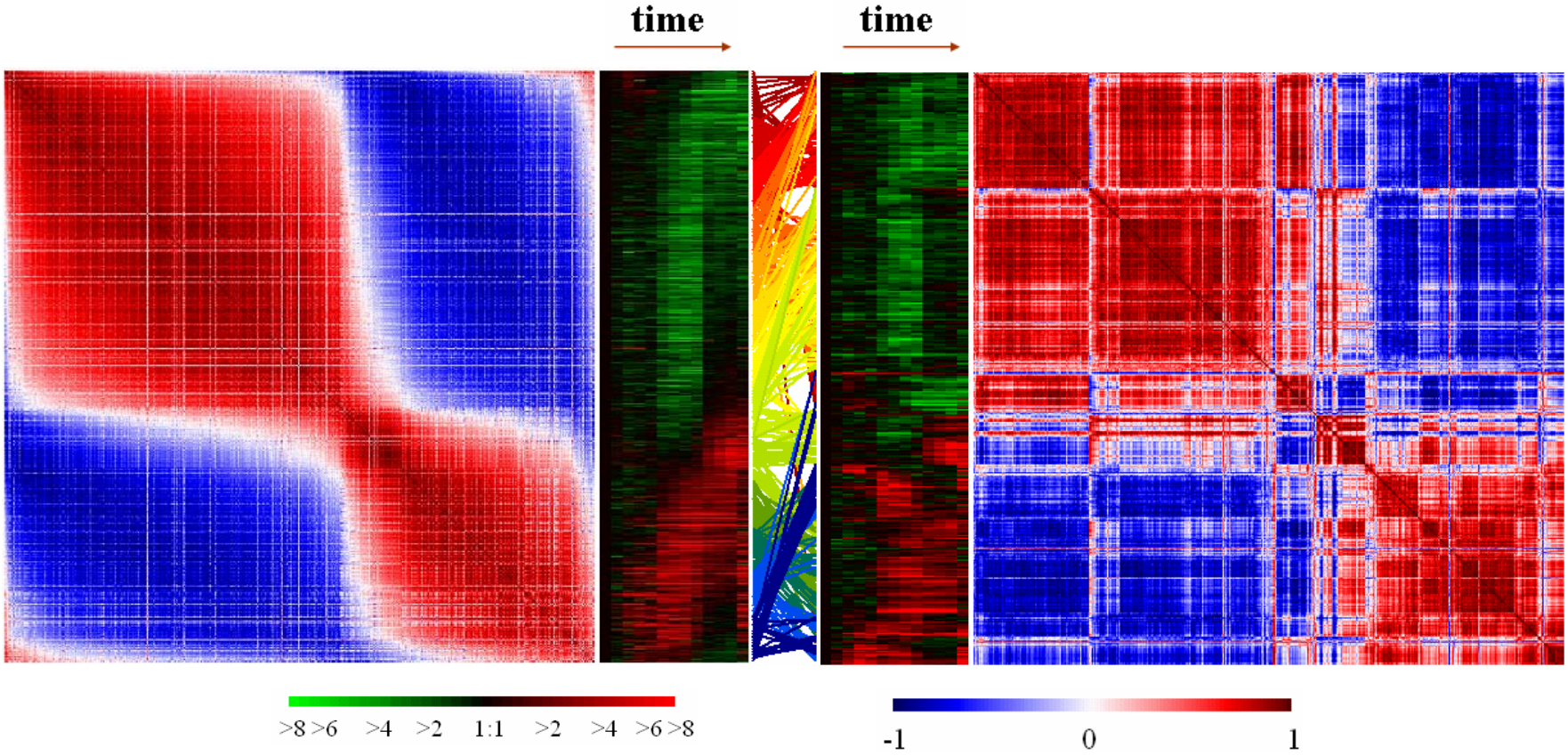
(1) 平均表現レベルによる(Cluster Software, Eisen et al 1998)
(2) 一次元SOMの結果を使用する
(3) …

