PART-I

Microarray Data Analysis

Basic

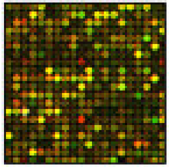
Course: 國立臺灣大學 資訊所
生物資訊與計算分子生物學
2005/12/06

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

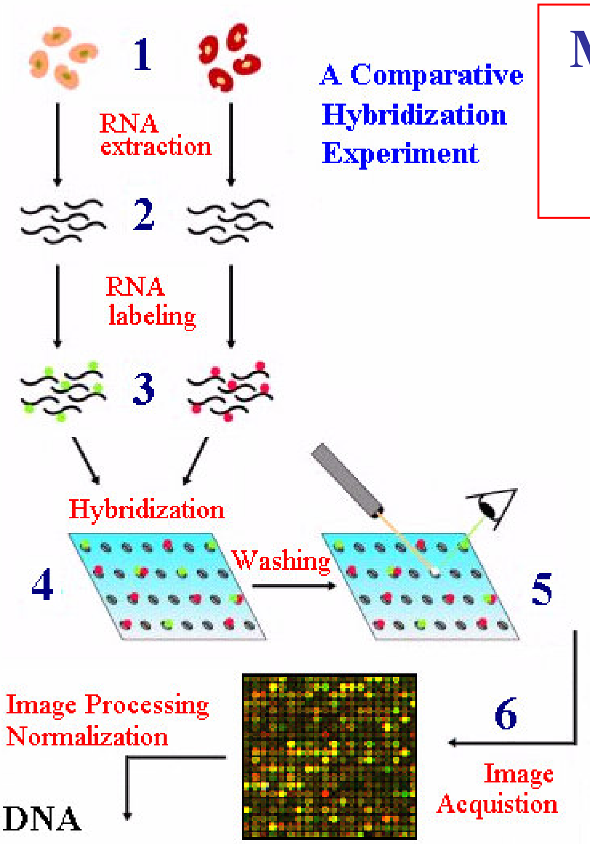
中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

Outlines

- Overview of Microarray Data Analysis
- *Graphical Presentation of Slide Data and Some Statistical Plots*
- *Preprocessing*: **Image Processing**, Normalization
- *Finding Differentially Expressed Genes*
  - Fold Changes Method and Hypothesis Testing
  - Multiple-testing Problem
- *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*

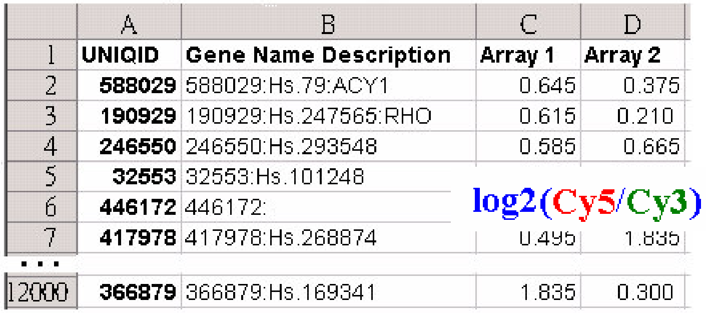


Overview of cDNA Microarray Experiment



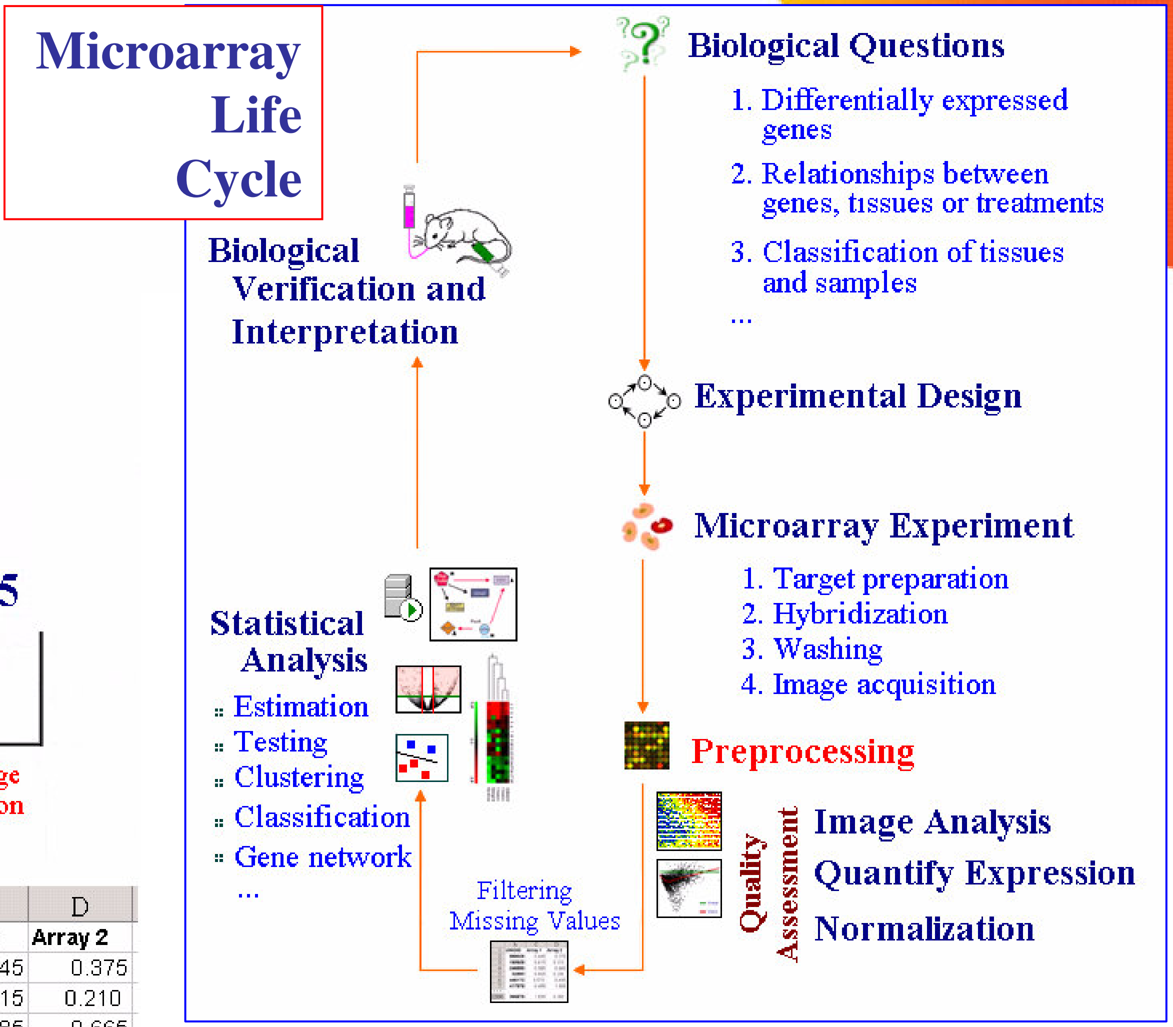
A Comparative Hybridization Experiment

cDNA Microarray Data



	A	B	C	D
1	UNIQID	Gene Name Description	Array 1	Array 2
2	588029	588029:Hs.79:ACY1	0.645	0.375
3	190929	190929:Hs.247565:RHO	0.615	0.210
4	246550	246550:Hs.293548	0.585	0.665
5	32553	32553:Hs.101248		
6	446172	446172:		
7	417978	417978:Hs.268874	0.495	1.835
…				
12000	366879	366879:Hs.169341	1.835	0.300

log2(Cy5/Cy3)



Microarray Life Cycle

Biological Questions

1. Differentially expressed genes
2. Relationships between genes, tissues or treatments
3. Classification of tissues and samples

…

Experimental Design

Microarray Experiment

1. Target preparation
2. Hybridization
3. Washing
4. Image acquisition

Preprocessing

Quality Assessment

- Image Analysis
- Quantify Expression
- Normalization

Filtering Missing Values

Statistical Analysis

- Estimation
- Testing
- Clustering
- Classification
- Gene network

…

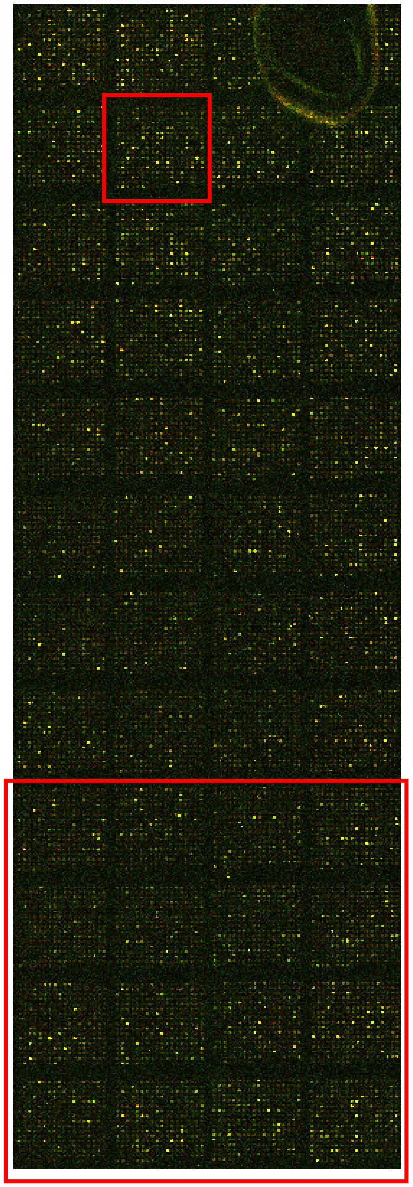
Biological Verification and Interpretation

