Microarray Data Analysis

Visualization, Clustering and Classification

國立台灣大學資訊所

Course: 生物資訊與計算分子生物學

2006/11/07

吳漢銘

hmwu@stat.sinica.edu.tw

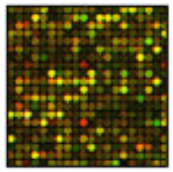
http://www.sinica.edu.tw/~hmwu

中央研究院 統計科學研究所

Institute of Statistical Science, Academia Sinica

Outlines

- **Exploratory Visualization Methods**
  - Principal Components Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - Dendrogram and HeatMap (Matrix Visualization)



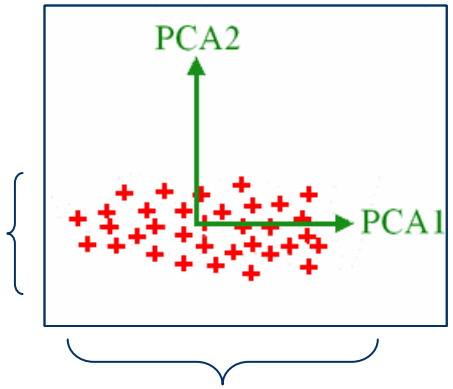
- **Analysis of Relationship Between Genes, Tissues or Treatments**
  - Hierarchical Clustering, K-Means Clustering
  - Self-Organizing Maps (SOM)
  - How Many Clusters?

- **Classification of Genes, Tissues or Samples**
  - Linear Discriminant Analysis (LDA)
  - Support Vector Machines (SVM)
- **Software**

Principal Component Analysis

(Pearson 1901; Hotelling 1933; Jolliffe 2002)

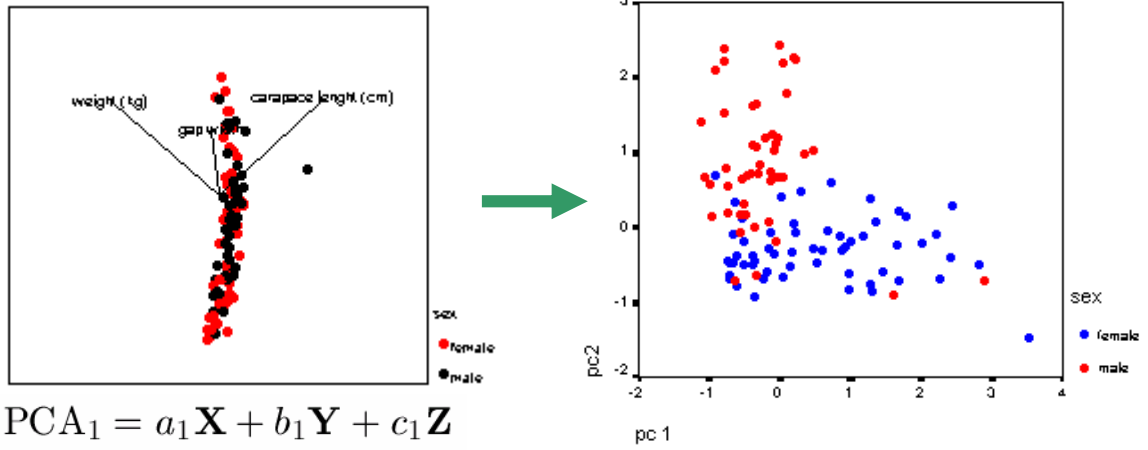
PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y}$$

Image source: 61BL4165 Multivariate Statistics, Department of Biological Sciences, Manchester Metropolitan University



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y} + c_1\mathbf{Z}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$

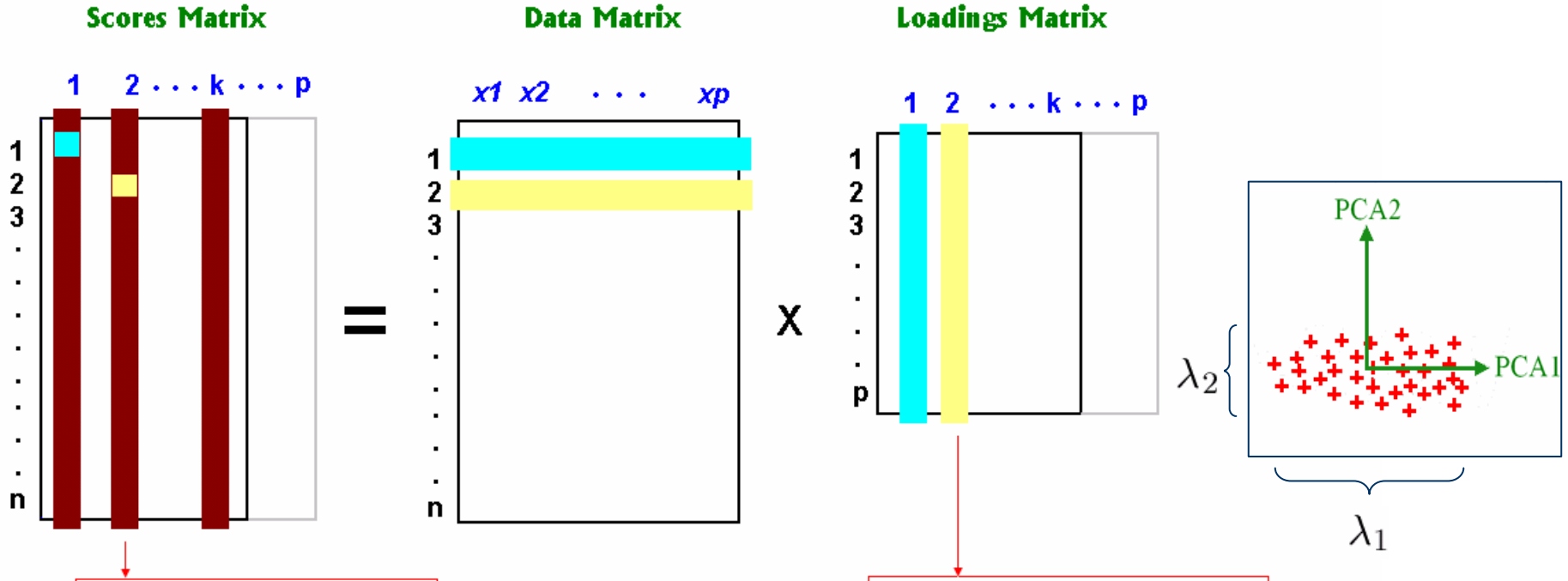
Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

$$\mathrm{PCA}_1 = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p$$

$$\mathrm{PCA}_2 = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p$$

PCA: Loadings and Scores

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



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

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

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

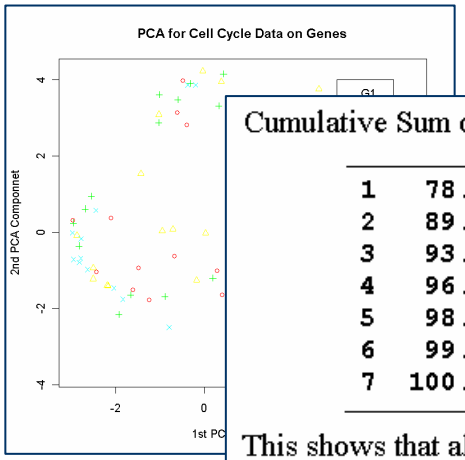
PCA (conti.)

Microarray Data Matrix

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

PCA on Conditions

MA Table	PCA-1	PCA-2	PCA-3
gene001	-0.18	-0.11	-0.03
gene002	0.51	-0.53	0.54
gene003	-0.35	-0.39	0.26
gene004	-0.18	-1.08	0.41
gene005	-0.62	-0.8	0.13
gene006	-0.09	-0.23	0.77
gene007	-0.38	-0.32	1.08
gene008	-0.88	-0.55	1.03
gene009	-1.26	0.45	0.41
gene010	0.12	-0.36	-0.16
gene011	-0.28	-0.44	2.13
gene012	-0.45	-0.23	0.82
gene013	-0.2	-0.43	0.44
gene014	0.03	-0.26	-0.68
gene015	-0.7	-0.76	0.5
gene016	-0.61	0.07	-0.04
gene017	-0.23	-0.71	0.01
gene018	0.1	0.1	0.11
gene019	-0.94	-0.97	0.24
gene020	-0.55	-0.53	0.86
gene021	-0.47	-0.87	-0.02
gene022	-0.34	-1.1	0.51
gene•••	-0.49	-0.2	0.91
gene n	-0.15	-1.04	-0.01

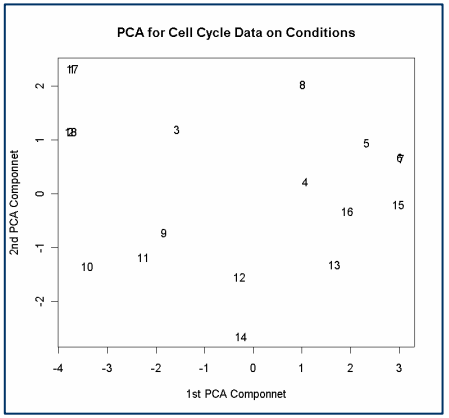


Cumulative Sum of the Variances:

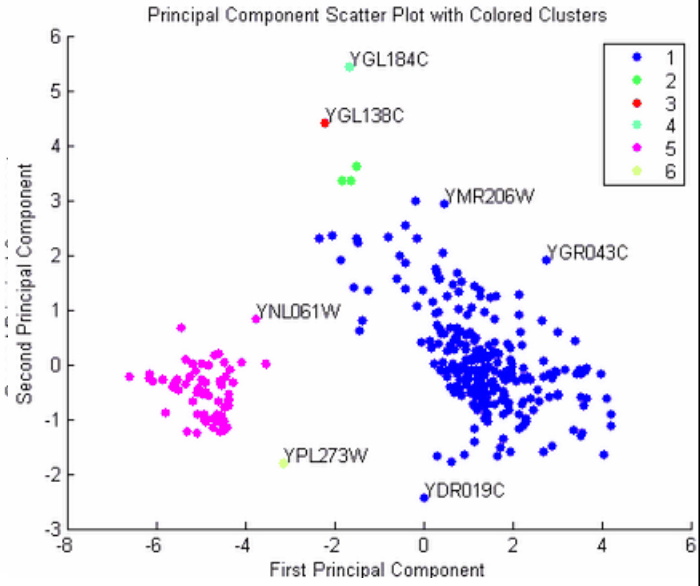
1	78.3719
2	89.2140
3	93.4357
4	96.0831
5	98.3283
6	99.3203
7	100.0000

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

PCA on Genes



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
PCA-1	0.18	0.3	-0.12	-0.44	0.19	-0.39	-0.61
PCA-2	-0.16	-0.58	-0.43	-0.22	0.53	0.69	0.08
PCA-3	0.16	-0.44	-0.93	-1.23	-0.62	0.62	1.3



Yeast Microarray Data is from
DeRisi, JL, Iyer, VR, and Brown, PO.(1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale"; Science, Oct 24;278(5338):680-6.