大域的 vs. 局所的整列

データ: 517遺伝子、13アレイ

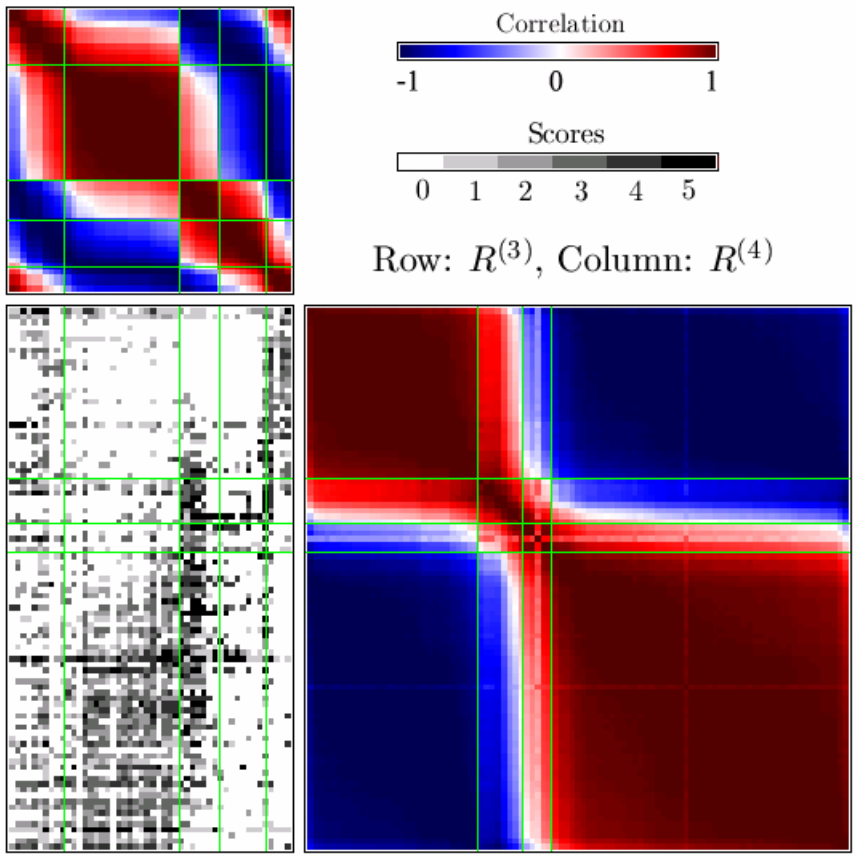
GAPランク2楕円整列

Michael Eisen(1998)木整列

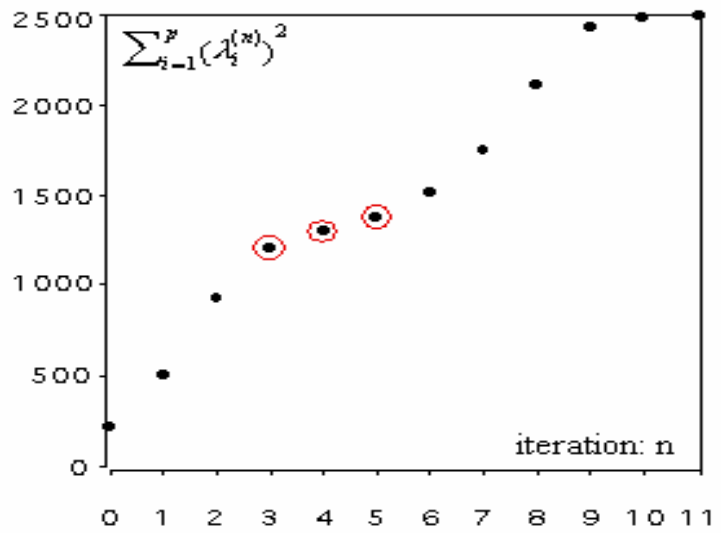


イメージ・ソース: チェン博士のスライド

並べ替えられた行列地図の分割

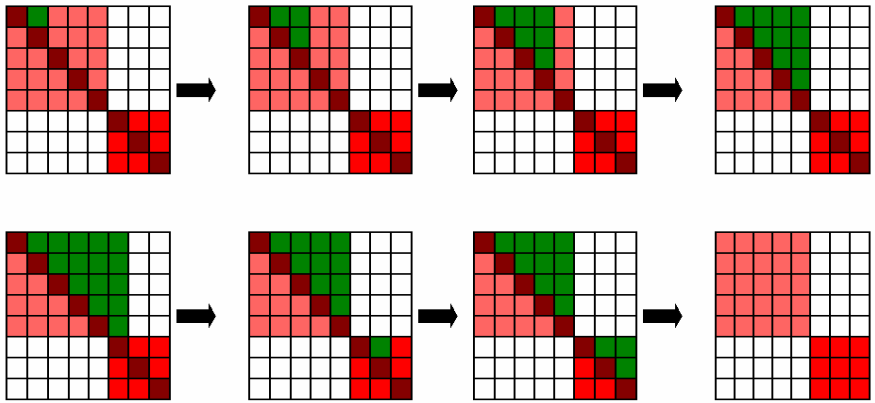


固有値の二乗の合計(相関の二乗の合計)

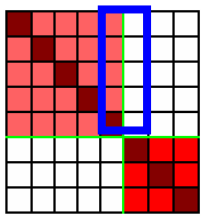


並べ替えられた行列地図の分割

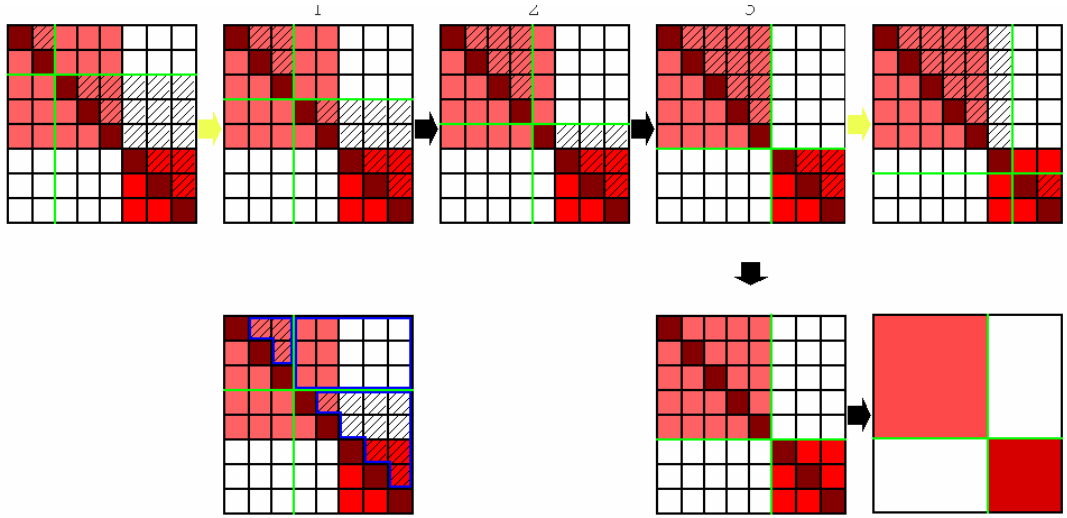
1方向ブロック探索



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

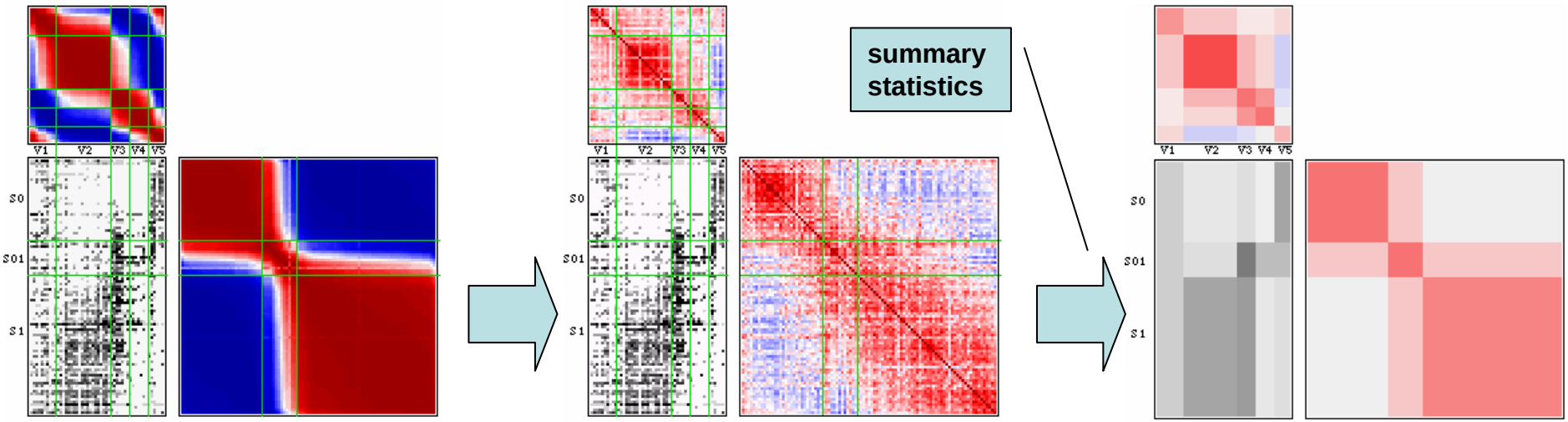
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