Spot color	*Signal strength*	*Gene expression*
Yellow	Control = Treated	Unchanged
Red	Control < Treated	Induced
Green	Control > Treated	Repressed

R=Rf-Rb
G=Gf-Gb

M=log2R/G
A=1/2 log2RG

Array Image



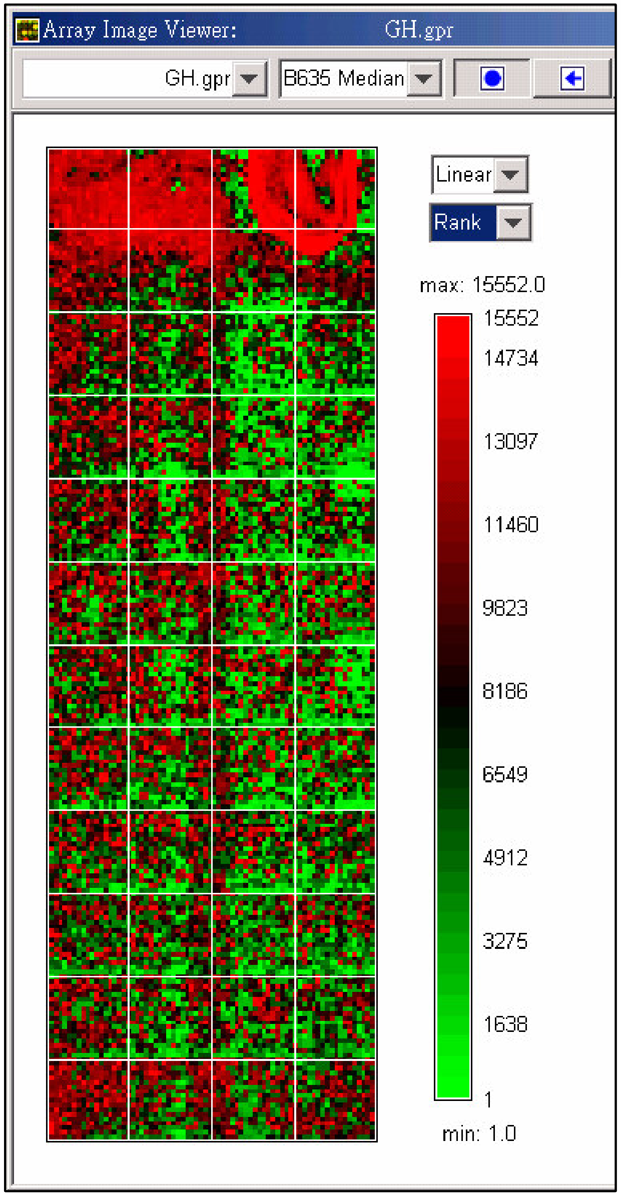
Blocks:
12 by 4

Features:
18 by 18

Signal
16-bit
0~65535

*.gpr

GAL



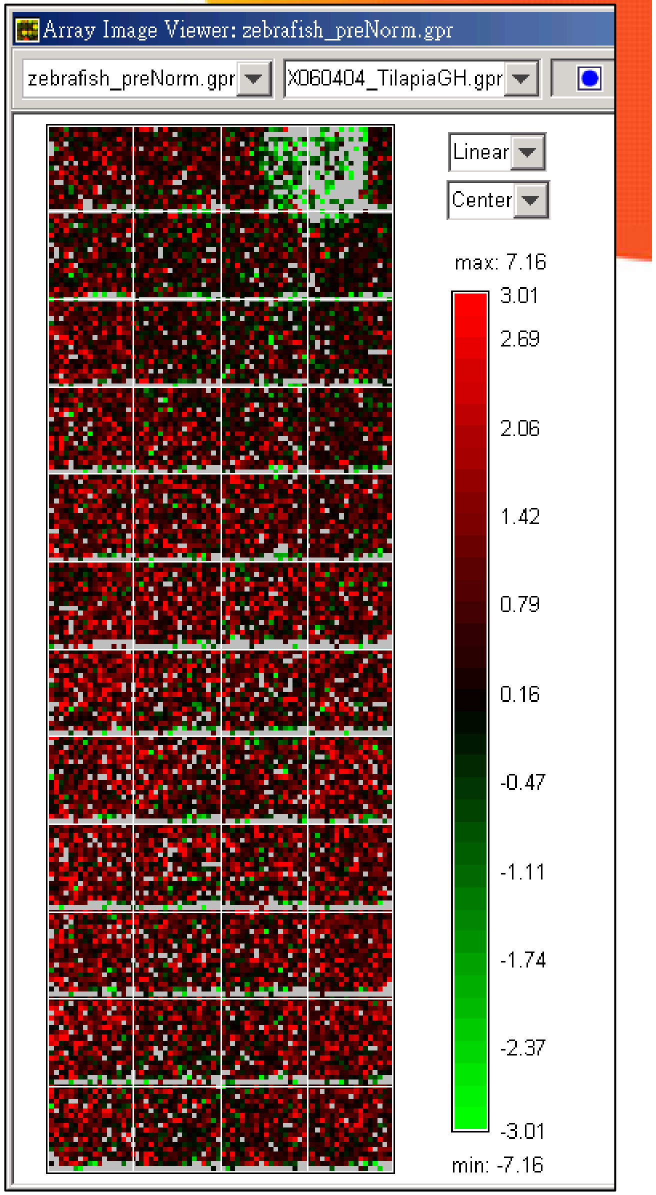
Array Image Viewer: GH.gpr

GH.gpr

B635 Median

Linear

Rank



Array Image Viewer: zebrafish_preNorm.gpr

zebrafish_preNorm.gpr

X060404_TilapiaGH.gpr

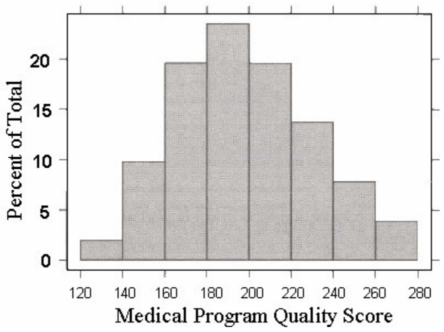
Linear

Center

Statistical Plots: Histogram

- 1/2h adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar.

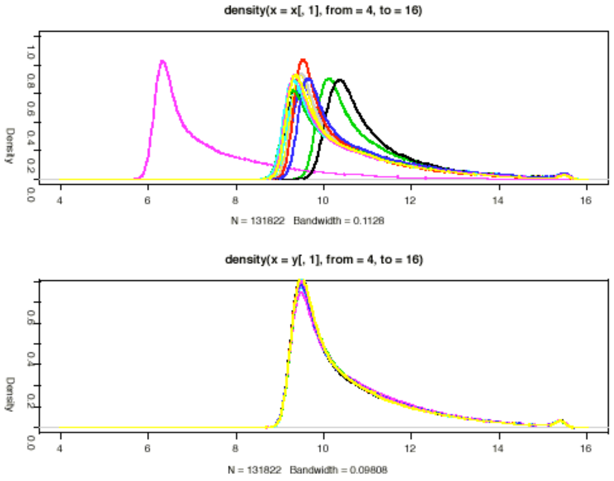
O. Bin origin at 120, bin widths of 20.

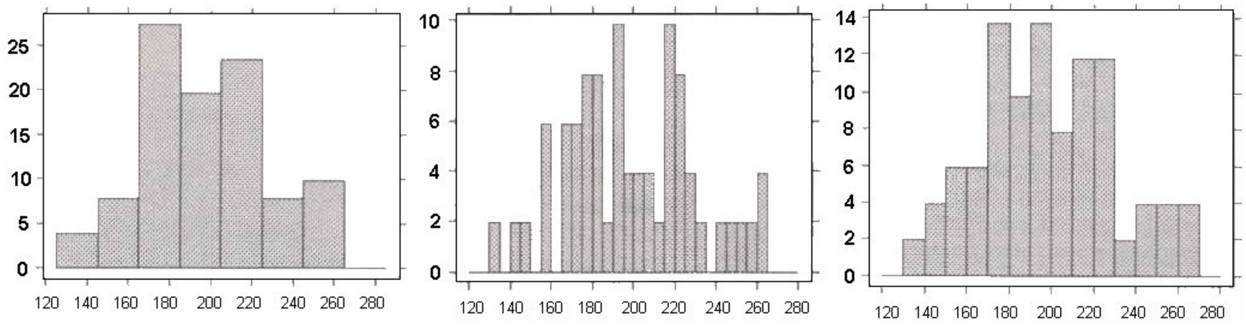


Disadvantage for displaying a variable's distribution:

- selection of origin of the bins.
- selection of bin widths.
- the very use of the bins is a distortion of information because any data variability within the bins cannot be displayed in the histogram.

Density Plots




A. Bin origin at 125, bin widths of 20.

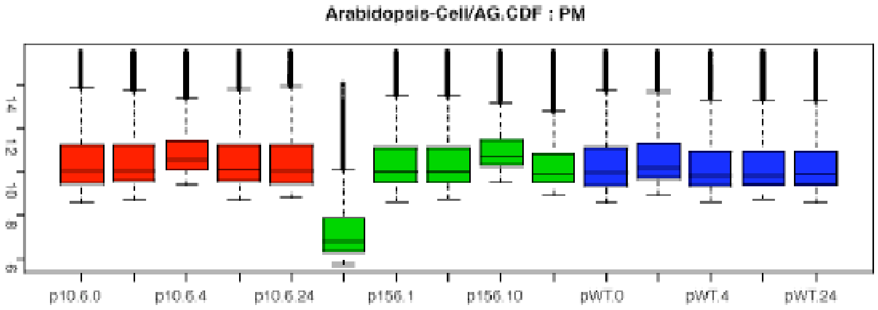
B. Bin origin at 120, bin widths of 5.