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)

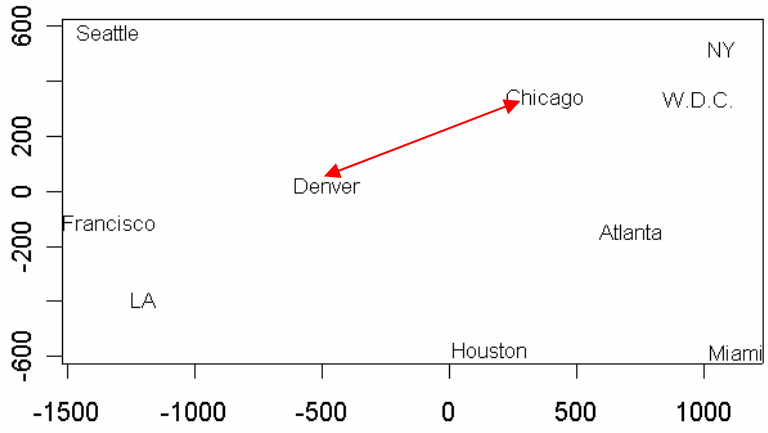
http://www.lib.utexas.edu/maps/united_states.html

- Classical MDS takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.
- projection from some unknown dimensional space to 2-d dimension.

Flying Mileages Between Ten U.S. Cities

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

MDS



MDS: Metric and Non-Metric Scaling

Question

Given a *dissimilarity matrix* D of certain objects, can we construct points in k-dimensional (often 2-dimensional) space such that

Goal of metric scaling

the Euclidean distances between these points approximate the entries in the dissimilarity matrix?

Goal of non-metric scaling

the order in distances coincides with the order in the entries of the dissimilarity matrix approximately?

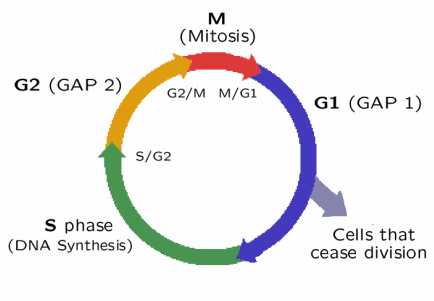
$$S = \sum_{i,j} (\hat{d}_{ij} - d_{ij})^2$$

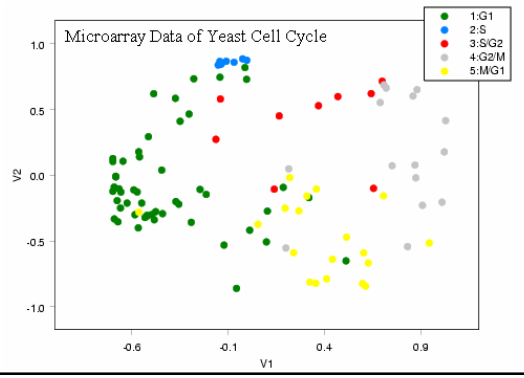
Mathematically: for given k, compute points $x_1,\ldots,x_n$ in k-dimensional space such that the object function is minimized.

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

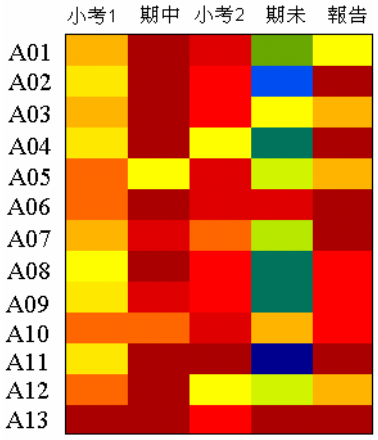
Microarray Data of Yeast Cell Cycle

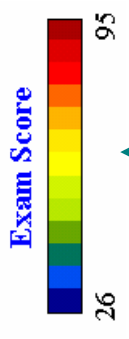
- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points.
- 2D MDS Configuration Plot for 103 known genes.





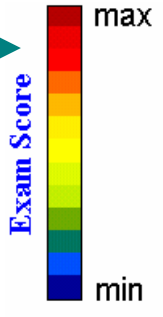
Heat Map
(Data Image, Matrix Visualization)

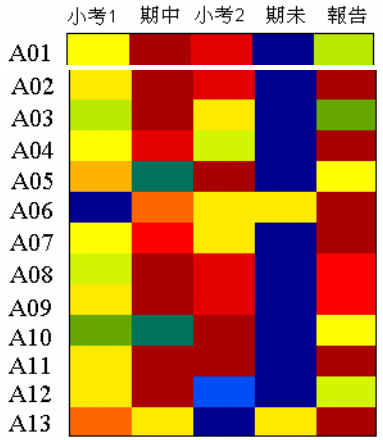




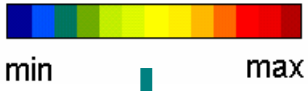
Range Matrix Condition

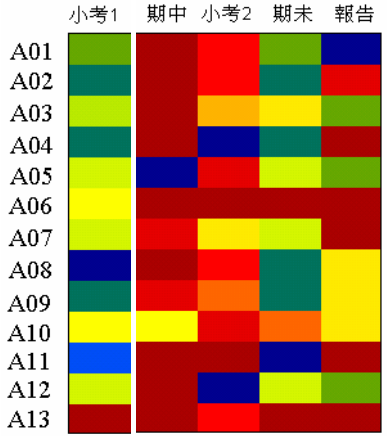
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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





Range Raw Condition



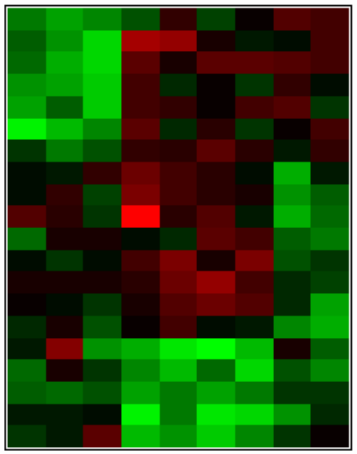


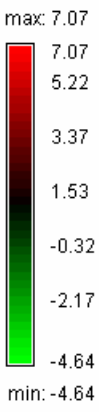
Range Column Condition

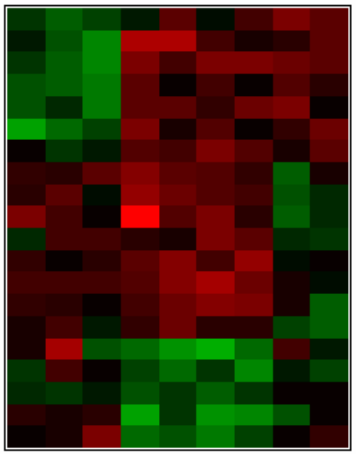
What about this one?

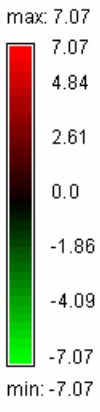
Heat Map (conti.)

	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28







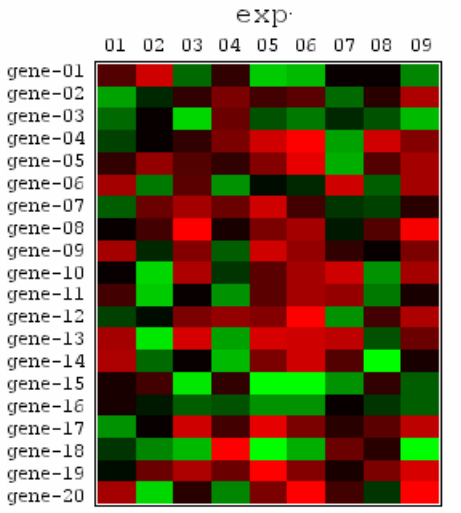


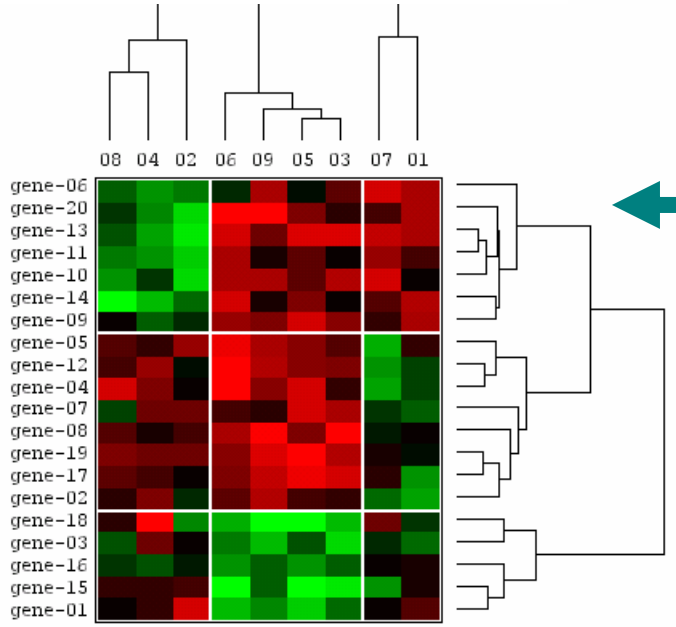
Range Matrix Condition

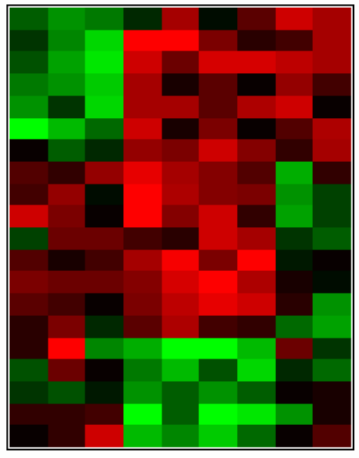
Center Matrix Condition

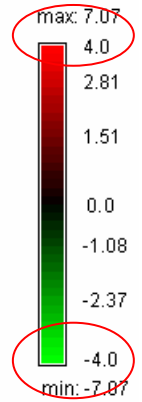
Gene Expression

Down-regulated — no differential expressed — Up-regulated





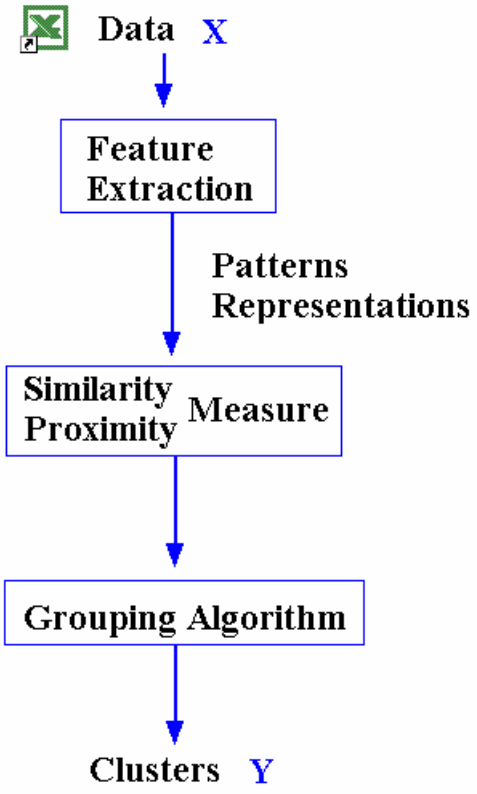




Clustering Analysis (Unsupervised Learning)

What is Clustering?

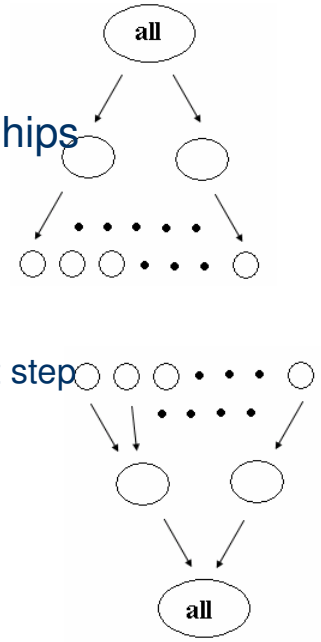
Cluster analysis is the organization of a collection of patterns into clusters based on similarity. The problem is to group a given collection of unlabeled patterns into meaningful clusters.