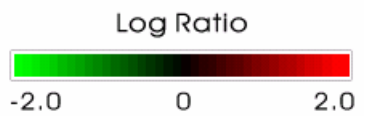
十分統計量

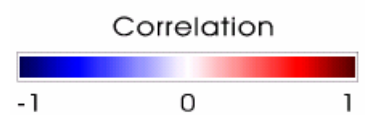
	小考1	期中考	小考2	期末考	報告
平均	71.77	86.54	80.38	53.46	83

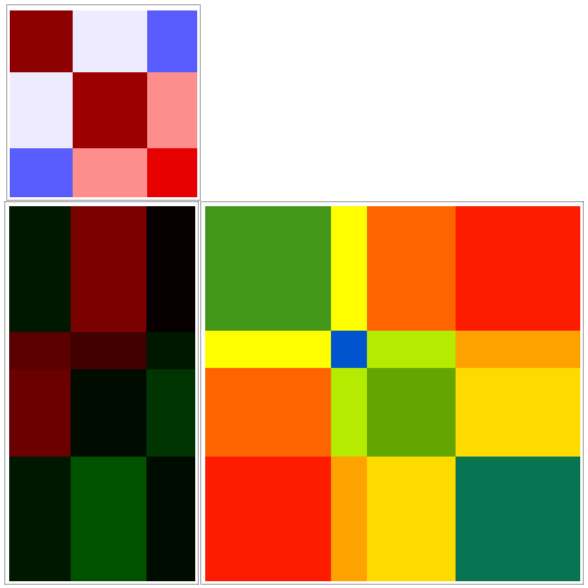
70	低平均	65.67	81.83	73.67	53.67	72
	高平均	77.83	90.67	86.67	53.67	94.17