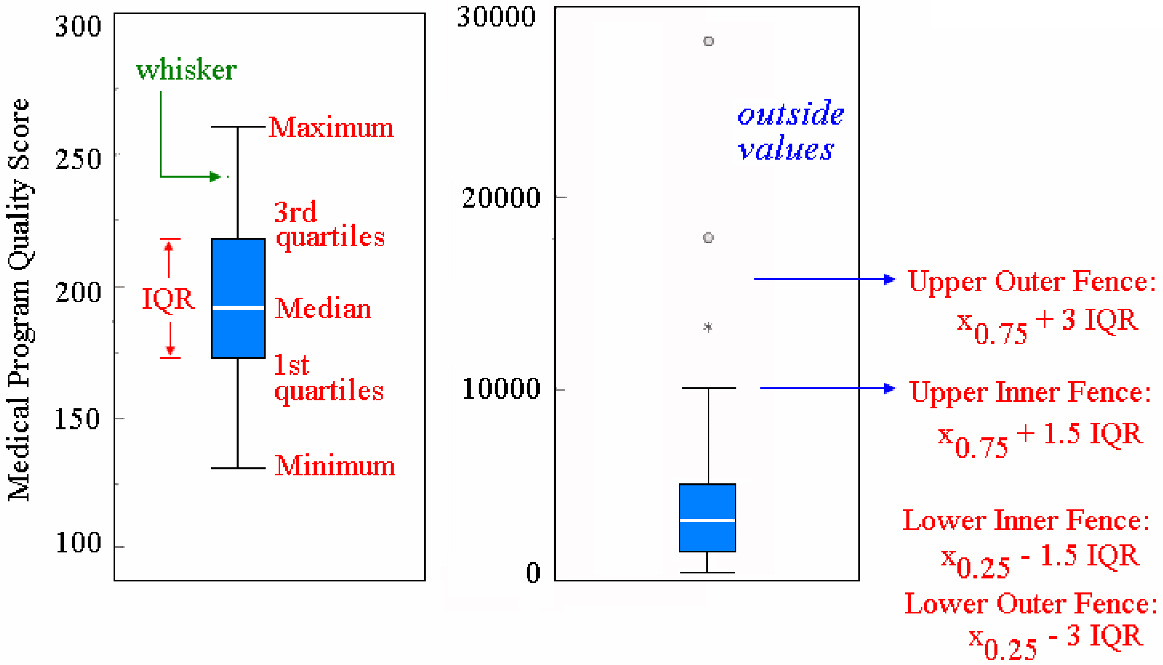
C. Bin origin at 120, bin widths of 10.

Figure Sources: Jacoby (1997).

Statistical Plots: Box Plots

- Box plots (Tukey 1977, Chambers 1983) are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.





Upper Outer Fence: $x_{0.75} + 3\ \mathrm{IQR}$

Upper Inner Fence: $x_{0.75} + 1.5\ \mathrm{IQR}$

Lower Inner Fence: $x_{0.25} - 1.5\ \mathrm{IQR}$

Lower Outer Fence: $x_{0.25} - 3\ \mathrm{IQR}$

The box plot can provide answers to the following questions:

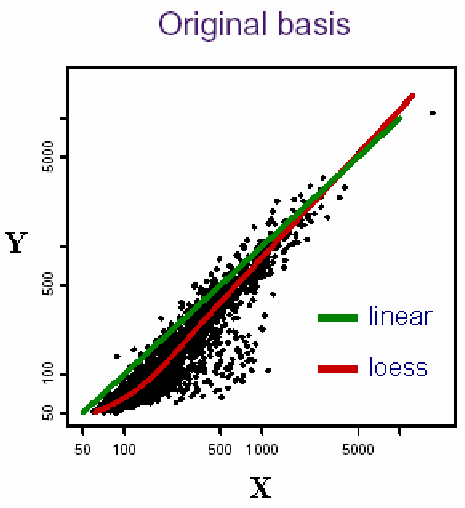
- Is a factor significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there any outliers?

Further reading: http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm

Scatterplot and MA plot

- **Features of scatter plot.**
  - the substantial correlation between the expression values in the two conditions being compared.
  - the preponderance of low-intensity values. (the majority of genes are expressed at only a low level, and relatively few genes are expressed at a high level)
- **Goals:** to identify genes that are differentially regulated between two experimental conditions.
- **Outliers in logarithm scale**
  - spreads the data from the lower left corner to a more centered distribution in which the prosperities of the data are easy to analyze.
  - easier to describe the fold regulation of genes using a log scale. In log2 space, the data points are symmetric about 0.

- **MA plots** can show the intensity-dependant ratio of raw microarray data.



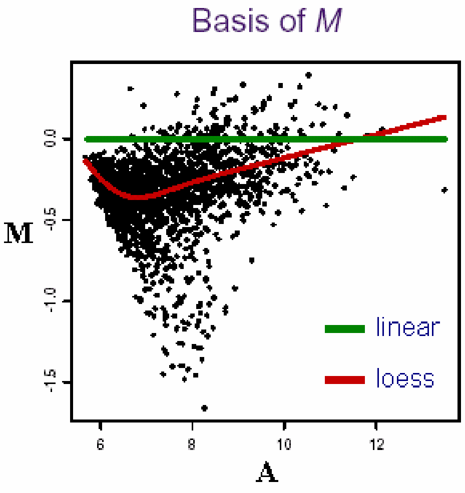
$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2}\log_2(X\,Y)$$

Oligo	cDNA

$X = PM_1$,
$Y = PM_2$
	X= Cy3
	Y= Cy5

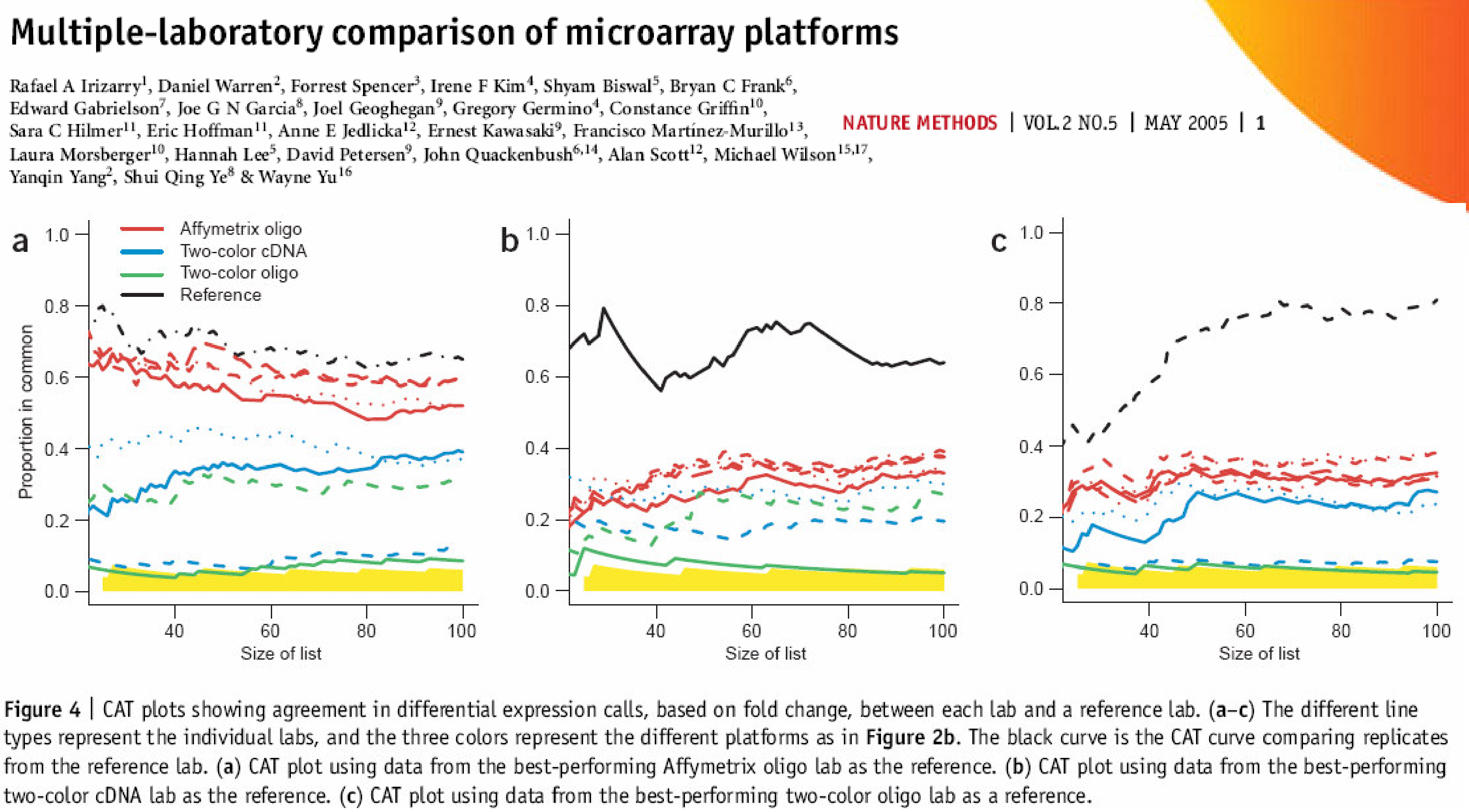
$X = PM_1 - MM_1$, $Y = PM_2 - MM_2$	



x-axis (mean log2 intensity): average intensity of a particular element across the control and experimental conditions.

y-axis (ratio): ratio of the two intensities.