Hierarchical clustering

The result is a tree that depicts the relationships between the objects.

- **Divisive clustering:**
  begin at step 1 with all the data in one cluster.
- **Agglomerative clustering:**
  all the objects start apart., there are n clusters at step 0.



Non-Hierarchical clustering

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

Two important properties of a clustering definition:

1. Most of data has been organized into **non-overlapping clusters**.
2. Each cluster has a within variance and one between variance for each of the other clusters. A good cluster should have a small within variance and large between variance.

Distance and Similarity Measure

Cov	x1	x2	x3	x4		x p
x1	1.00	0.48	0.10	-0.10		-0.28
x2	0.48	1.00	0.41	0.22		-0.23
x3	0.10	0.41	1.00	0.36		-0.05
x4	-0.10	0.22	0.36	1.00		0.10
x p	-0.28	-0.23	-0.05	0.10		1.00

Proximity Matrix

Data Matrix x y

Data	x1	x2	x3	x4	...	x p
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-1.22	-0.74	1.34	1.50		0.29
subject09	-0.73	-1.06	-0.79	-0.02		0.44
subject10	-0.58	-0.40	0.13	0.58		0.02
subject11	-0.50	-0.42	0.66	1.05		0.06
subject12	-0.86	-0.29	0.42	0.46		0.10
subject13	-0.16	0.29	0.17	-0.28		-0.55
subject14	-0.36	-0.03	-0.03	-0.08		-0.25
subject15	-0.72	-0.85	0.54	1.04		0.24
subject16	-0.78	-0.52	0.26	0.20		0.48
subject17	0.60	-0.55	0.41	0.45		-0.66
⋮						
subject n	-2.29	-0.64	0.77	1.60		0.55
mean	0.07	-0.04	0.44	0.31	...	-0.21

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

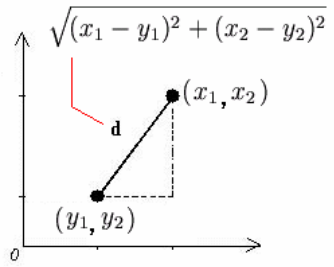
$$y = (y_1, y_2, \cdots, y_n)$$

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

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



- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

More Similarity Measures

Dissimilarity/Similarity Measure for Quantitative Data

Kendall's tau

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

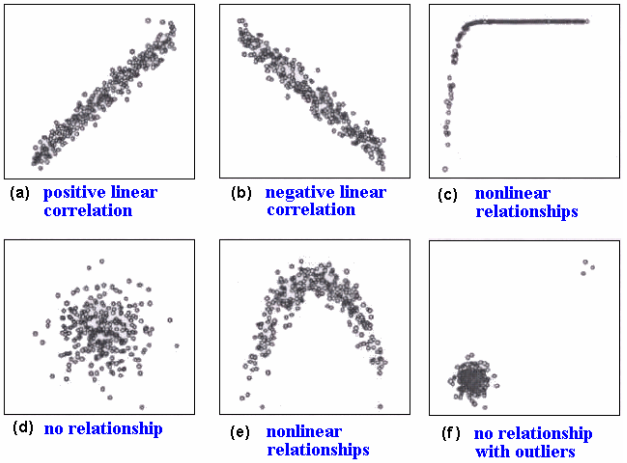
- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

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

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

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

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



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

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

- Pearson's rho measures the strength of a linear relationship [(a), (b)].
- Spearman's rho and Kendall's tau measure any monotonic relationship between two variables [(a), (b) ,(c)].
- If the relationship between the two variables is non-monotonic, all three correlation coefficients fail to detect the existence of a relationship [(e)].
- Both Spearman's rho and Kendall's tau are rank-based non-parametric measures of association between variable X and Y.
- The rank-based correlation coefficients are more robust against outliers.

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

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

Hierarchical Clustering and Dendrogram

(Kaufman and Rousseeuw, 1990)

Example:

Average-Linkage

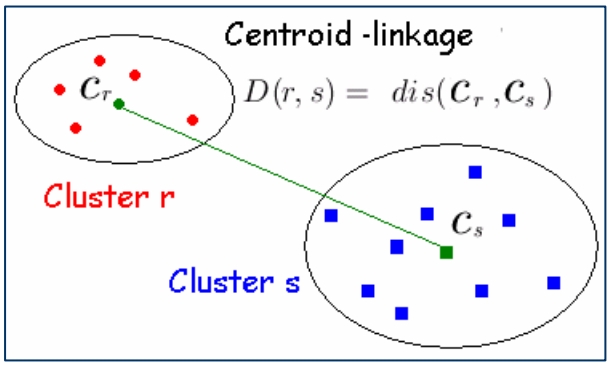
	a	b	c	d	e
a	0	2	6	10	9
b		0	5	9	8
c			0	4	5
d				0	3
e					0

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

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

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

UPGMC (Unweighted Pair-Groups Method Centroid)



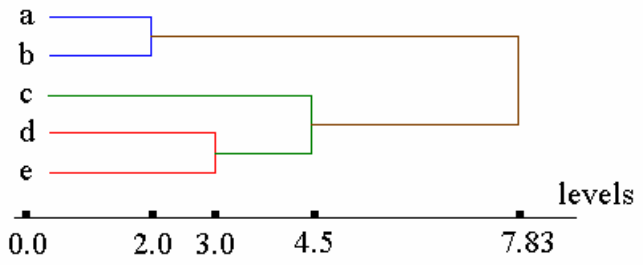
$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c) + D(b,c)]$$

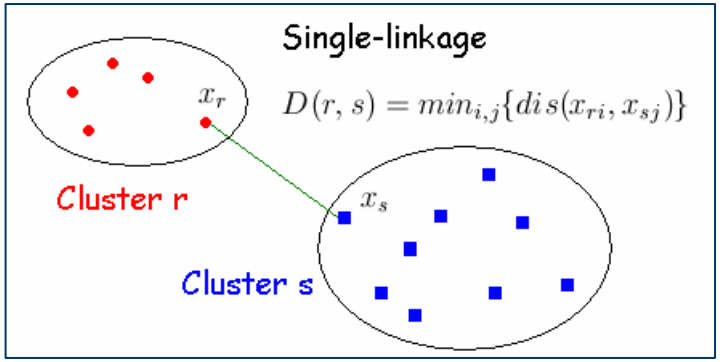
$$= \frac{1}{2}(6+5) = 5.5$$

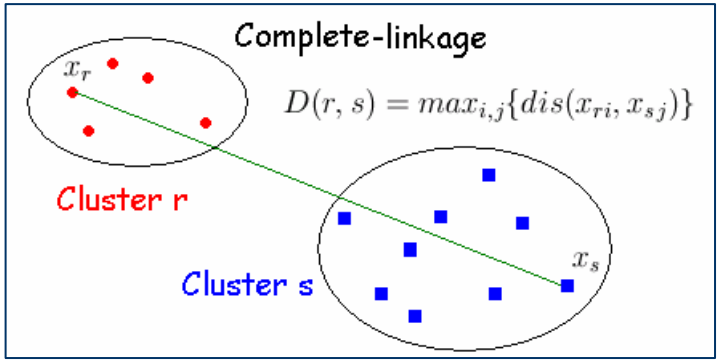
$$D(\{a,b\},\{d,e\})$$

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

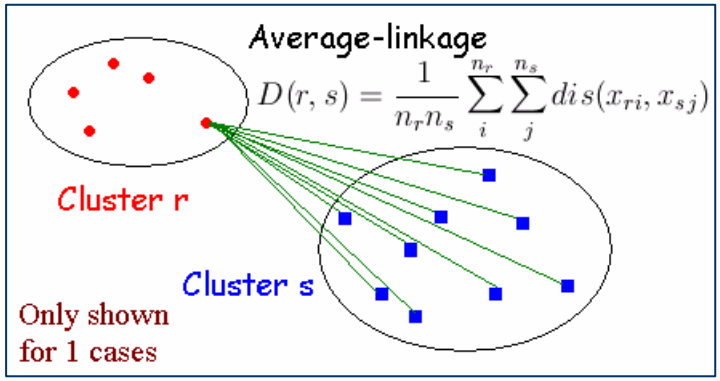
$$= \frac{1}{4}(10+9+9+8) = 9$$



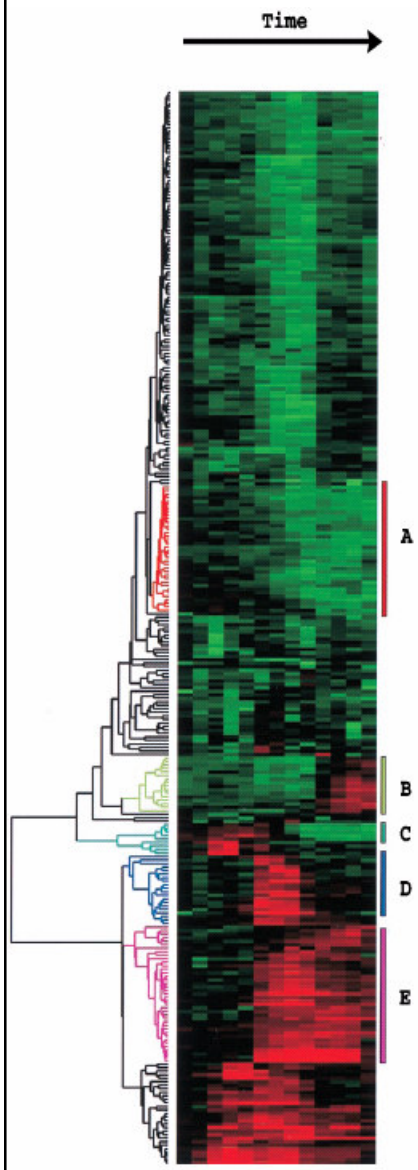




UPGMA (Unweighted Pair-Groups Method Average)



Hierarchical Clustering



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

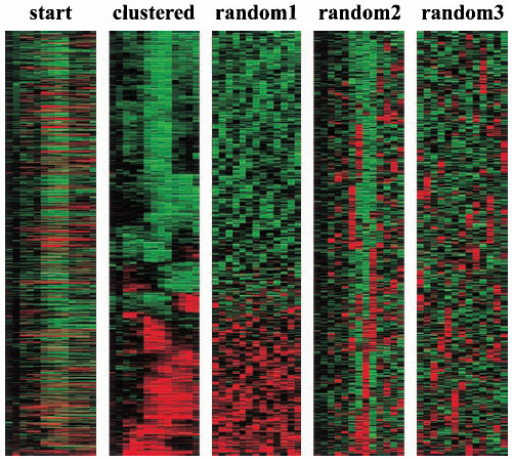
Cluster analysis and display of genome-wide expression patterns

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

FIG. 1. Clustered display of data from time course of serum stimulation of primary human fibroblasts. Experimental details are described elsewhere (11). Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8,600 distinct human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated green for log ratios −3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. As described in detail in ref. 11, the sequence-verified named genes in these clusters contain multiple genes involved in (*A*) cholesterol biosynthesis, (*B*) the cell cycle, (*C*) the immediate–early response, (*D*) signaling and angiogenesis, and (*E*) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes. A larger version of this image, with gene names, is available at http://rana.stanford.edu/clustering/serum.html.

Software:
Cluster and TreeView

FIG. 3. To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).