十分グラフ







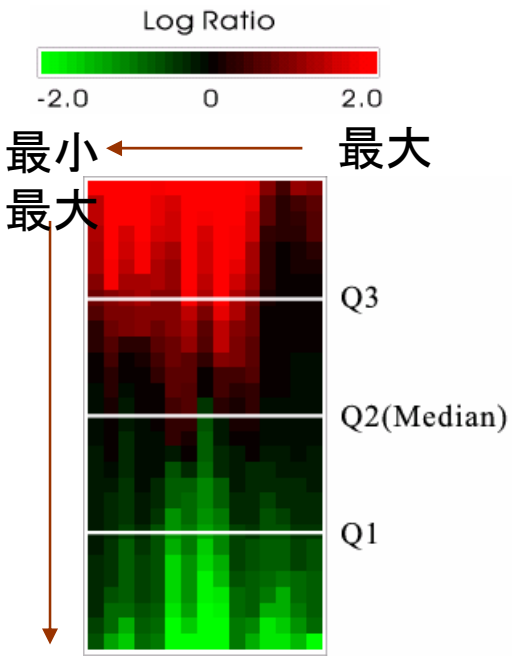


要約表示

(1) 観測値-観測値
(2) 変数-変数
(3) 観測値-変数

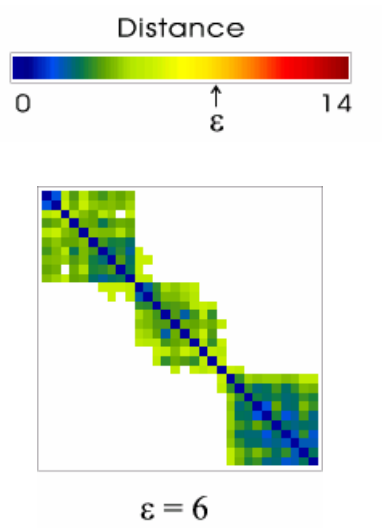
一般化と柔軟性

沈殿表示



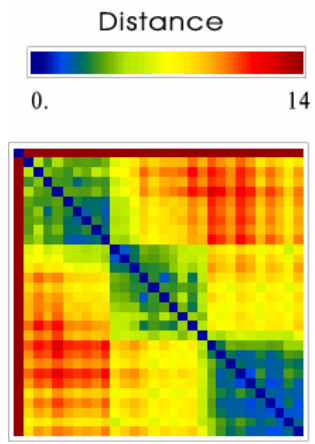
四分位数が計算されたボックスプロットによって与えられたものと同様の情報

部分表示

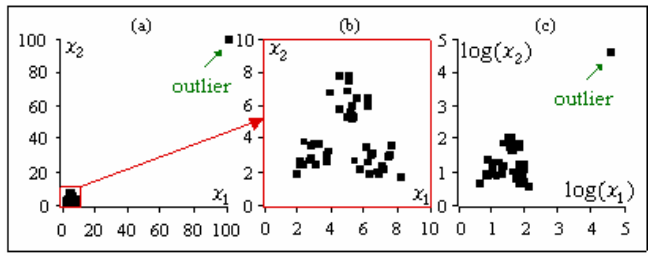


ある条件を満たす数値だけを表示

制限表示

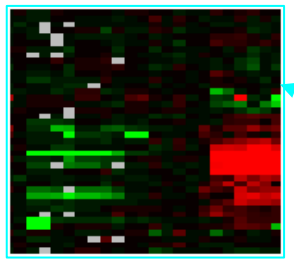


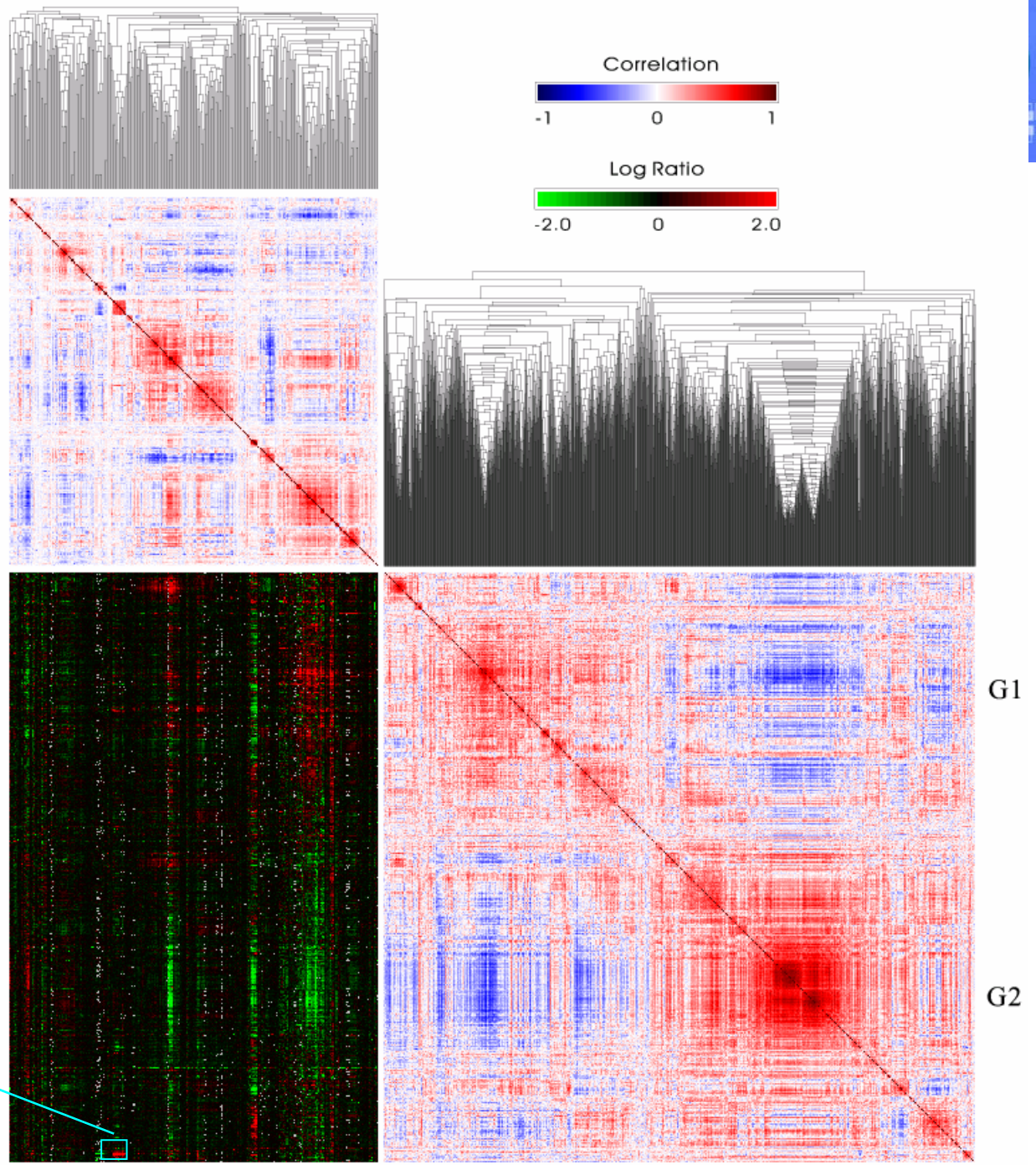
統計グラフの分解能



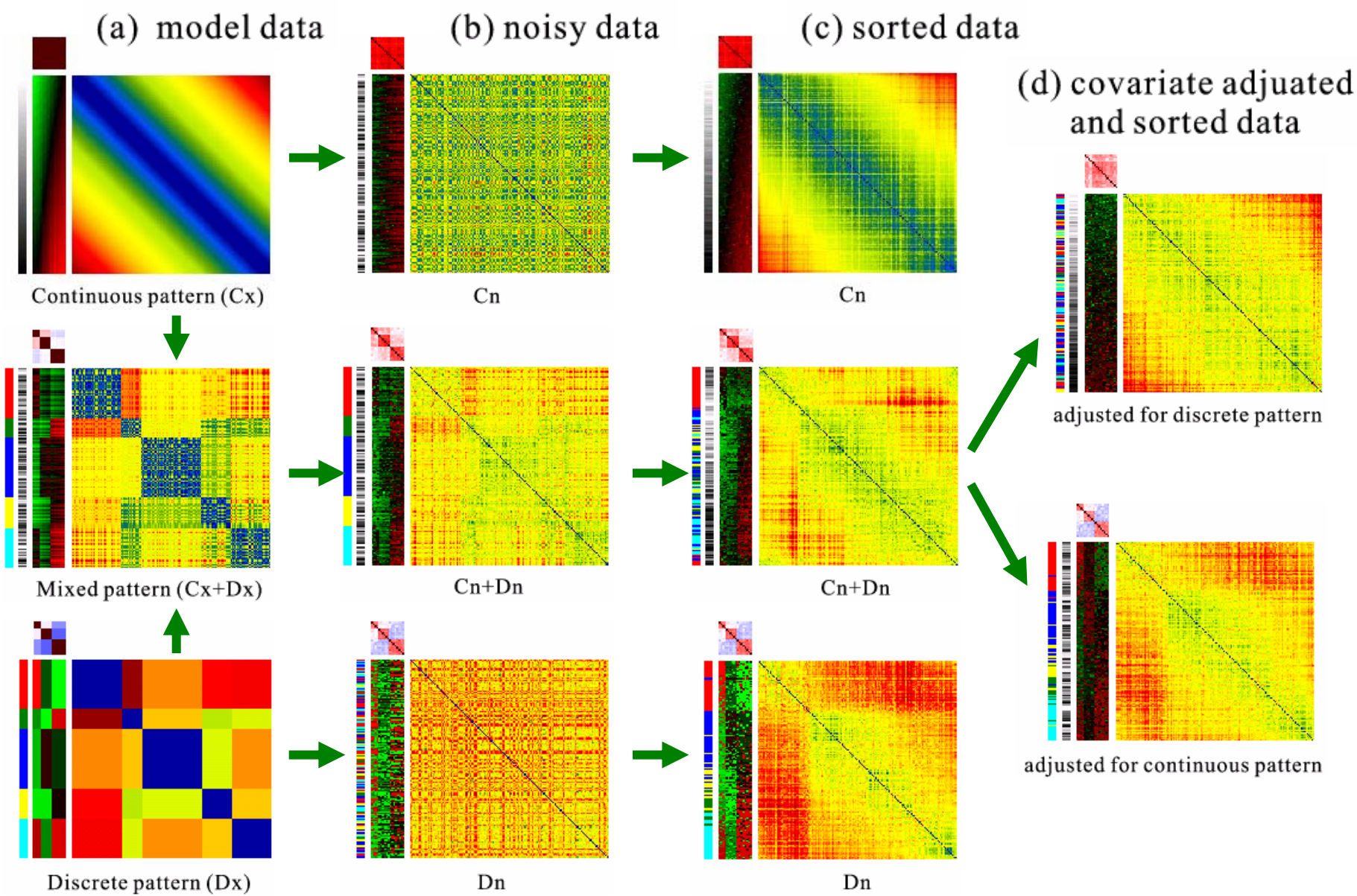
例

- 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



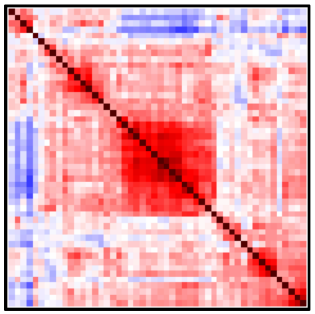


モジュール: 共変量調整のためのMV



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

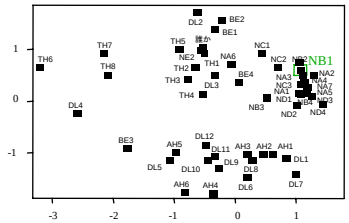
(1) 類似度行列を入力



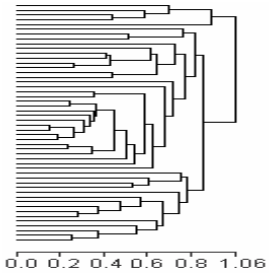
(例: ピアソンの相関係数)