CATplot



Multiple-laboratory comparison of microarray platforms

Rafael A Irizarry[1], Daniel Warren[2], Forrest Spencer[3], Irene F Kim[4], Shyam Biswal[5], Bryan C Frank[6], Edward Gabrielson[7], Joe G N Garcia[8], Joel Geoghegan[9], Gregory Germino[4], Constance Griffin[10], Sara C Hilmer[11], Eric Hoffman[11], Anne E Jedlicka[12], Ernest Kawasaki[9], Francisco Martínez-Murillo[13], Laura Morsberger[10], Hannah Lee[5], David Petersen[9], John Quackenbush[6,14], Alan Scott[12], Michael Wilson[15,17], Yanqin Yang[2], Shui Qing Ye[8] & Wayne Yu[16]

NATURE METHODS | VOL.2 NO.5 | MAY 2005 | 1

Figure 4 | CAT plots showing agreement in differential expression calls, based on fold change, between each lab and a reference lab. (**a–c**) The different line types represent the individual labs, and the three colors represent the different platforms as in **Figure 2b**. The black curve is the CAT curve comparing replicates from the reference lab. (**a**) CAT plot using data from the best-performing Affymetrix oligo lab as the reference. (**b**) CAT plot using data from the best-performing two-color cDNA lab as the reference. (**c**) CAT plot using data from the best-performing two-color oligo lab as a reference.

- Correspondence at the top plot
- Nick Cox’s catplot (Cox 2004): Stata module for plots of frequencies, fractions or percents of categorical data.
- SAS: The catplot macro is designed to plot observed and/or predicted values for logit models fit by the CATMOD procedure.

Normalization

Assume Microarray Image has been processed appropriately.

Ensure that the data is of high quality and suitable for analysis.

- Removing Flagged Features
- Background Subtraction
- Taking Logarithm

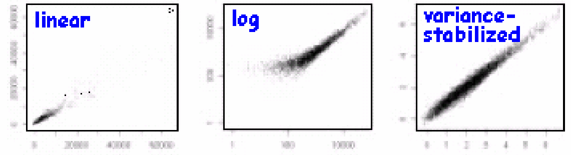
Quantification of Expression

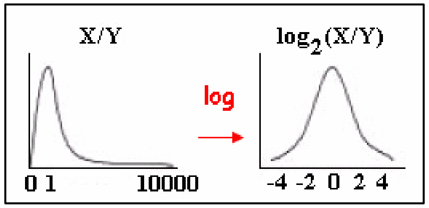
Red intensity (Cy5) = Rfg - Rbg
Green intensity (Cy3) = Gfg - Gbg

→ log2 ratio
= Log2(Cy5/Cy3)

- **What is normalization?**
  - Non-biological factor can contribute to the variability of data, in order to reliably compare data from multiple probe arrays, differences of non-biologocal origin must be minimized.
  - Normalization is a process of reducing unwanted variation across chips. It may use information from multiple chips.
- **Why normalization?**
  - Normalization corrects for overall chip brightness and other factors that may influence the numerical value of expression intensity, enabling the user to more confidently compare gene expression estimates between samples.

cDNA microarray red vs green intensities





Main idea of the normalization

Remove the systematic bias in the data as comp letely possible while preserving the variation in the gene expression that occurs because of biologically relevant changes in transcription.

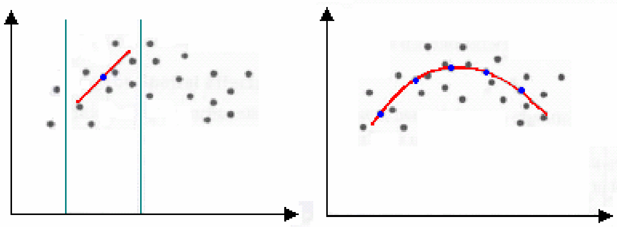
Basic Assumption

1. The average gene does not change in its expression level in the biological sample being tested.
2. Most genes are not differentially expressed or up- and down-regulated genes roughly cancel out the expression effect.

Normalization Methods: loess

- Loess normalization (Bolstad *et al*., 2003) is based on **MA plots**. Two arrays are normalized by using a lowess smoother.
- **Skewing** reflects experimental artifacts such as the
  - contamination of one RNA source with genomic DNA or rRNA,
  - the use of unequal amounts of radioactive or fluorescent probes on the microarray.
- Skewing can be corrected with local normalization: fitting a local regression curve to the data.

Loess regression
(locally weighted polynomial regression)



1. For any two arrays i, j with probe intensities x_{ki} and x_{kj} where $k = 1, \ldots, p$ represents the probe
2. we calculate $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \frac{1}{2}\log_2(x_{ki}x_{kj})$.
3. A normalization curve is fitted to this M versus A plot using loess.

   Loess is a method of local regression (see Cleveland and Devlin (1988) for details).
4. The fits based on the normalization curve are $\hat{M}_k$
5. the normalization adjustment is $M'_k = M_k - \hat{M}_k$.
6. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k+\frac{M'_K}{2}}$ and $x'_{kj} = 2^{A_K-\frac{M'_k}{2}}$.

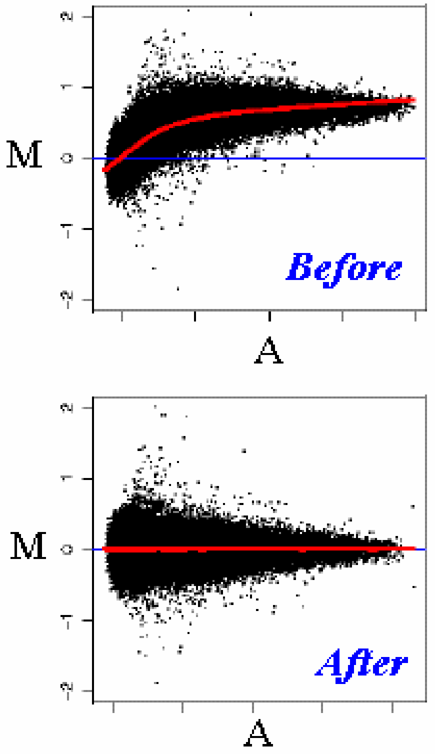
$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2}\log_2(X\,Y)$$

Oligo	cDNA

$X = PM_1$,
$Y = PM_2$
	X= Cy3
	Y= Cy5

$X = PM_1 - MM_1$, $Y = PM_2 - MM_2$	



Normalization Methods

Within-array Normalization

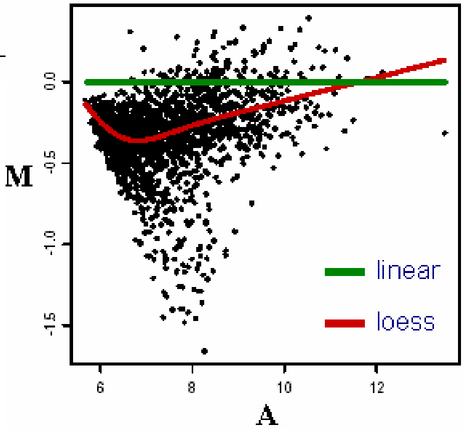
	Subject be used for estimating normalization curve			
Location Normalization				
Method	allGenes	Print-tip i	2D Location (x, y)	SelectedGenes (Controls, Housekeeping, MSP, Invariant set)
constant	global normalization $N = M - c$ c : mean, median	print-tip normalization $N = M - c_i$		
loess (Robust scatterplot smoother: loess, spline,...)	global loess normalization $N = M - c(A)$ c : loess curve	print-tip loess normalization $N = M - c_i(A)$	2D loess normalization $N = M - c(x, y) - c(A)$	$N = M - p_A c_{\text{MSP}}(A) - (1 - p_A) c_i(A)$
Scale Normalization				
MAD	global scale normalization $N = s \times M$ $s = 1/mad(A)$	print-tip scale normalization $N = s_i \times M$ $s_i = 1/mad_i(A)$		
STD	standardization $N = M - ave(M)/std(A)$			

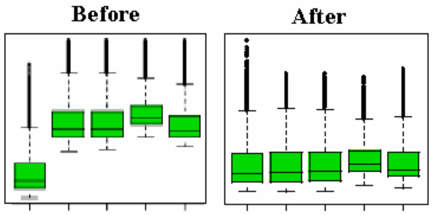
Smyth and Speed (2003)

Between-array Normalization

Scale-normalization: scaling of the M-values from a series of arrays so that each array has the same $MAD = \text{median}|M - \text{median}(M)|$

Paired-array Normalization (Dye-swap)





Reference for Normalization Methods

- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. (2002), "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," Nucleic Acids Res 2002 Feb 15;30(4)
- Christopher Workman, Lars Juh Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Henrik Bjørn Nielser, Hans-Henrik Saxild, Claus Nielsen, Søren Brunak, Steen Knudsen, A new non-linear normalization method for reducing variability in DNA microarray experiments, Genome Biology 2002 3(9): research0048.1-0048.16
- Colantuoni C., Zeger S., Pevsner J., "SNOMAD (Standardization and Normalization of Microarray Data): web accessible gene expression data analysis," Bioinformatics, in press,
- Wang Y, Lu J, Lee R, Gu Z, Clarke R. (2002), "Iterative normalization of cDNA microarray data," IEEE Trans Inf Technol Biomed 2002 Mar;6(1):29-37
- Bilban M, Buehler LK, Head S, Desoye G, Quaranta V. (2002), "Normalizing DNA microarray data," Curr Issues Mol Biol 2002 Apr;4(2):57-64
- Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. (2001), "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," Nucleic Acids Res 2001 Jun 15;29(12):2549-57
- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzel, H. (2000), "Normalization strategies for CDNA microarrays," Nucleic Acids Res, 28, E47.
- Thomas B. Kepler, Lynn Crosby, and Kevin T. Morgan (2000), "Normalization and Analysis of DNA Microarray Data by Self-Consistency and Local Regression", Santa Fe Institute.
- Eickhoff, B., Korn, B., Schick, M., Poustka, A. and van der Bosch, J. (1999), "Normalization of Array Hybridization Experiments in differential gene expression analysis," Nucleic Acids Res, 27, e33.
- Yue Wang, Jianping Lu, Richard Lee, Zhiping Gu, and Robert Clarke, Iterative Normalization of cDNA Microarray Data
- Sanchez-Cabo F, Cho KH, Butcher P, Hinds J, Trajanoski Z, Wolkenhauer O. Is LOWESS a Panacea in the Normalization of Microarray Data? Applied Bioinformatics. 2003
- Quackenbush, J. (2002). Microarray data normalization and transformation. Nat Genet 32 Suppl, 496-501.
- TC Kroll, S W. (2002), "Ranking: a closer look on globalisation methods for normalisation of gene expression arrays", Nucleic Acids Research, 30(11):e50.
- Ding, Y., Wilkins, D. (2004). The Effect of Normalization on Microarray Data Analysis, DNA and Cell Biology, 22:10, 635-642.