K-Means Clustering

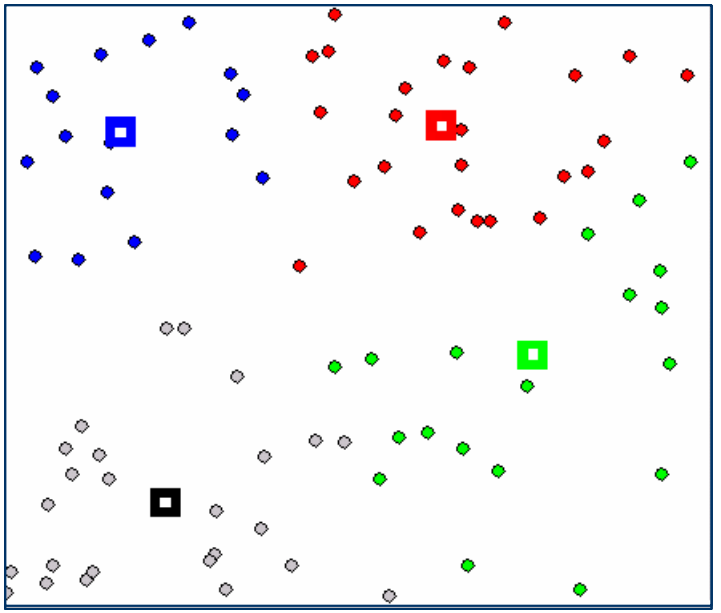
- K-means is a partition methods for clustering.
- Data are classified into k groups as specified by the user.
- Two different clusters cannot have any objects in common, and the k groups together constitute the full data set.

Optimization problem:

Minimize the sum of squared within-cluster distances

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

Converged



The K-Means Algorithm

1. The data points are randomly assigned to one of the K clusters.
2. The position of the K centroids are determined (initial group centroids).
3. For each data point:
   - Calculate the distance from the data point to each cluster.
   - Assign data point to the cluster that has the closest centroid.
4. Repeat the above step until the centroids no longer move.

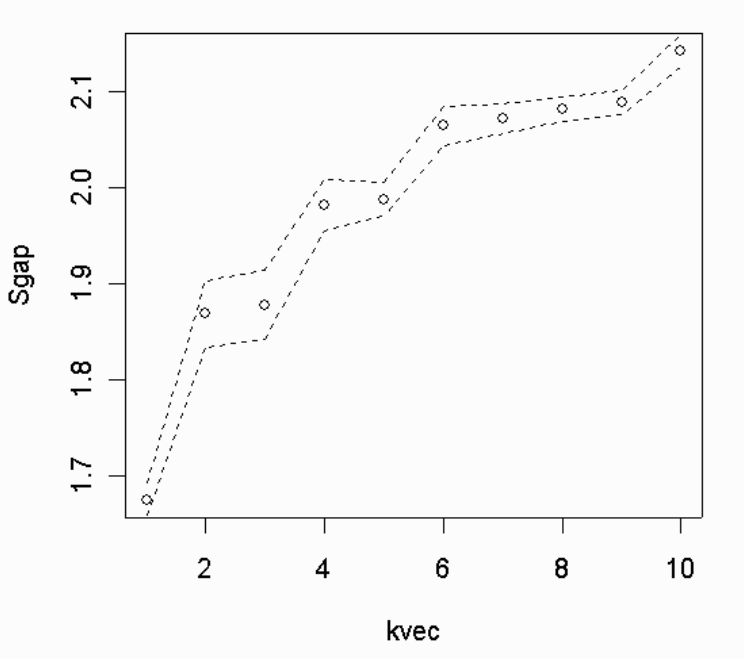
The choice of initial partition can greatly affect the final clusters that result.

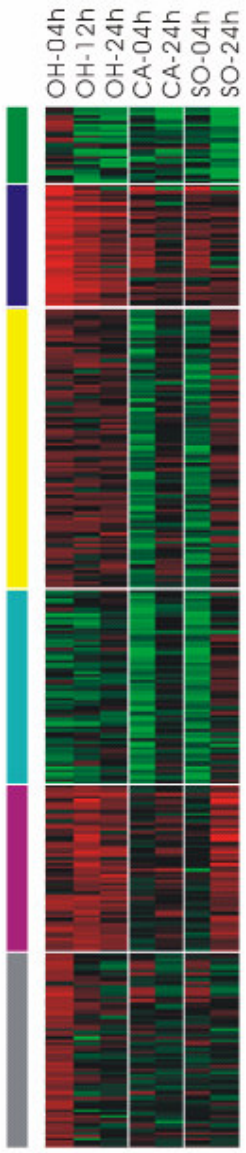
K-Means Clustering

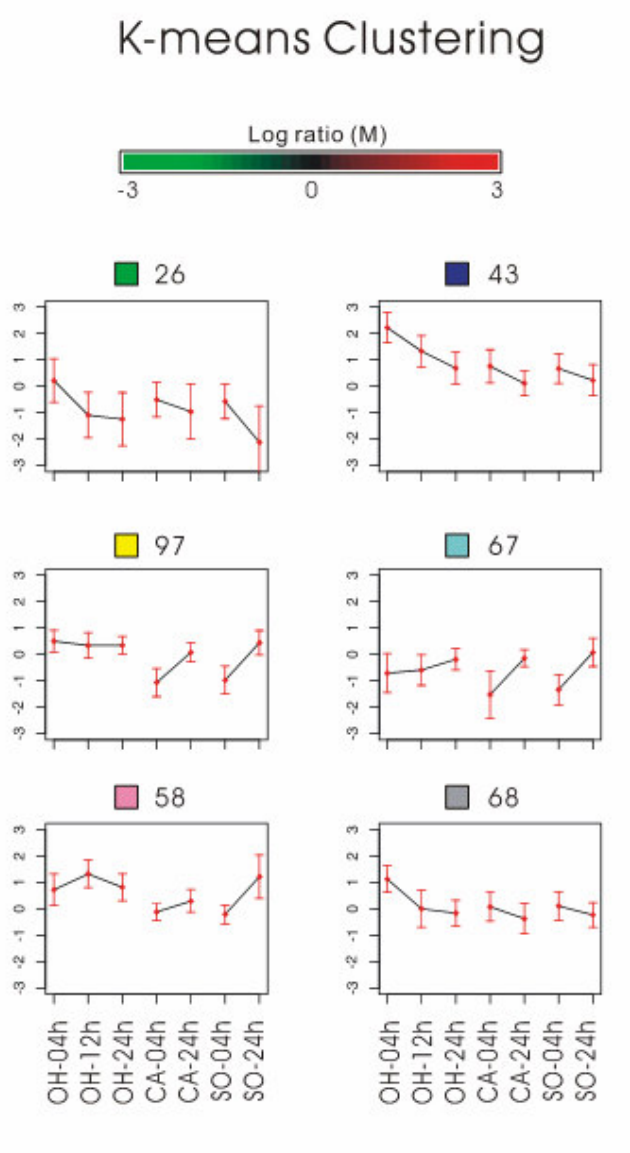
- Data

Baseline: Culture Medium (CM-00h)
OH-04h, OH-12h, OH-24h
CA-04h, CA-24h
SO-04h, SO-24h

- A set of 359 genes was selected for clustering.

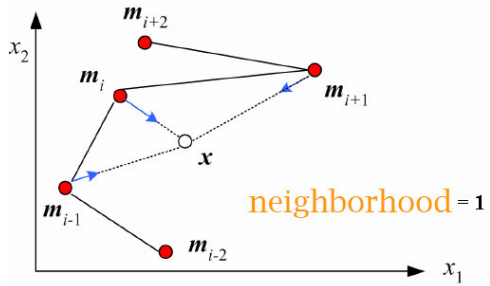






Self-Organizing Maps (SOM)

- SOMs were developed by **Kohonen** in the early **1980's**, original area was in the area of speech recognition.
- ***Idea:*** Organise data on the basis of **similarity** by putting entities **geometrically** close to each other.



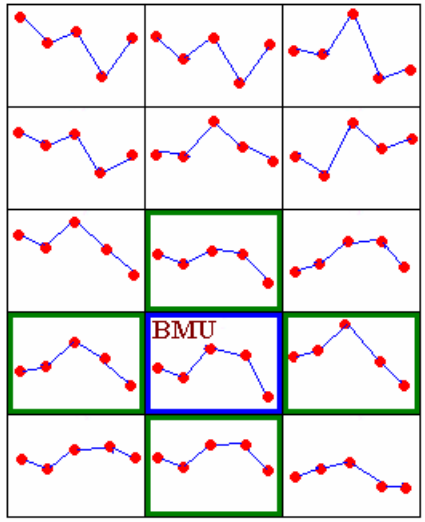
- SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display.

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

Learning process:

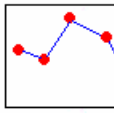
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,] & i \in N_c(t) \\ \mathbf{w}_i(t), \quad \text{o.w.} & \end{cases}$$

5 x 3 output node



Data Matrix

Table	X01	X02	X03	•••	Xp
obs 001	-0.48	-0.42	0.87		-0.35
obs 002	-0.39	-0.58	1.08		-0.58
obs 003	0.87	0.25	-0.17		-0.13
obs 004	1.57	1.03	1.22		-1.02
obs 005	-1.15	-0.86	1.21		-0.44
obs 006	0.04	-0.12	0.31		0.08
obs 007	2.95	0.45	-0.40		-0.76
obs 008	-1.22	-0.74	1.34		-0.55
obs 009	-0.73	-1.06	-0.79		0.03
obs 010	-0.58	-0.40	0.13		-0.45
obs 011	-0.50	-0.42	0.66		0.01
obs 012	-0.86	-0.29	0.42		-0.63
obs 013	-0.16	0.29	0.17		-0.04
obs •••					
obs n	-1.79	0.94	2.13		-0.66



X01 ••• Xp

input node

Incrementally decrease the learning rate and the neighborhood size, and repeat

Algorithm of SOM

Step 0: Initialize weights $\mathbf{w}_i(t)$.
Set topological neighborhood parameters $N_c(t)$.
Set learning rate parameters $\alpha(t)$ and $h_{ci}(t)$.

Step 1: For each input vector $\mathbf{x}(t)$, do

a. Finding a BMU: $\|\mathbf{x}(t) - \mathbf{w}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{w}_i(t)\|$

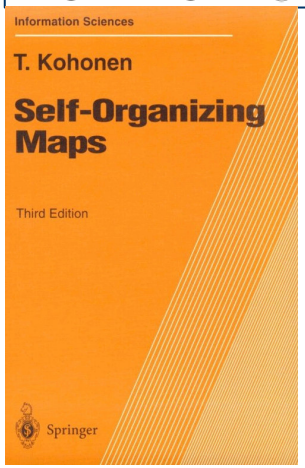
b. Learning process:

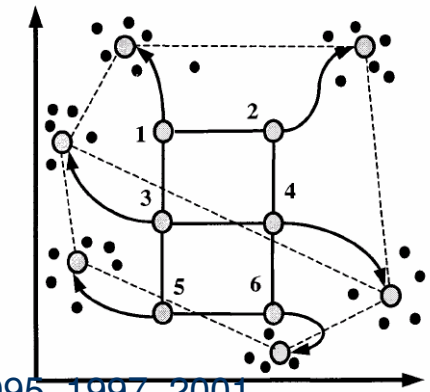
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

c. Go to the next unvisited input vector. If there are no unvisited input vector left then go back to the very first one and go to Step 2.