統計的モデリング

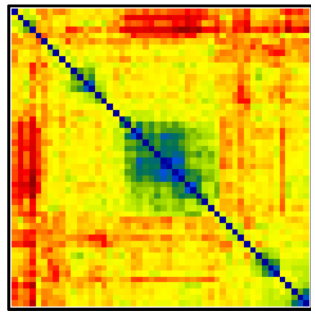
多次元的尺度構成法(MDS)

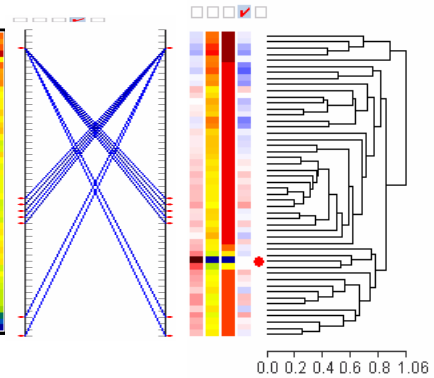


階層的クラスタリング木(HCT)



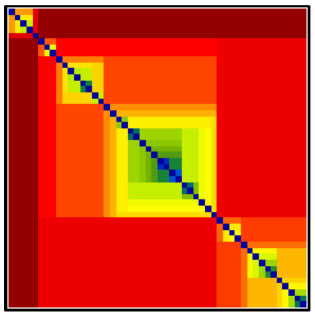
(2) 変換された非類似度行列





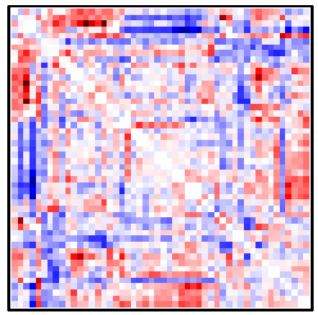
(例： 距離)

(例: Cophenetic行列)

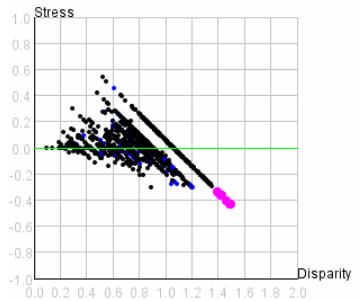


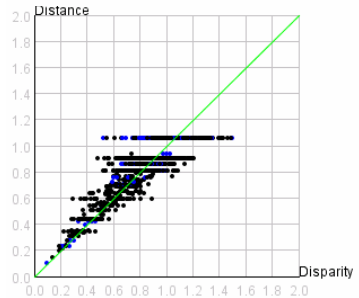
(3) 距離行列を出力

(例えば、残差行列)