Finding Differentially Expressed Genes

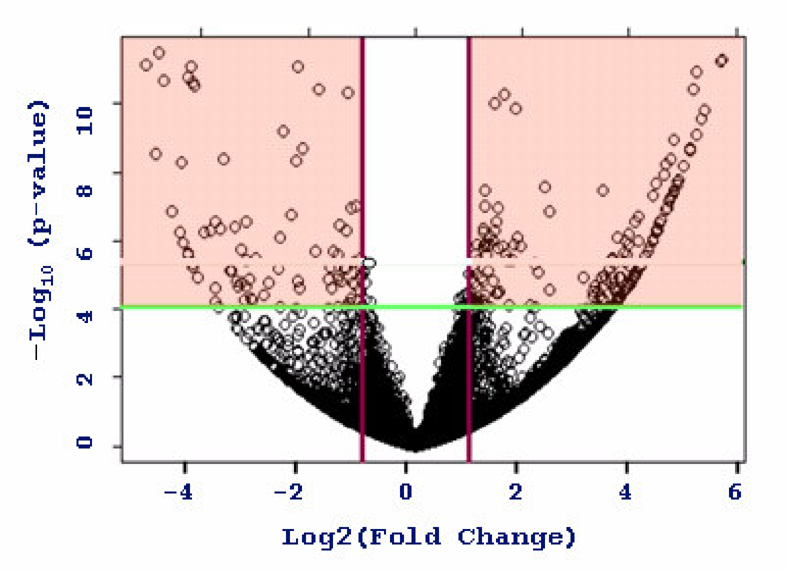
- Select a statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence.

(Primary Importance): only a limited number of genes can be followed up in a typical biological study.

- Choose a critical-value for the ranking statistic above which any value is considered to be significant.

Volcano Plot

For a volcano plot the Y variate is typically a probability (in which case a -log10 transform is used) or less commonly a p-value. The X variate is usually a measure of differential expression such as a log-ratio.



Fold-Change Method

Calculate the expression ratio in control and experimental cases and to rank order the genes. Chose a threshold, for example at least 2-fold up or down regulation, and selected those genes whose average differential expression is greater than that threshold.

Problems: it is an arbitrary threshold.

- In some experiments, no genes (or few gene) will meet this criterion.
- In other experiments, thousands of genes regulated.
- bg=100, s1=300, s2=200. => subtract bg => s1=200, s2=100 ==> 2-fold.
  (s2 close to bg, the difference could represent noise. It is more credible that a gene is regulated 2-fold with 10000, 5000 units)
- The average fold ratio does not take into account the extent to which the measurements of differential gene expression vary between the individuals being studied.
- The average fold ratio does not take into account the number of patients in the study, which statisticians refer to as the sample size.

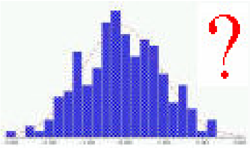
Define which genes are significantly regulated might be to choose 5% of genes that have the largest expression ratios.

Problems:

- It applies no measure of the extent to which a gene has a different mean expression level in the control and experimental groups.
- Possible that no genes in an experiment have statistically significantly different gene expression.

Hypothesis Testing

Decide which genes are significantly regulated in a microarray experiment.

Microarray Data	Paired data	Unpaired data	Complex data *More than two Groups*
Parametric Hypothesis Testing	■ z-test **Dependent samples** ■ *t-test*	■ *two-sample* **Independent samples** *t-test*	■ One-Way Analysis of Variance (ANOVA)
	Assumptions and Test for Normality ■ Histogram, QQplot ■ Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test		
Non-Parametric Hypothesis Testing	■ Sign test, ■ Wilcoxon signed-rank test	■ Wilcoxon rank-sum test, (Mann-Whitney U test).	

Bootstrap Analysis, Permutation Test

This Topic will be covered in PART-II.

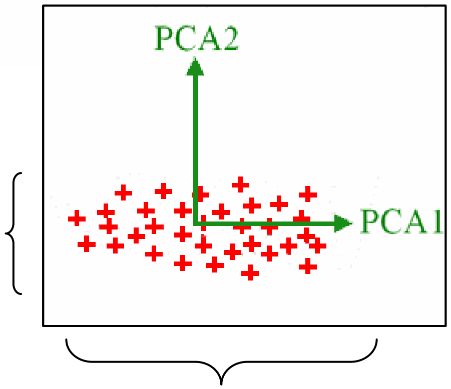
Exploratory Visualization Methods

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

Principal Component Analysis

(Pearson 1901; Hotelling 1933; Jolliffe 2002)

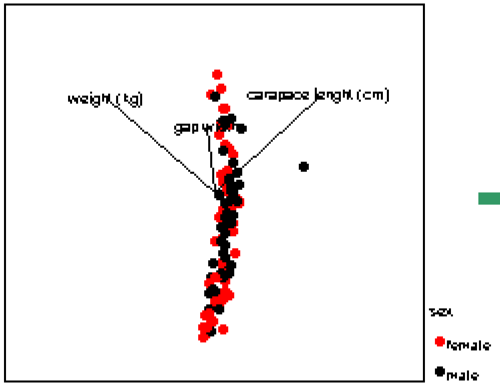
PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).

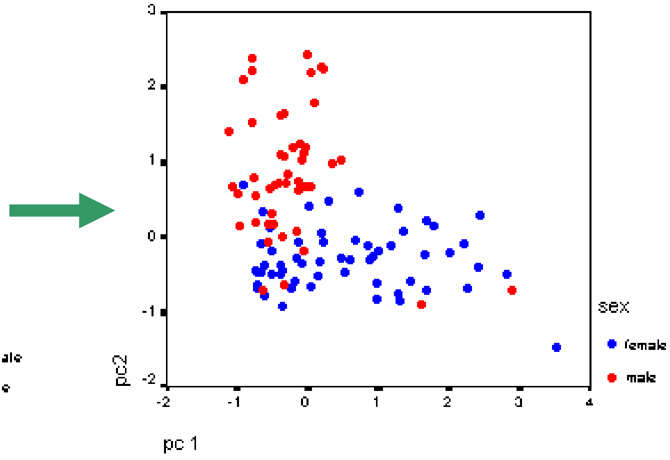


$$\mathrm{PCA}_1 = \boxed{a_1}\mathbf{X} + \boxed{b_1}\mathbf{Y}$$

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

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





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

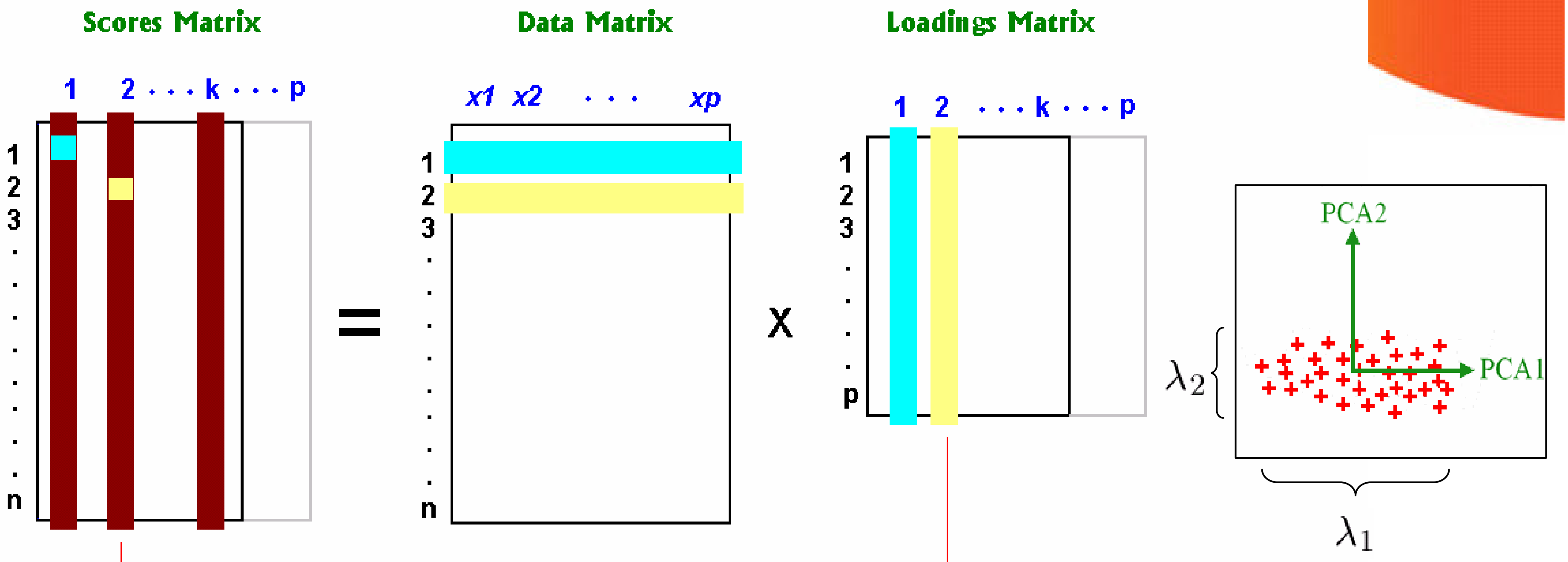
$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$