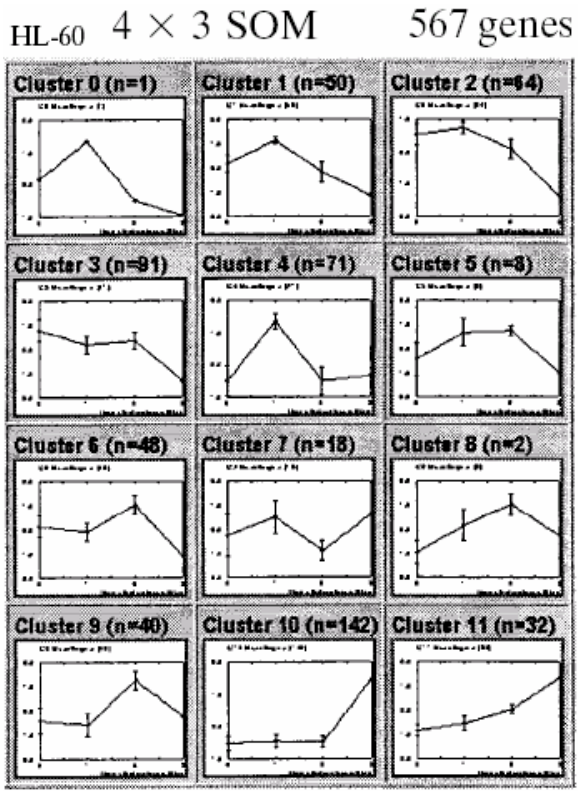
Step 2: Incrementally decrease the learning rate and the neighborhood size, and repeat Step 1.

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





1995, 1997, 2001



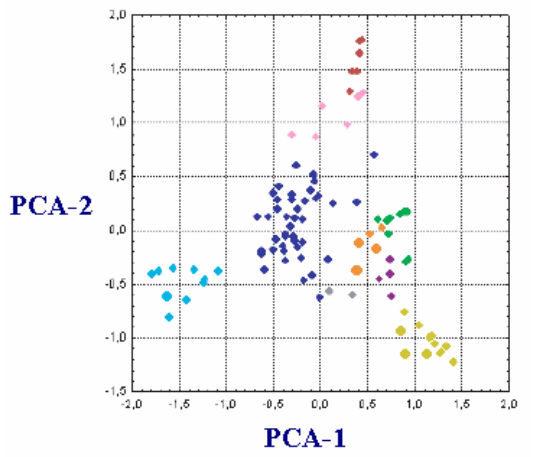
Macrophage Differentiation in HL-60 cells

Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96:2907-2912.

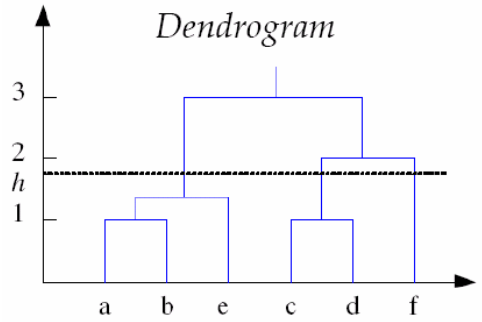
Choosing the Number of Clusters

(1) K is defined by the application.

(2) Plot the data in two PAC dimensions.

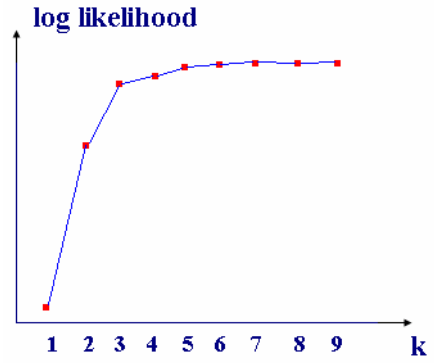


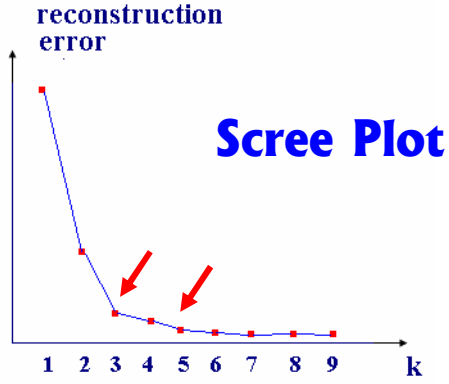
(4) Hierarchical clustering: look at the difference between levels in the tree.



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

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





Scree Plot

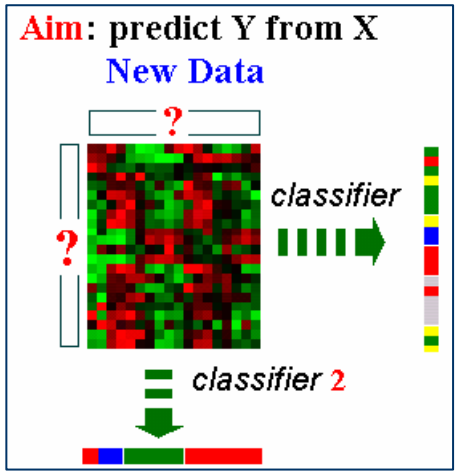
Calinski and Harabasz (1974): CH(k)
Hartigan (1975): $H(k)$
Krzanowski and Lai (1985): KL(k)
Kaufman and Rousseeuw (1990): $s(i)$

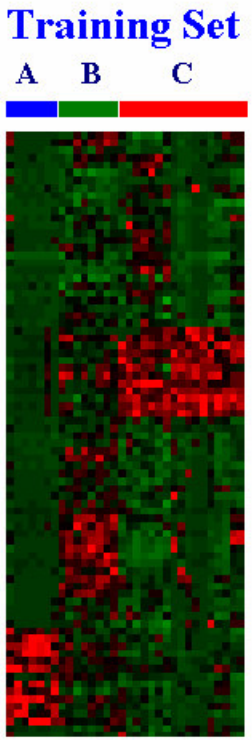
J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

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

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

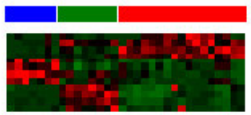
Classification of Genes, Tissues or Samples (Supervised Learning)



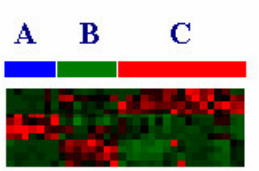


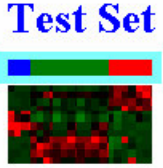
Possible to
1. classification for genes
2. classification for samples (arrays)

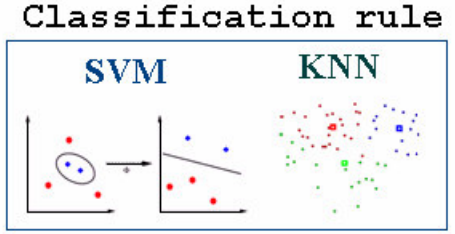
Gene Selection Methods



Construct







Assign class labels

predicted → classification error

true →

prediction error

Linear Discriminant Analysis (LDA)

- LDA (Fisher, 1936) finds the linear combinations **xa** of the gene expression profiles **x = (x1,…, xp)** with large ratios of between-groups to within-groups sum of squares.

Genes (variables)

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

mRNA samples (observations)

$X_{[n \times p]}$: data matrix.

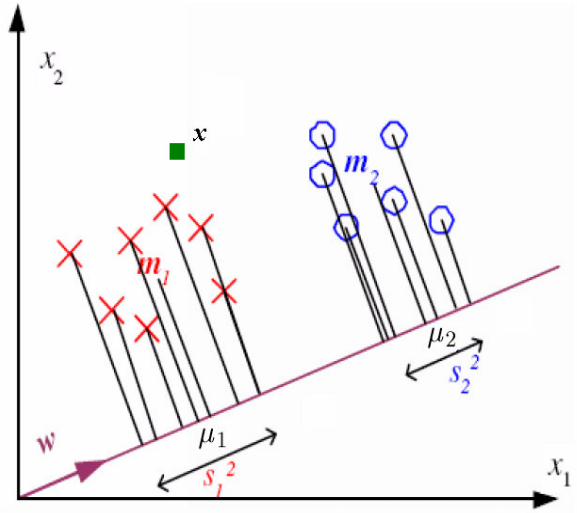
Aim : $\underset{\mathbf{a}}{\text{Max}}\,(\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a})$

Xa : linear combination of the columns of X.

$\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$: ratio of between-groups to within-groups sum of squares.

$B_{[p \times p]}$: matrices of between-groups sum of squares.

$W_{[p \times p]}$: matrices of within-groups sum of squares.



Solution:

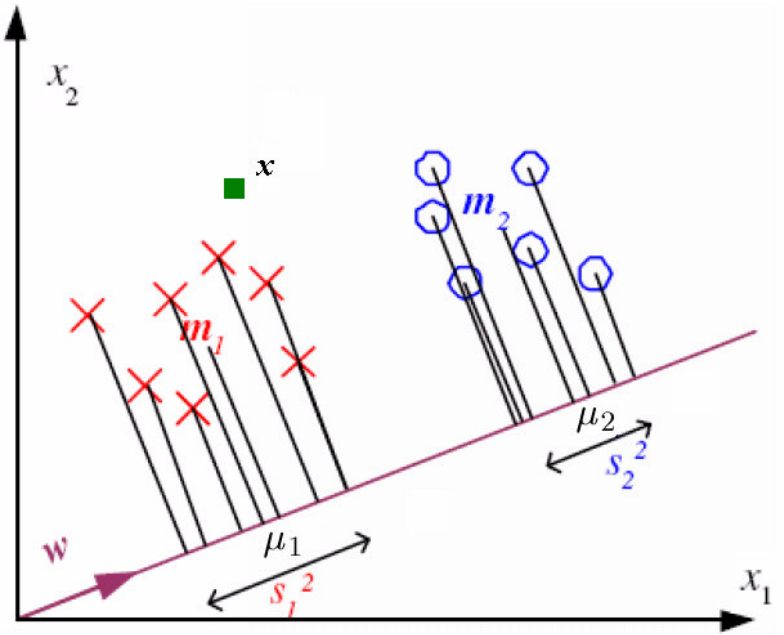
The matrix $W^{-1}B$ has at most $s = \min(K-1, p)$ non-zero eigenvalues,

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$, with corresponding linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$.

The *discriminant variables* $u_l = \mathbf{x}\mathbf{v}_l$, $l = 1, \ldots, s$.

LDA: Classification

- Fisher's linear discriminant is optimal if the classes are normally distributed.
- After projection, for the two classes to be well separated we would like the means to be as far apart as possible and the examples of classes be scatteres in as small a



Classification Rules:

For an observation $\mathbf{x} = (x_1, \ldots, x_d)$

$d_k(\mathbf{x}) = ((\mathbf{x} - \bar{\mathbf{x}}_k)\mathbf{w})^2$

denote its (squared) Euclidean distance, in terms of the discriminant variables, from the $1 \times d$ vector of class k averages $\bar{\mathbf{x}}$ for the learning set $\mathcal{L}$.

The predicted class for observation $\mathbf{x}$ is

$\mathcal{C}(\mathbf{x}, \mathcal{L}) = \text{argmin}_k \, d_k(\mathbf{x})$,

the class whose mean vector is closest to $\mathbf{x}$ in the space of discriminant variables.

Fisher's Criterion for the Gene Selection

Lymphoma dataset

three most prevalent adult lymphoid malignancies 人類淋巴腫瘤

B-cell chronic lymphocytic leukemia (B-CLL) : 29 cases B細胞慢性淋巴性白血病

follicular lymphoma (FL) : 9 cases 濾泡型淋巴瘤

diffuse large B-cell lymphoma (DLBCL) : 43 cases 瀰漫性大B細胞淋巴癌

gene expression data for $p = 4,682$ genes in $n = 81$ mRNA samples.