(4) ストレス行列





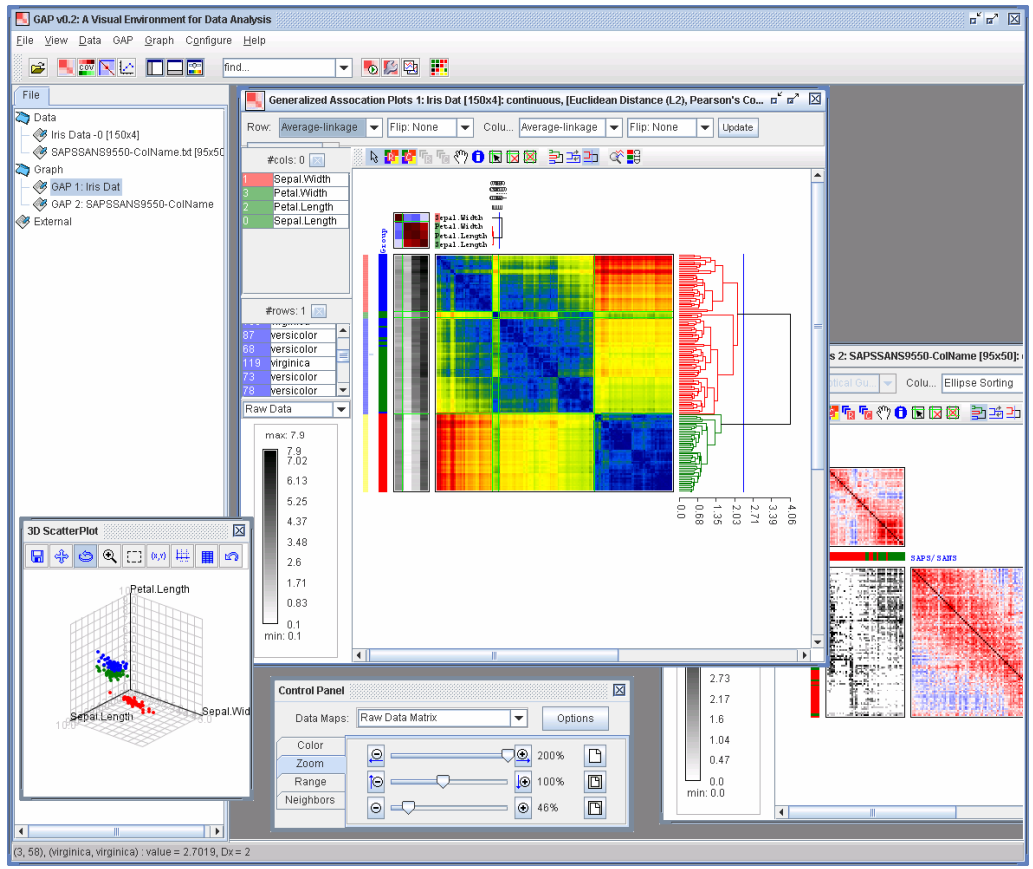
GAP Software verison0.2

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



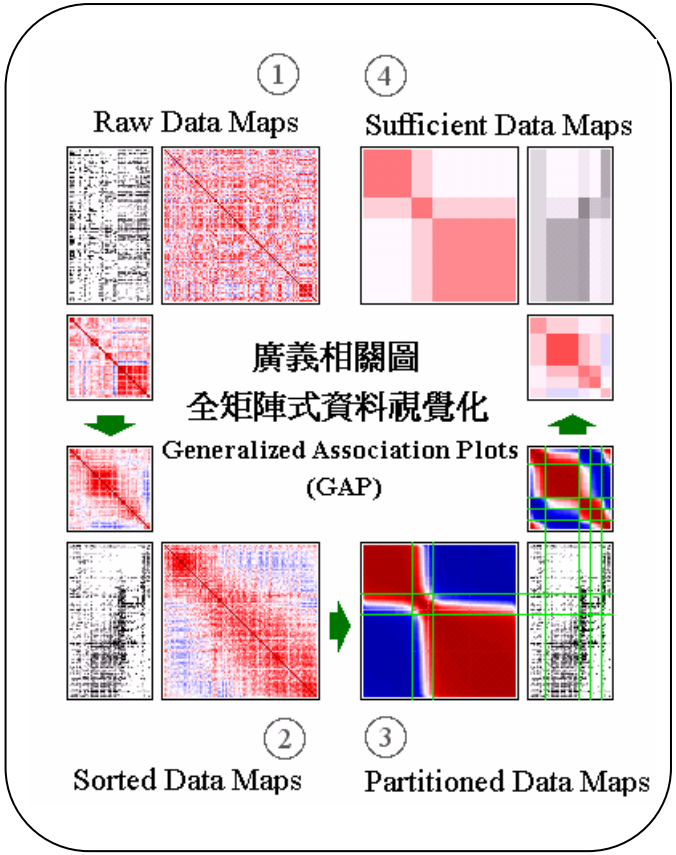
http://gap.stat.sinica.edu.tw/Software/GAP

結論

MVはデータ行列の色による整列表現である

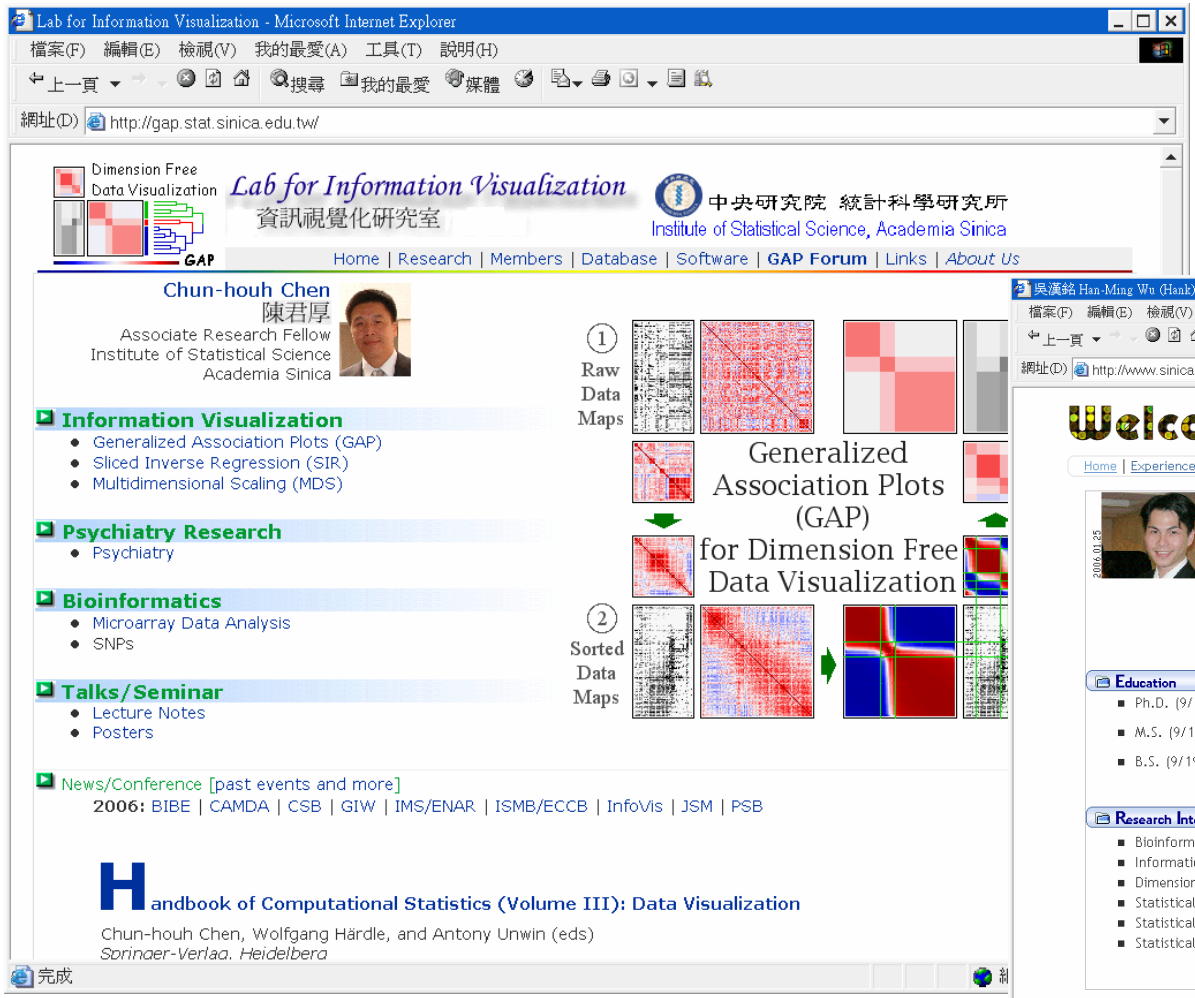
MV表示は5つのレベルの情報を提供する:

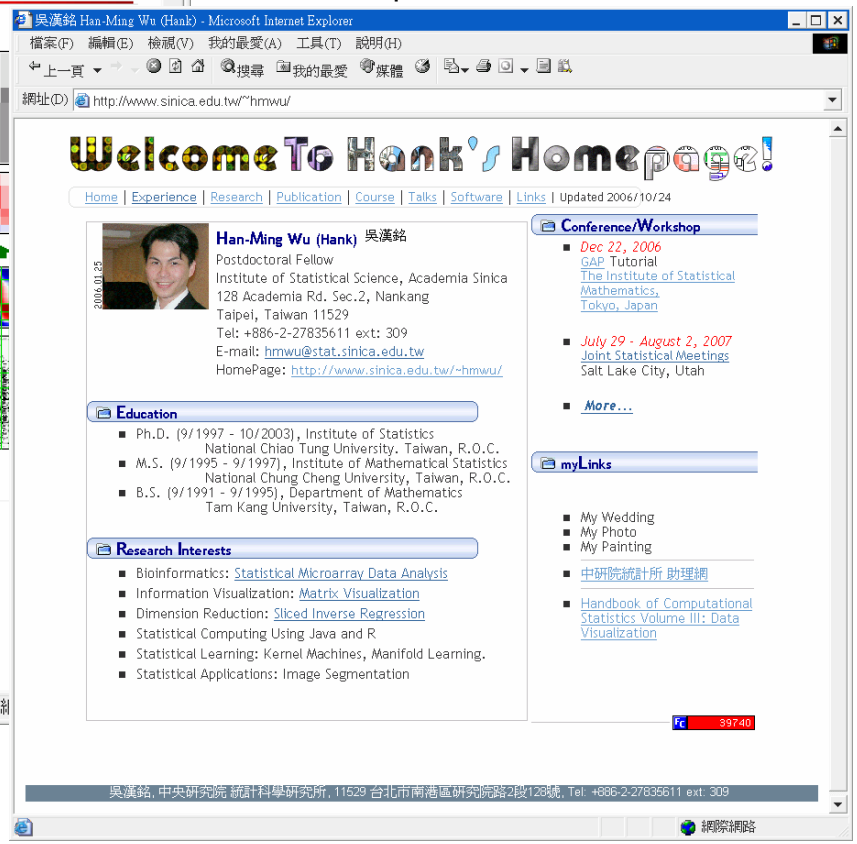
1. あらゆるサンプル/変数の組み合わせの生スコア
2. すべての変数に対する個々のサンプルのスコアベクトル、およびすべてのサンプルに対する個々の変数のベクトル
3. あらゆるサンプル-サンプルと変数-変数の関係のための相関スコア
4. 変数の組分け構造とサンプルのクラスタリング効果
5. 変数グループにおけるサンプル-クラスタの相互パターン



- 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





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

情報可視化研究室

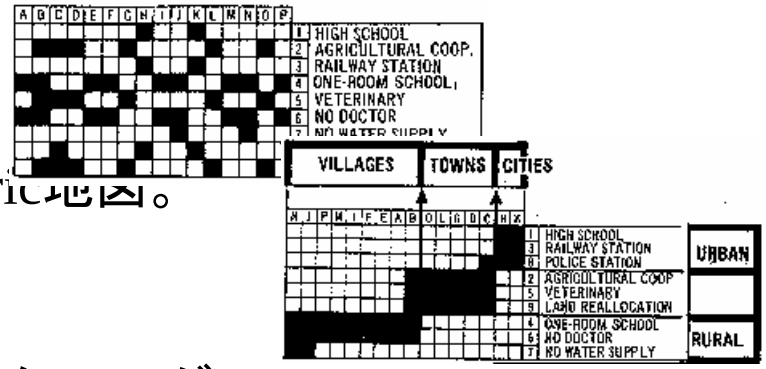
cchen@stat.sinica.edu.tw
http://gap.stat.sinica.edu.tw

付録

MVの関連

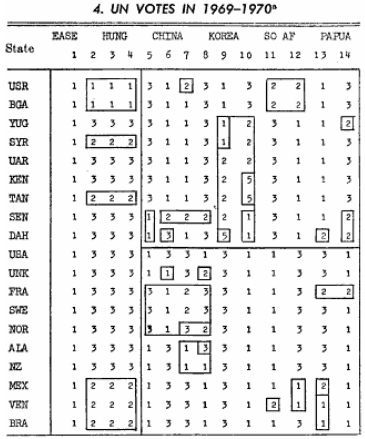
概念:

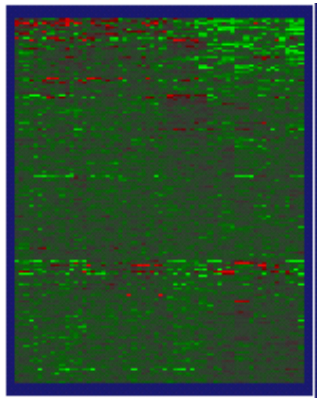
- ベルタン(1967): 整列可能行列
- カーマイケルとスニース(1969): taxometric地図。



データアレイのクラスタリング:

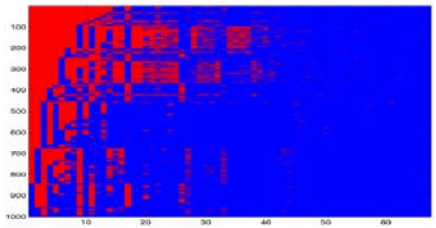
- Hartigan(1972): データ行列の直接クラスタリング
- Tibshirani(1999): ブロッククラスタリング
- Lenstra(1974): 巡回セールスマン問題
- スレイグル他 (1975): 最短経路

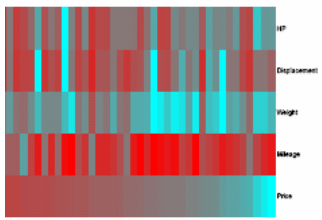


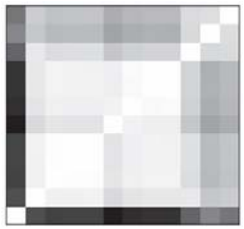


色表現:

- 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



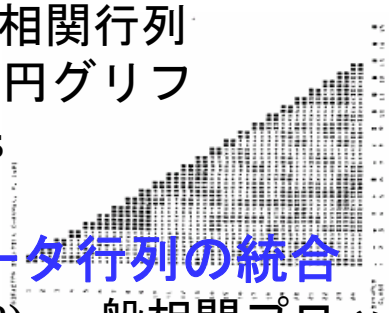


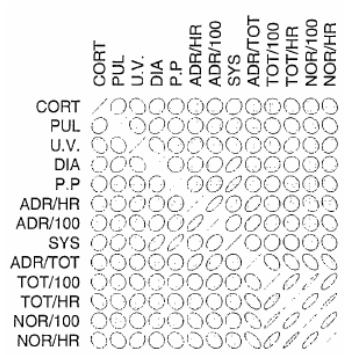


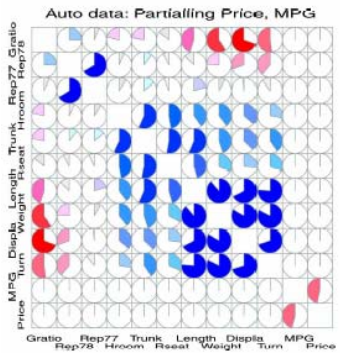
関連(続き)