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.

$$\boxed{\begin{aligned}\mathrm{PCA}_1 &= a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p \\ \mathrm{PCA}_2 &= a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p\end{aligned}}$$

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

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

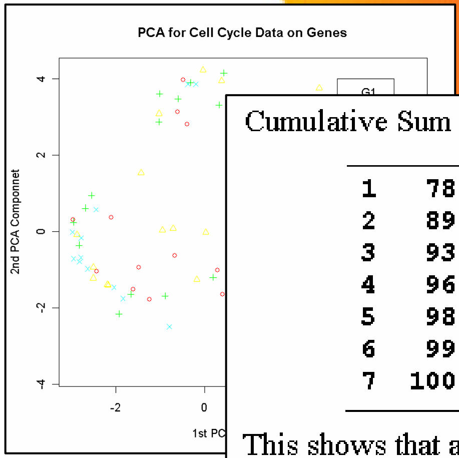
PCA (conti.)

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



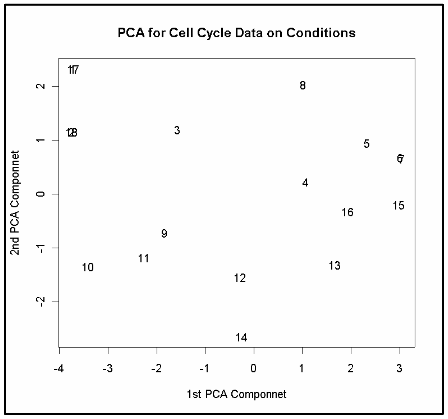
Cumulative Sum of the Variances:

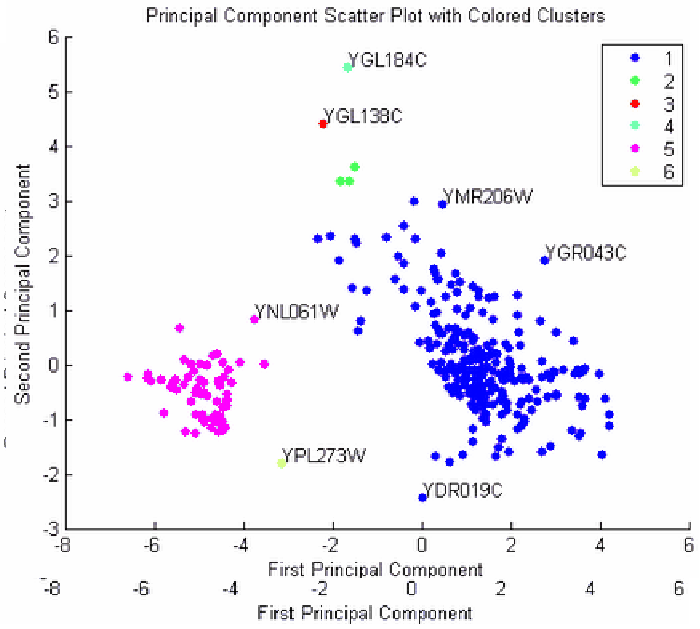
1	78.3719
2	89.2140
3	93.4357
4	96.0831
5	98.3283
6	99.3203
7	100.0000

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

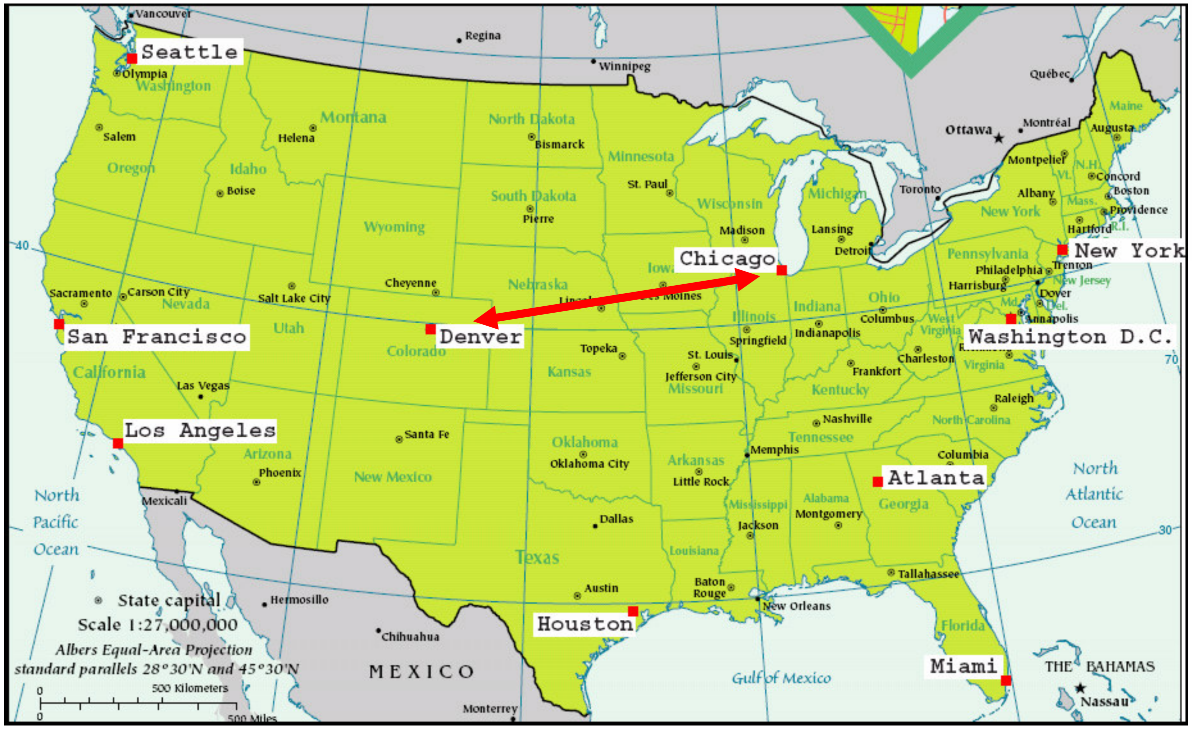




Yeast Microarray Data is from
DeRisi, JL, Iyer, VR, and Brown, PO.(1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale"; Science, Oct 24;278(5338):680-6.

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)



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

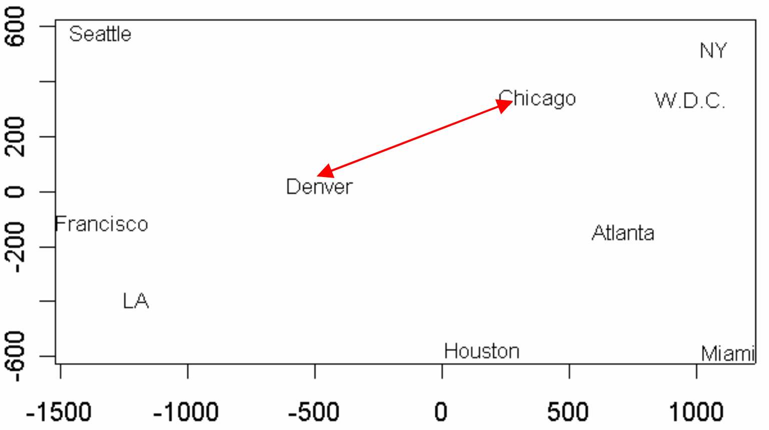
?

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

- 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.



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?

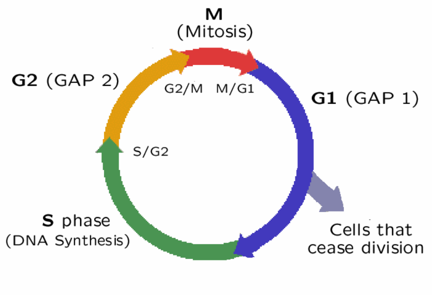
$$S = \sum_{i,j} (\hat{d}_{ij} - d_{ij})^2$$

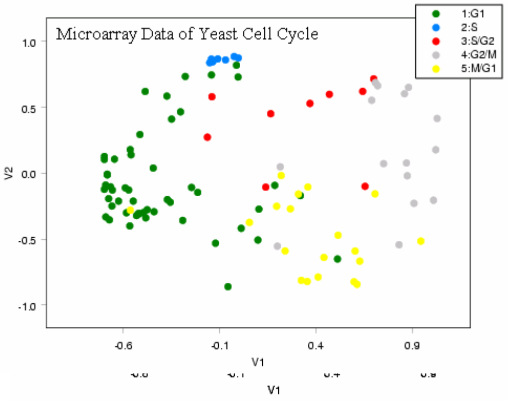
Mathematically: for given ***k***, compute points $\boldsymbol{x_1,\ldots,x_n}$ in ***k***-dimensional space such that the object function is minimized.