Gene selection

For a gene j

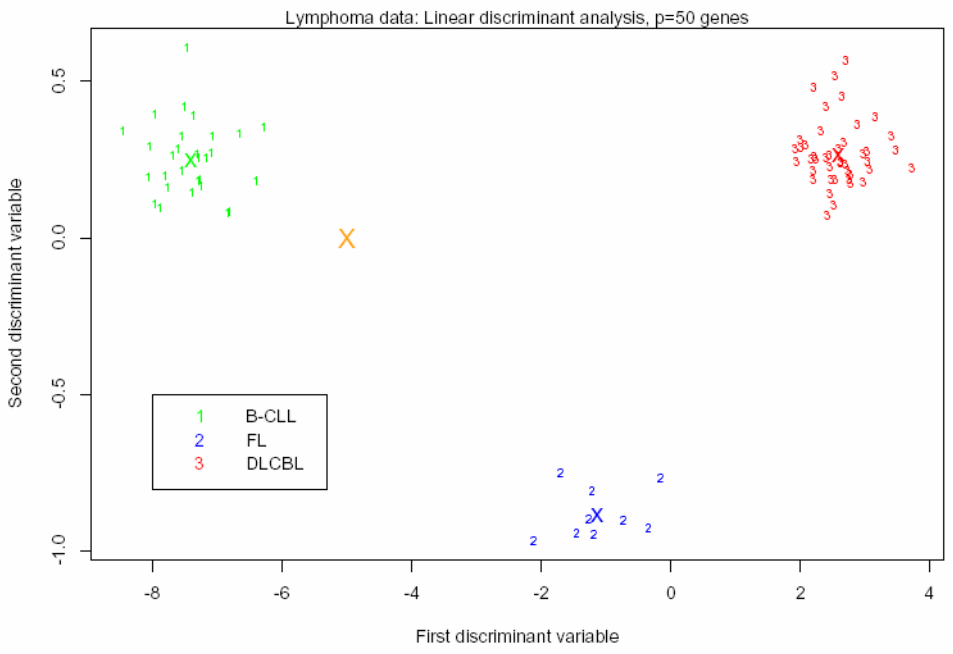
$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

$\bar{x}_{\cdot j}$ denotes the average expression level of gene j across all samples.

$\bar{x}_{kj}$ denotes the average expression level of gene j across samples belonging to class k.

Select

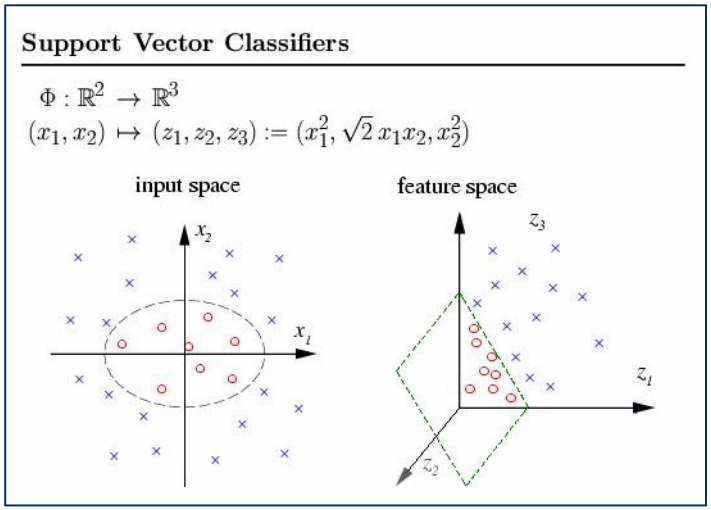
the p genes with the largest BSS/WSS ratios.



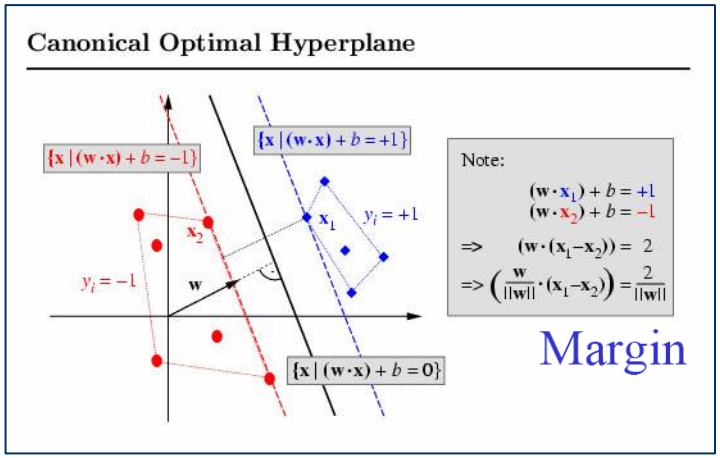
Dudoit S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 97 (457), 77-87.

Support Vector Machine (SVM)

SVMs (Vapnik, 1995) map the data (input space) into high dimensional space (feature space) through a kernel function ϕ and then find a hyperplane $\mathbf{w}$ to separate two groups (binary classification).



Kernel Machines



Quadratic Optimization Problem

- To find the optimal hyperplane
  (solve the quadratic optimization problem)
  To minimize the quadratic form $|W|^2 = (W * W)$ subject to the linear constraints $y_i((x_i * W) + b_0) \geq 1$

decision function

$f(\mathbf{X}) = \text{sign}\,((\mathbf{X} * W) + b_0)$

Multi-class problem

Two approaches for multi-class classification:

- **one-against-others**: The kth SVM model is constructed with all of the samples in the kth class with one group, and all other samples with the other group.
- **one-against-one**: The SVM trained model is constructed by using any two of classes. Therefore, there are total $K(K-1)/2$ classifiers.

Software

SVMTorch, Collobert and Bengio, 2001
LIBSVM, Chang and Lin, 2002

SVM

Brown et al. (2000). Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines, PNAS 97(1), 262-267.

Assume: Genes of similar function yield similar expression pattern.

Data

Yeast Gene Expression
[2467x 80] out of
[6,221x 80] has
accurate functional
annotations.

Tricarboxylic acid
Respiration
Ribosome
Proteasome
Histone
Helix-turn-helix

Table 1. Comparison of error rates for various classification methods

Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	Parzen	4	12	5	2,446	6
	FLD	9	10	7	2,441	5
	C4.5	7	17	0	2,443	−7
	MOC1	3	16	1	2,446	−1
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	7	7	23	2,430	39
	D-p 3 SVM	6	8	22	2,431	38

Table 3. Predicted functional classifications for previously unannotated genes

Class	Gene	Locus	Comments
TCA	YHR188C		Conserved in worm, *Schizosaccharomyces pombe*, human
	YKL039W	PTM1	Major transport facilitator family; likely integral membrane protein; similar YHL017w not co-regulated.
Resp	YKR016W		Not highly conserved, possible homolog in *S. pombe*
	YKR046C		No convincing homologs
	YPR020W	ATP20	Subsequently annotated: subunit of mitochondrial ATP synthase complex
	YLR248W	CLK1/RCK2	Cytoplasmic protein kinase of unknown function
Ribo	YKL056C		Homolog of translationally controlled tumor protein, abundant, conserved and ubiquitous protein of unknown function

⋮

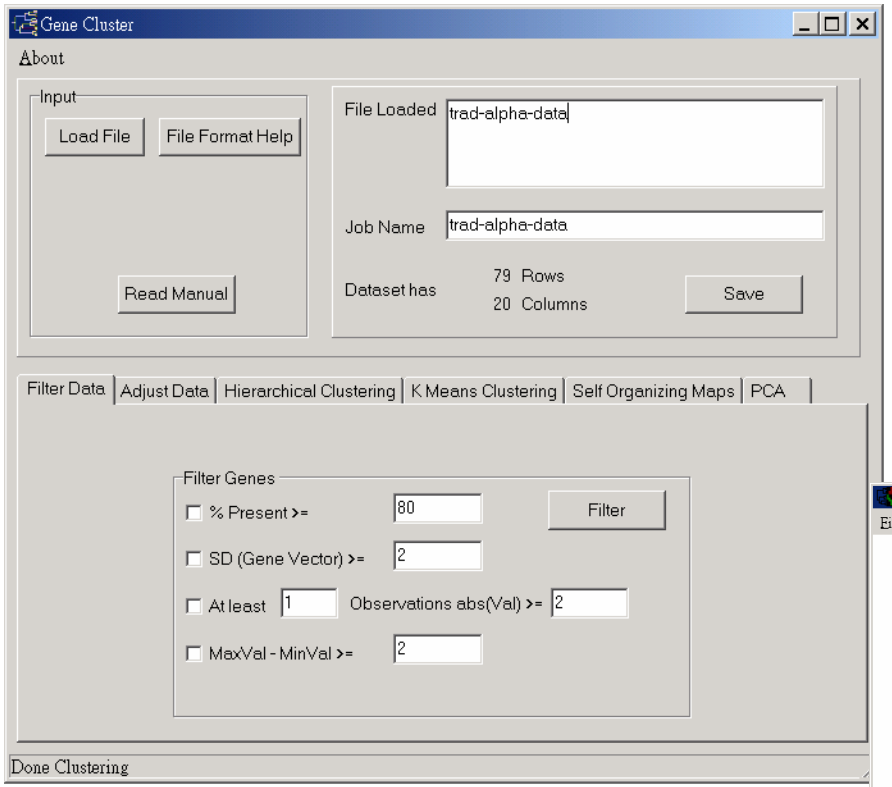
Kernel Machines:
http://www.kernel-machines.org
Support Vector Machines:
http://www.support-vector.net
MATLAB Support Vector Toolbox:
http://www.isis.ecs.soton.ac.uk/resources/svminfo
SVM Application List:
http://www.clopinet.com/isabelle/Projects/SVM/applist.html

Software

Statistical Analysis and Visualization

- *Freeware/Shareware*
  - Cluster and TreeView
  - The Bioconductor
  - GAP
- *Commercial*
  - Matlab: Bioinformatics ToolBox
  - GeneSpring

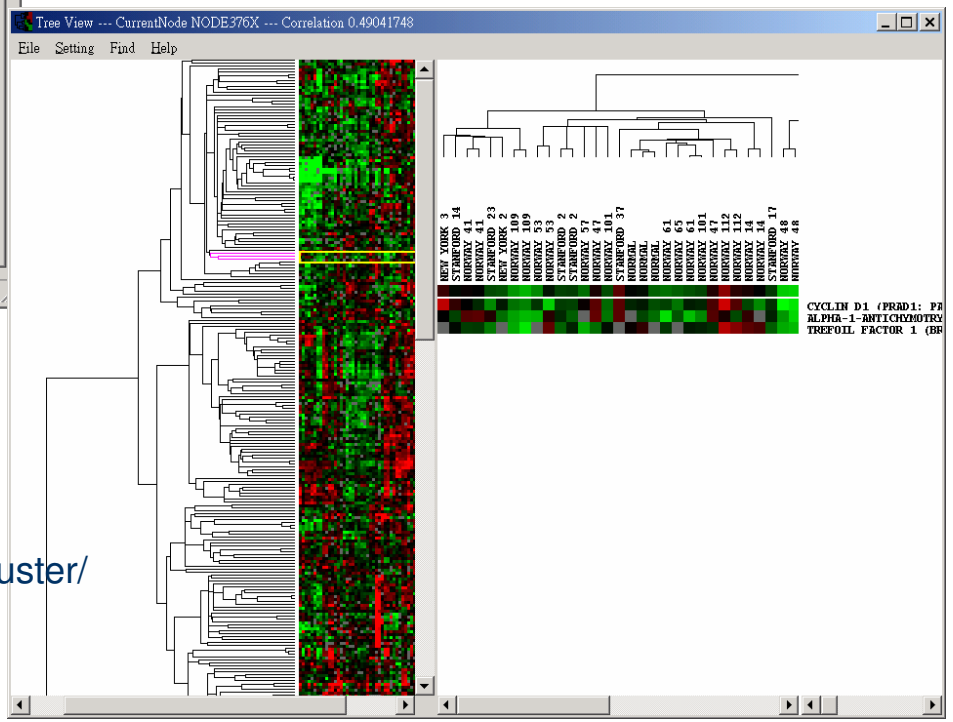
Cluster and TreeView



Eisen Lab

http://rana.lbl.gov/EisenSoftware.htm

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci.* 95(25):14863-8.