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

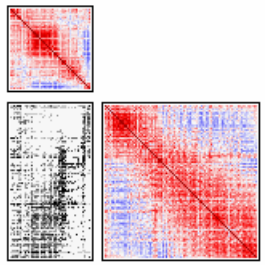






2つの類似度行列による生データ行列の統合

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

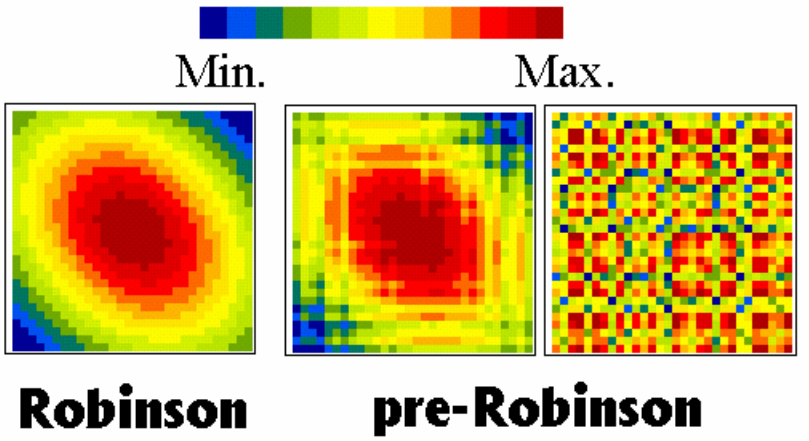


変数とサンプルの整列

- 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



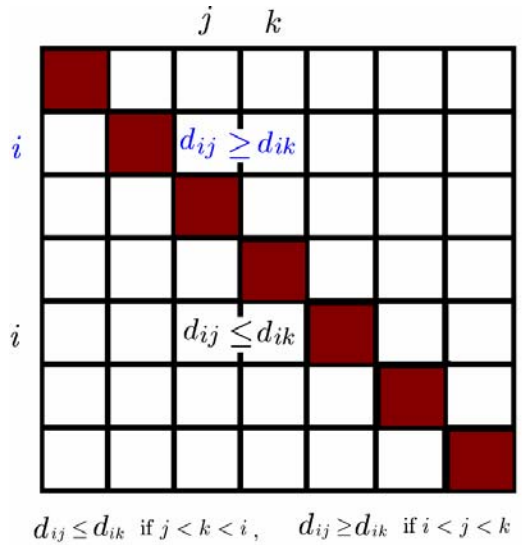
グローバルな評価基準: 非ロビンソン測度

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

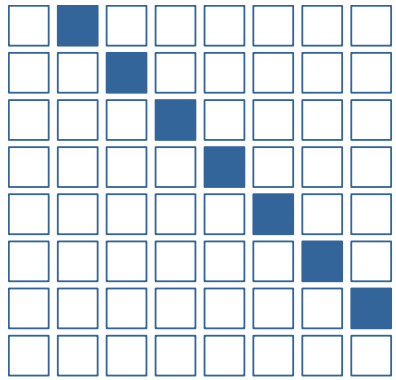
$$AR(s) = \sum_{i=1}^{p} \Big[\sum_{j<k<i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}| \Big],$$

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

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



局所的評価基準: 最小長さ損失関数



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

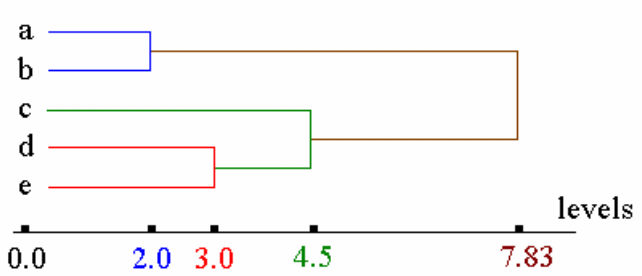
階層的なクラスタリング木

(クープマンとRousseeuw、1990)

例: 平均リンケージ

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

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

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

UPGMA(重みのない対グループ法平均)

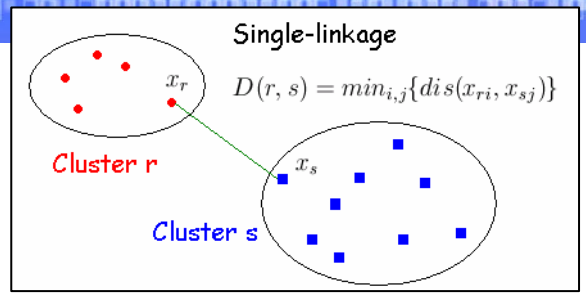
	{a, b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

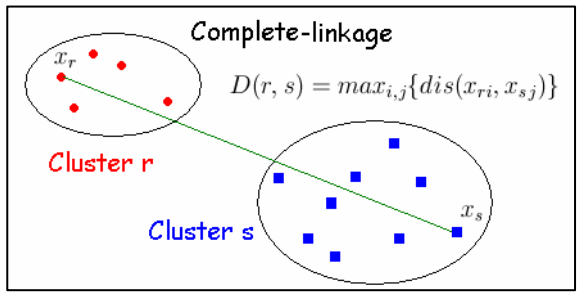
$$D(\{a,b\},\{d,e\})$$

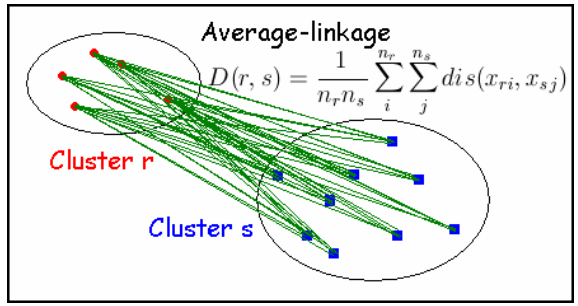
$$= \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$

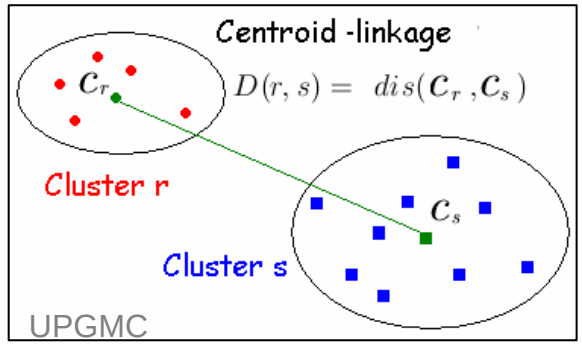
$$= \frac{1}{4}(10+9+9+8) = 9$$

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









Reference

- 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.