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

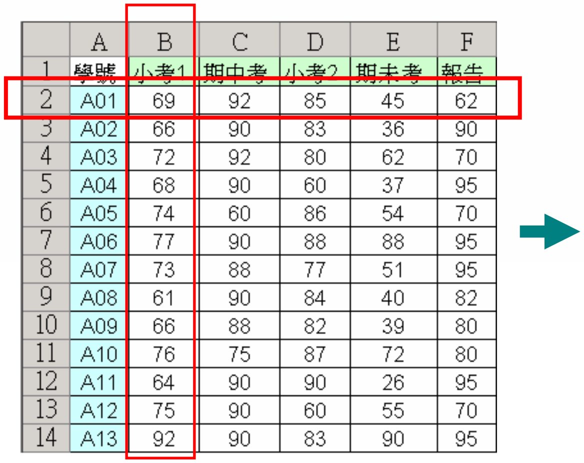
Microarray Data of Yeast Cell Cycle

- **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.**

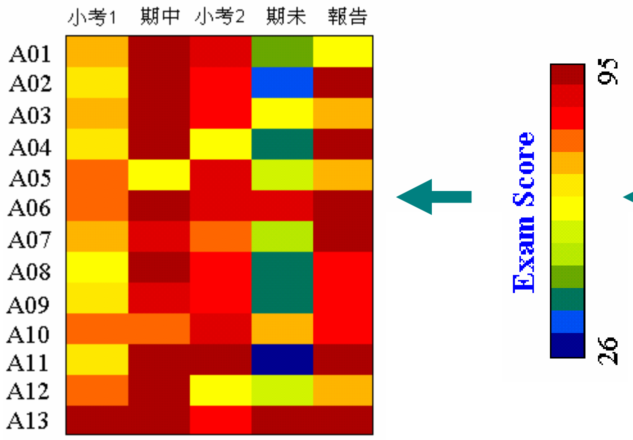




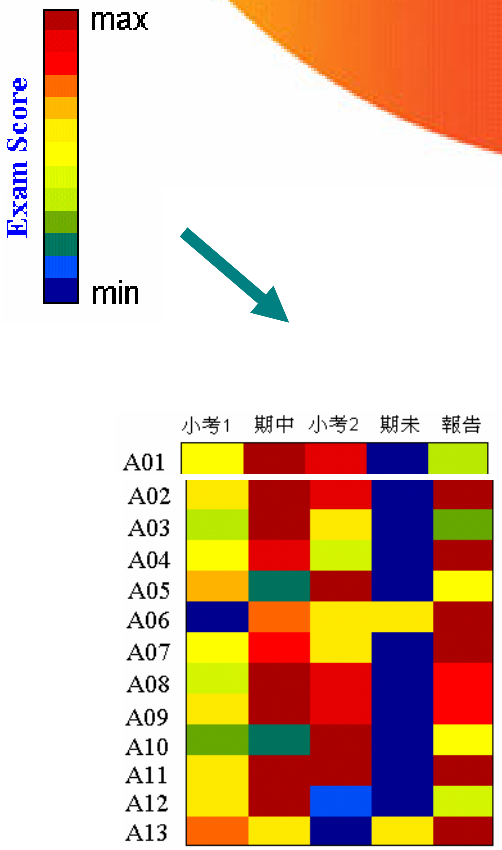
Heat Map



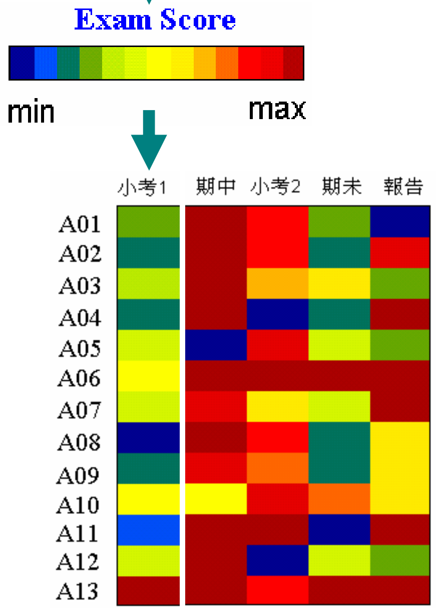
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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



Range Matrix Condition

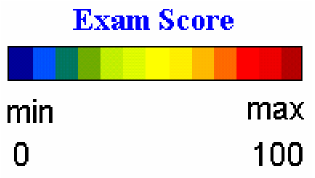


Range Raw Condition

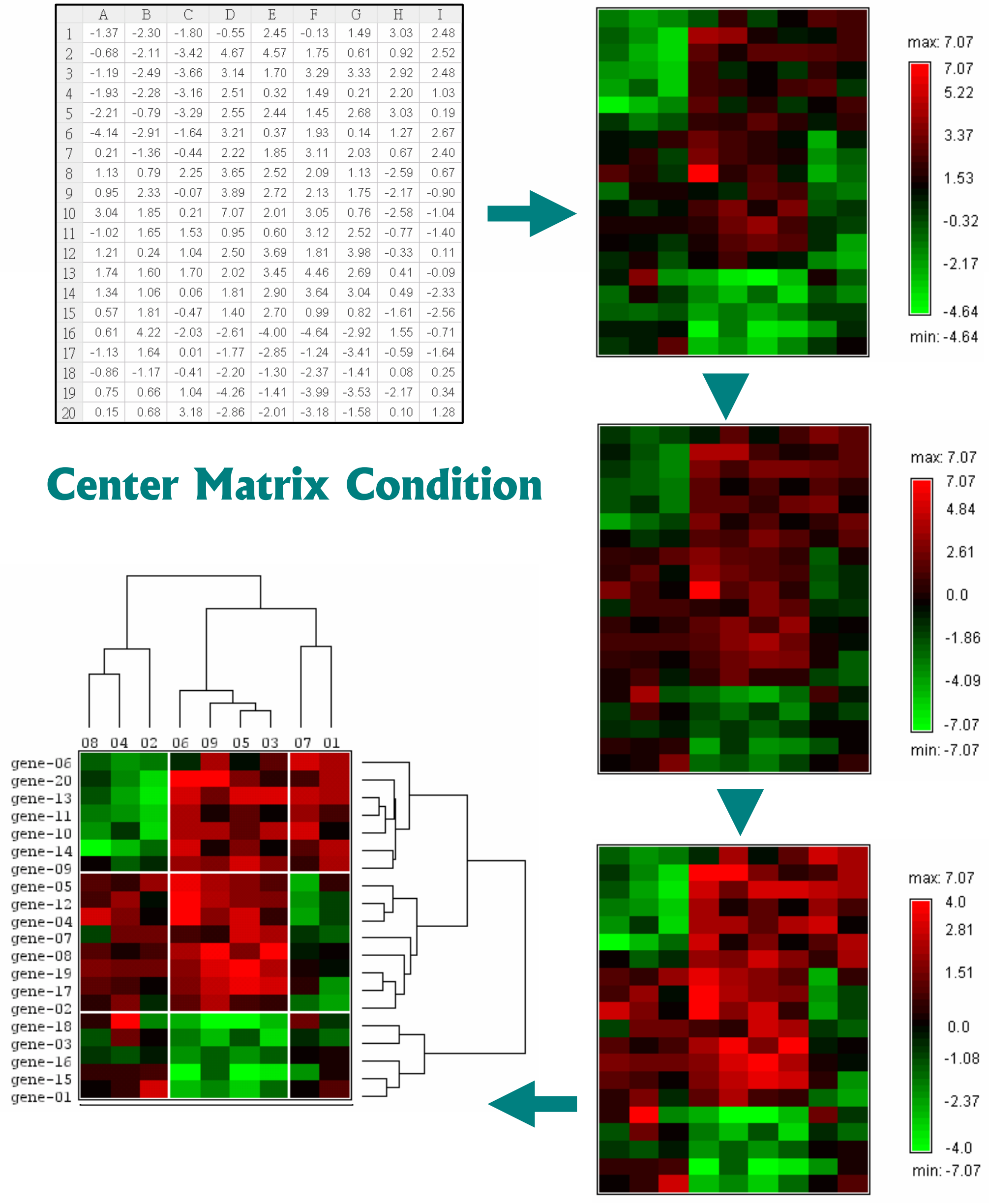


Range Column Condition

What about this one?



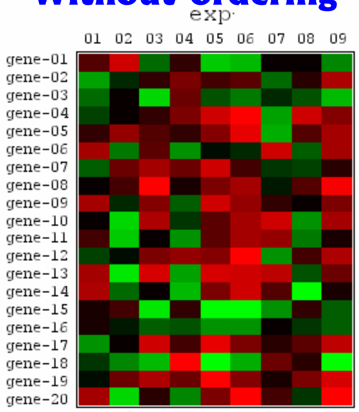
Heat Map (conti.)



	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

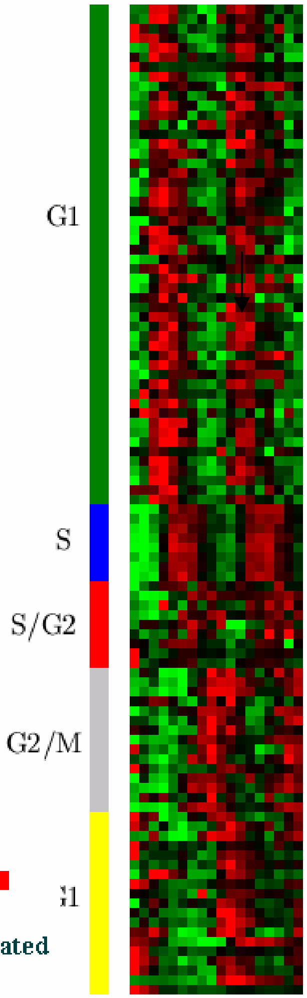
Center Matrix Condition

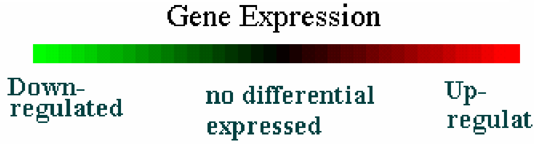
Without ordering



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.