De Hoon, M.J.L.; Imoto, S.; Nolan, J.; Miyano, S.; "**Open source clustering software**". Bioinformatics, 20 (9): 1453--1454 (2004)

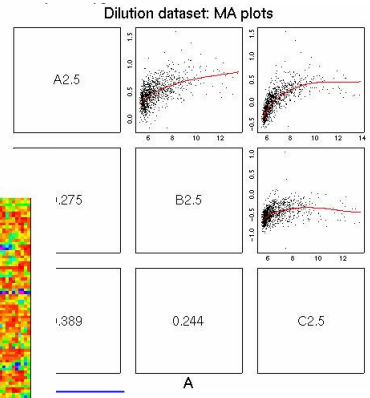
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

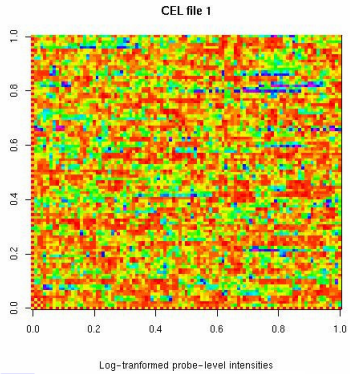
The Bioconductor

Package

AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Rgraphviz
Ruuid

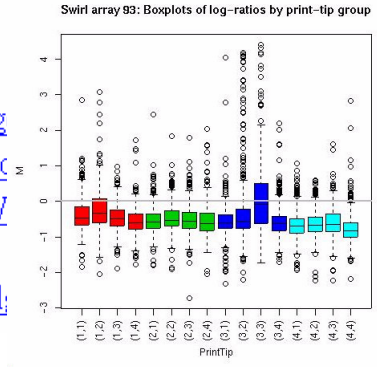
genefilter
geneplotter
globaltest
gpls
graph
hexbin
limma





daMA
edd
externalVector
factDesign
gcrma

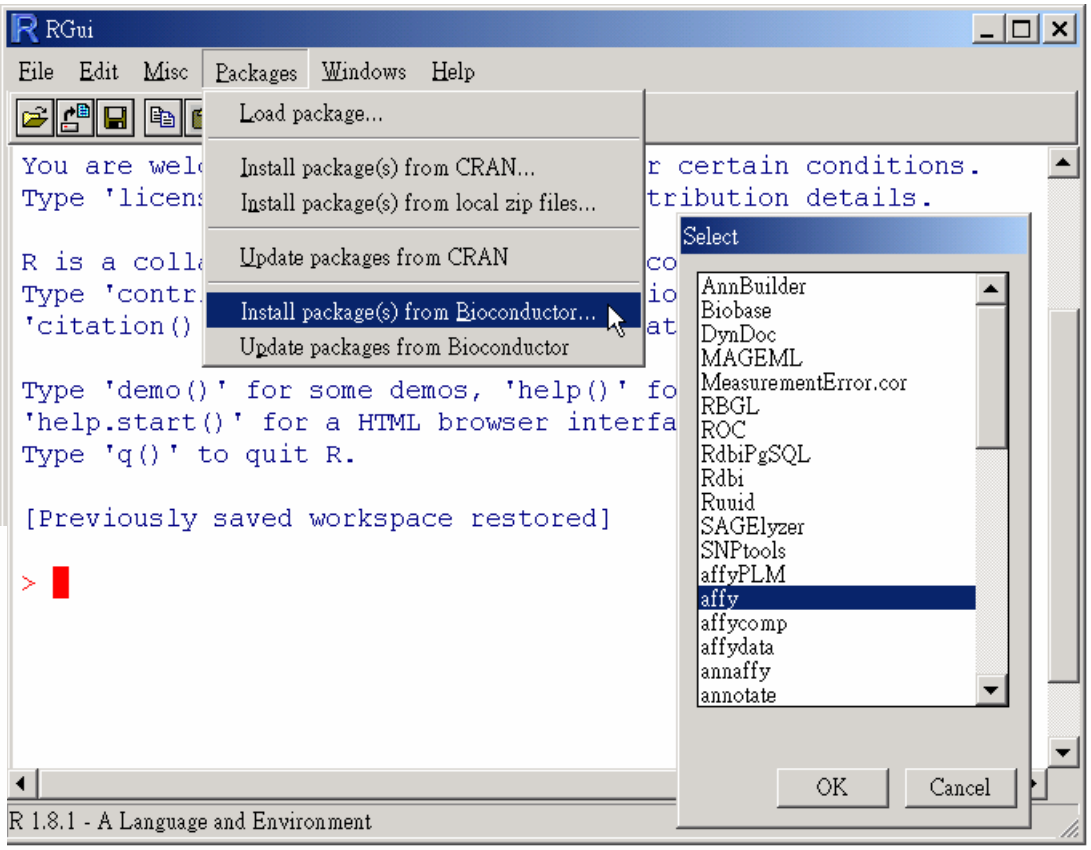
sigg
splic
tkW
vsn
wid



The Bioconductor
version 1.5 (2004-11-01)
http://www.bioconductor.org

The R Project for Statistical Computing

R version 2.1.0 (**2005-04-18**)
http://www.r-project.org



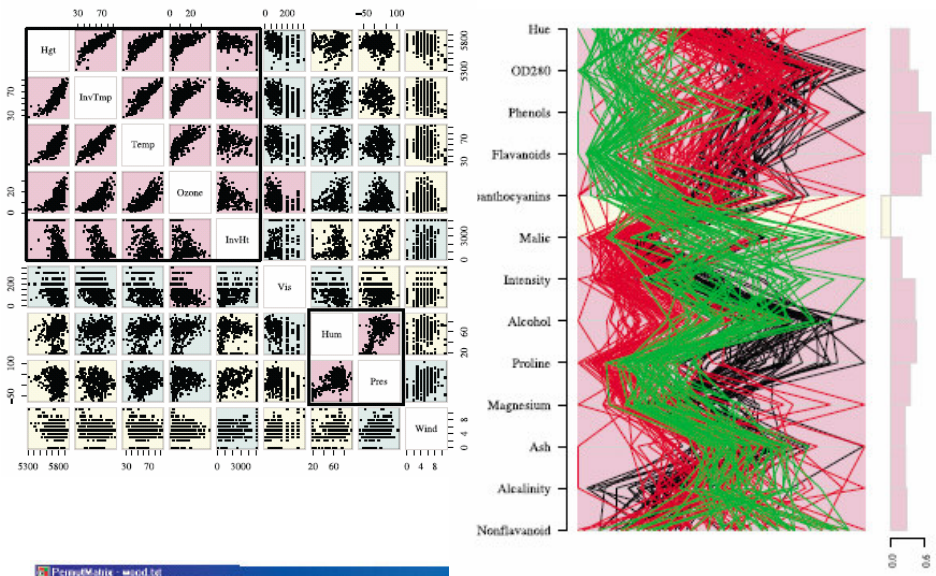
Gclus, PermutMatrix

- **gclus: Clustering Graphics**

  (R package)

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

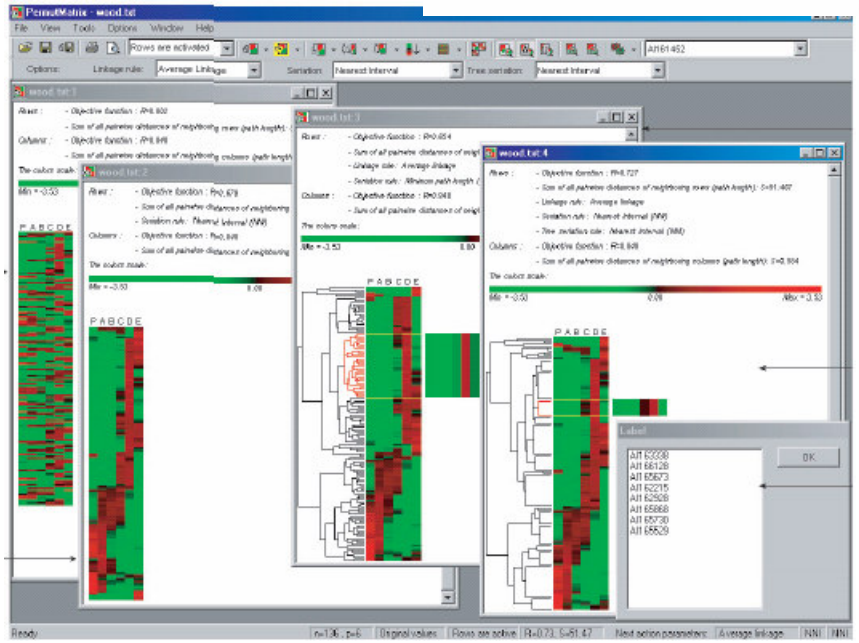
  Catherine B. Hurley, (2004), Clustering Visualizations of Multidimensional Data, Journal of Computational & Graphical Statistics, Vol. 13, No. 4, pp.788-806



- **PermutMatrix**

  http://www.lirmm.fr/~caraux/PermutMatrix

  Caraux, G., and Pinloche, S. (2005), "Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order," Bioinformatics, 21, 1280-1281.



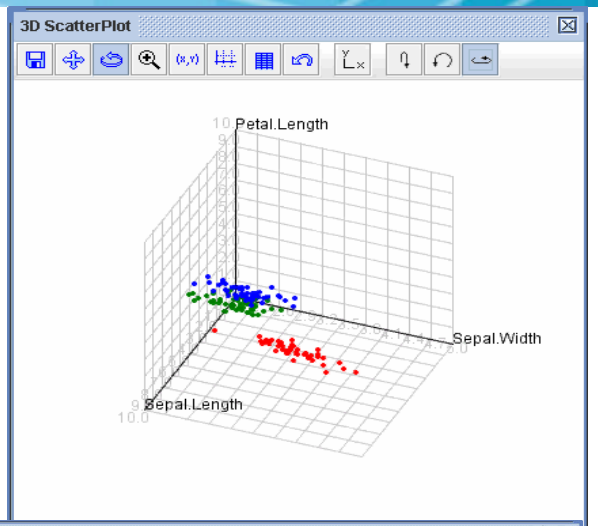
GAP (Generalized Association Plots)

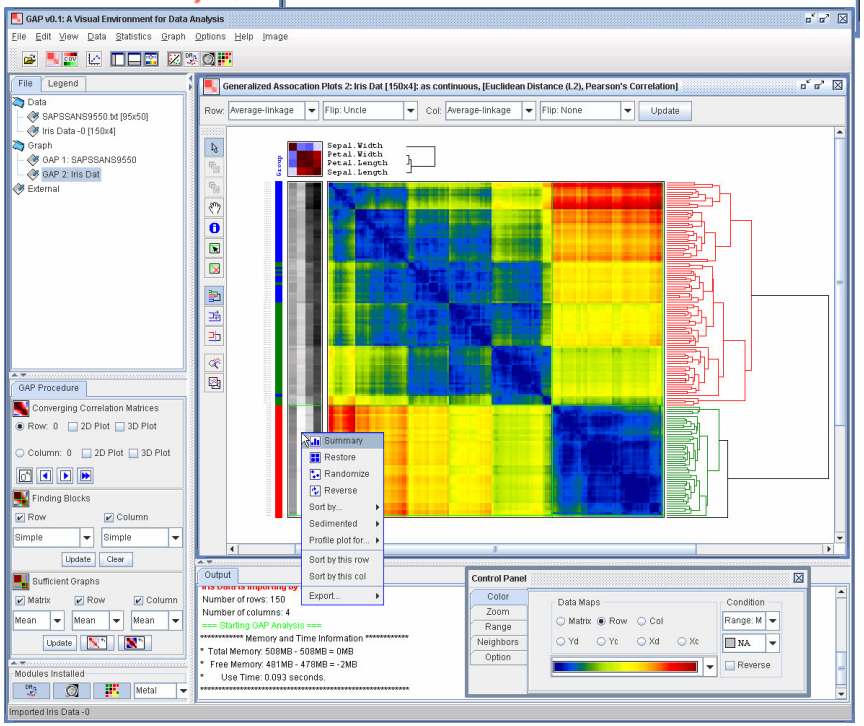
Generalized Association Plots

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

Statistical Plots

- 2D Scatterplot, 3D Scatterplot (Rotatable)





Chen, C. H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. Statistica Sinica 12, 7-29.
Wu, H. M., Tien, Y. J. and Chen, C. H. (2006). GAP: a Graphical Environment for Matrix Visualization and Information Mining.

Web Site

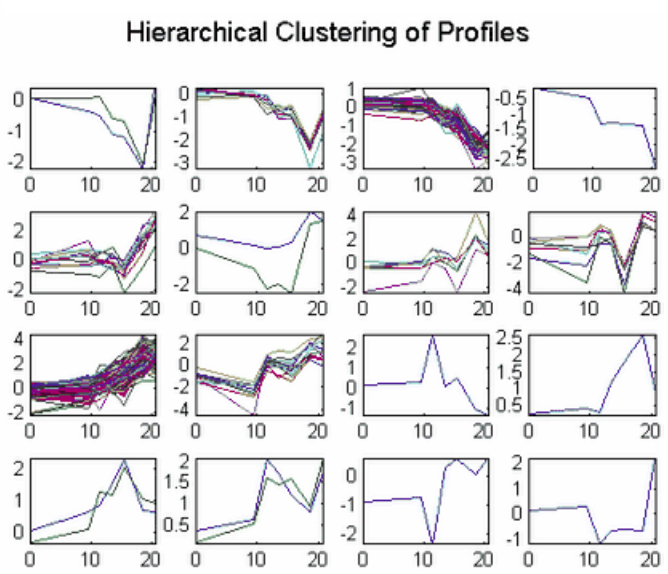
http://gap.stat.sinica.edu.tw/Software/GAP

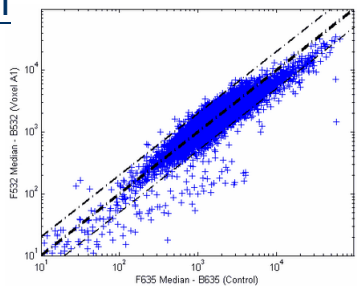
Matlab: Bioinformatics ToolBox

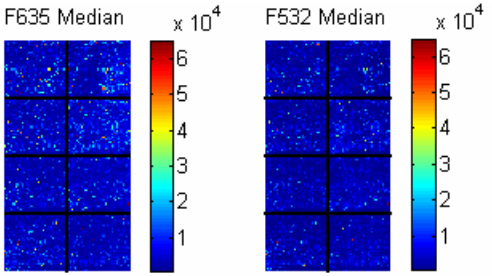
The MathWorks
Bioinformatics Toolbox

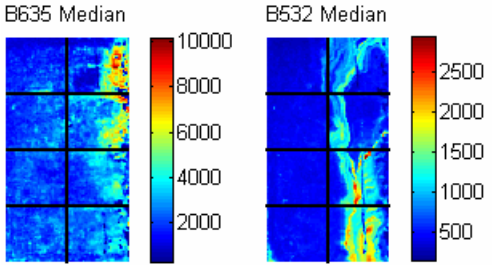
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html

- Data Formats and Databases — Access online databases, read and write to files with standard genome and proteome formats such as FASTA and PDB.
- Sequence Alignments — Compare nucleotide or amino acid sequences using pairwise and multiple sequence alignment functions.
- Sequence Utilities and Statistics — Manipulate sequences and determine physical, chemical, and biological characteristics.
- Microarray Analysis — Read, filter, normalize, and visualize microarray data.
- Protein Structure Analysis — Determine protein characteristics and simulate enzyme cleavage reactions.
- Prototype and Development Environment — Create new algorithms, try new ideas, and compare alternatives.
- Share Algorithms and Deploy Applications — Create GUIs and stand-alone applications.



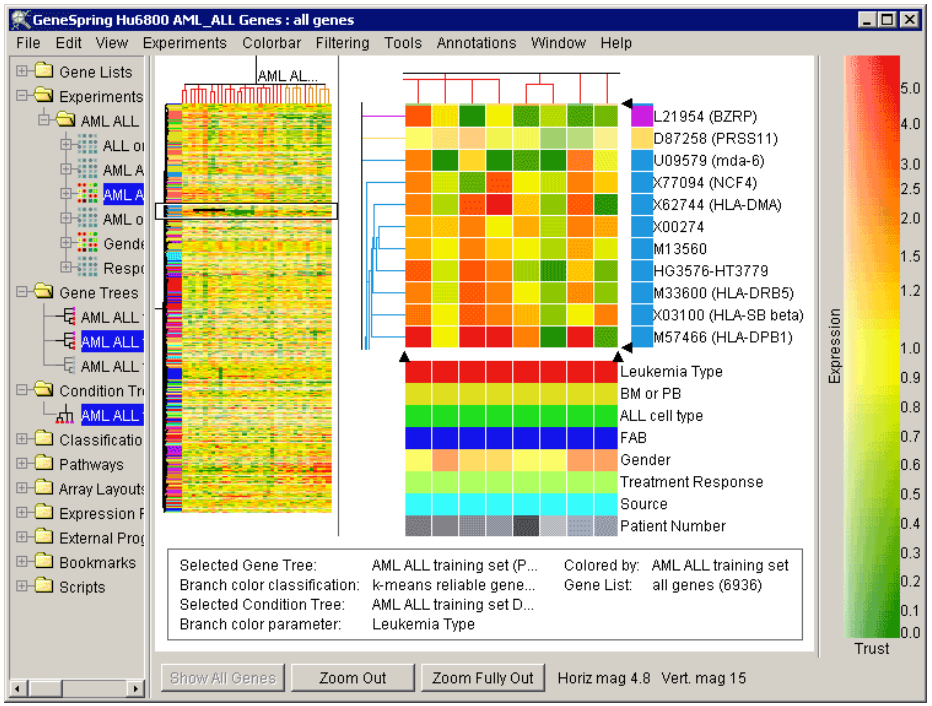


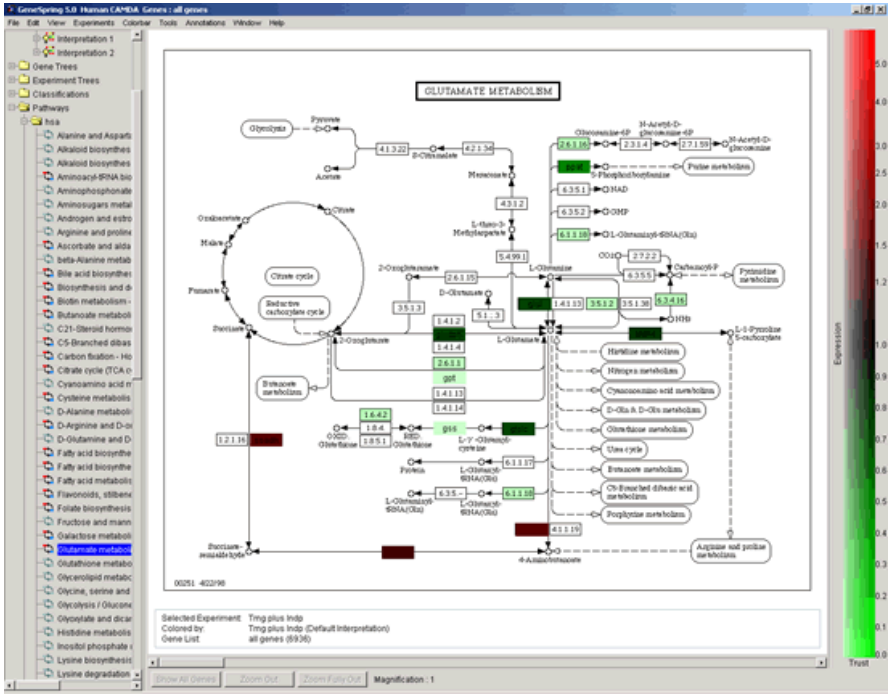




GeneSpring GX v7.3

- RMA or GC-RMA probe level analysis
- Advanced Statistical Tools
- Data Clustering
- Visual Filtering
- 3D Data Visualization
- Data Normalization (Sixteen)
- Pathway Views
- Search for Similar Samples
- Support for MIAME Compliance
- Scripting
- MAGE-ML Export





Images from
http://www.silicongenetics.com

Agilent Technologies

2004 Articles Citing GeneSpring®

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

More than 700 papers

Useful Links and Reference

Y.F.Leung's FUNCTIONAL GENOMICS

Home | Functional Genomics | Microarray | Bioinformatics | Proteomics | Genome mapping Complex disease mapping Linkage analysis | About Y.F.Leung

http://ihome.cuhk.edu.hk/~b400559/

AFFYMETRIX

http://www.affymetrix.com

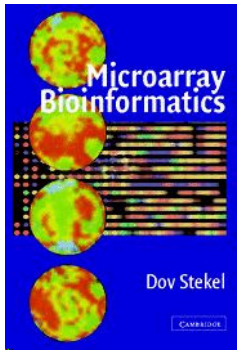
BIOINFORMATICS

http://bioinformatics.oupjournals.org

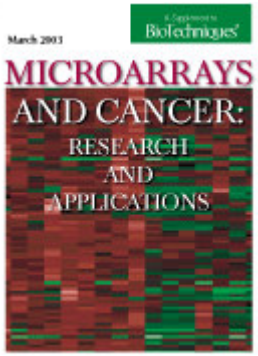
Bibliography on Microarray Data Analysis

http://www.nslij-genetics.org/microarray/

Stekel, D. (2003). Microarray bioinformatics, New York : Cambridge University Press.



- Speed Group Microarray Page: Affymetrix data analysis
  http://www.stat.berkeley.edu/users/terry/zarray/Affy/affy_index.html
- Statistics and Genomics Short Course, Department of Biostatistics Harvard School of Public Health.
  http://www.biostat.harvard.edu/~rgentlem/Wshop/harvard02.html
- Statistics for Gene Expression
  http://www.biostat.jhsph.edu/~ririzarr/Teaching/688/
- Bioconductor Short Courses
  http://www.bioconductor.org/workshop.htm



Microarrays and Cancer: Research and Applications
http://www.biotechniques.com/microarrays/

Google