Analysis of Relationship Between Genes, Tissues or Treatments

- Hierarchical Clustering, K-Means Clustering
- Self-Organizing Maps (SOM)
- How Many Clusters?

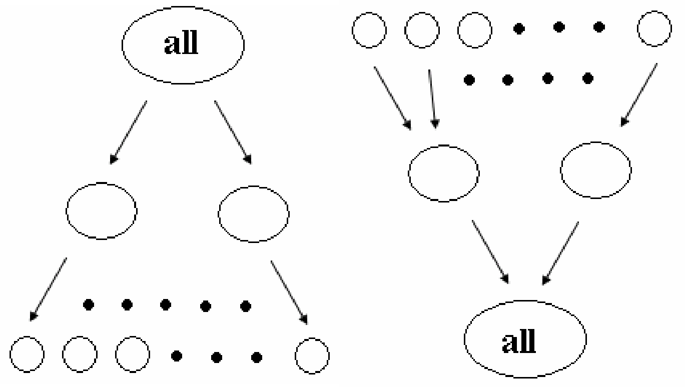
Clustering Analysis (Unsupervised Learning)

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 can be perform using agglomerative and divisive approaches. The result 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.



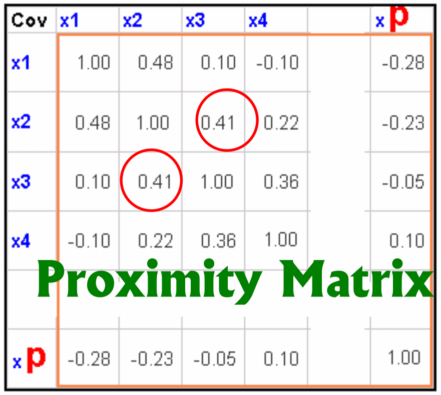
Non-Hierarchical clustering

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

- Only use similarities of instances, without any other requirement of the data.
- all attribute have the same scale.

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		xp
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
xp	-0.28	-0.23	-0.05	0.10		1.00

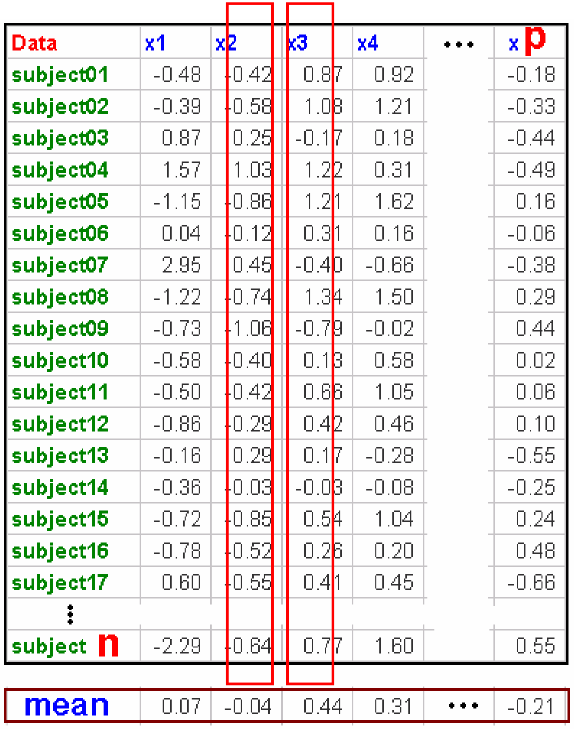
Proximity Matrix

Pearson Correlation Coefficient

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

Data Matrix

x y →

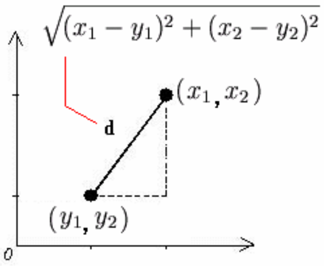


Data	x1	x2	x3	x4	•••	xp
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-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

$$\boxed{\begin{array}{l} x = (x_1, x_2, \cdots, x_n) \\ \\ y = (y_1, y_2, \cdots, y_n) \end{array}}$$

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)

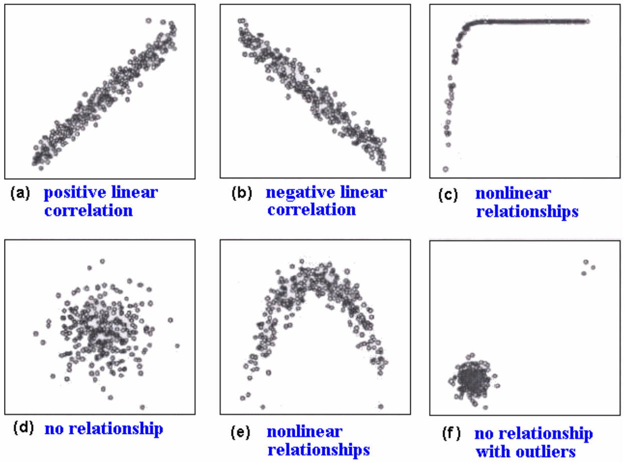
- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

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	$s(i,j) = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\mathrm{var}(x_i)\,\mathrm{var}(x_j)}}$
Spearman correlation (r_i is ranked x_i)	$s(i,j) = \frac{\mathrm{cov}(r_i, r_j)}{\sqrt{\mathrm{var}(r_i)\,\mathrm{var}(r_j)}}$
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

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

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

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

Hierarchical Clustering and Dendrogram (Kaufman and Rousseeuw, 1990)

UPGMC (Unweighted Pair-Groups Method Centroid)

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

↓

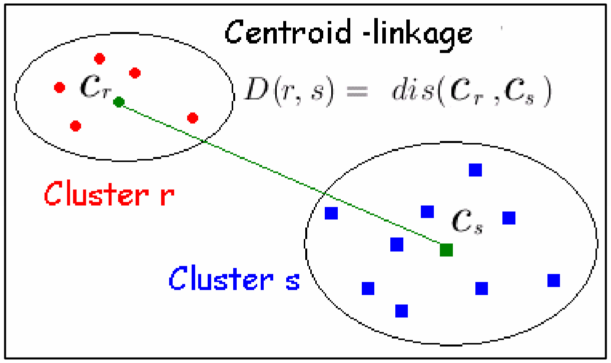
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

↓

	{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


Centroid -linkage

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

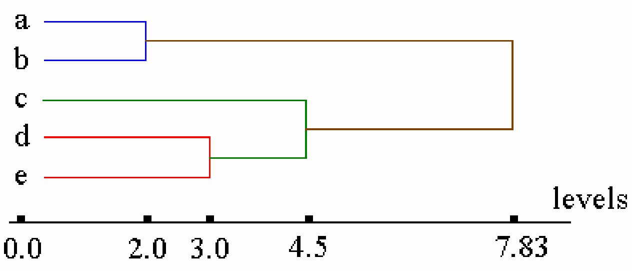
$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)]$$

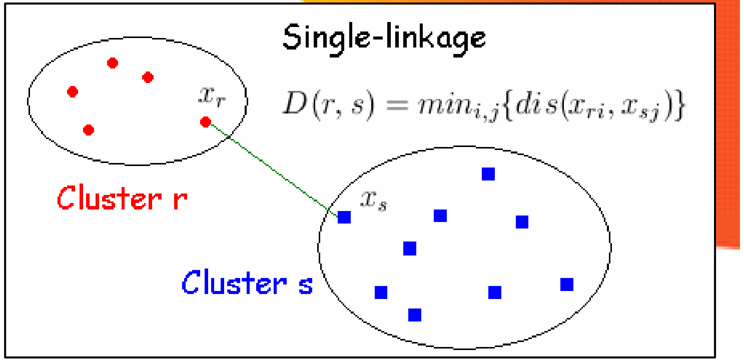
$$= \frac{1}{2}(6+5) = 5.5$$

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

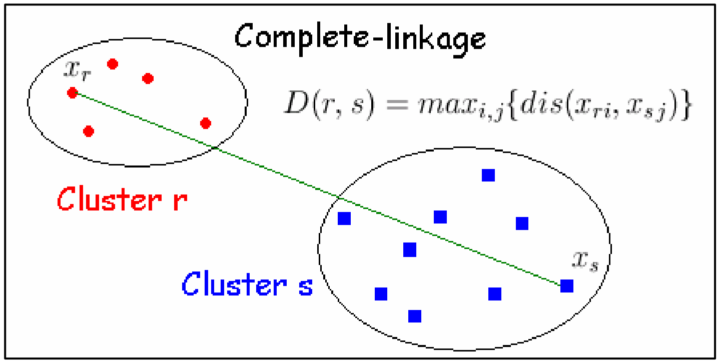
$$= \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$

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



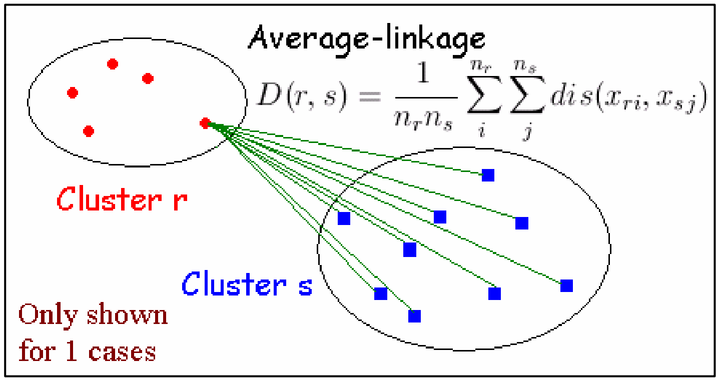

Single-linkage

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


Complete-linkage

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

UPGMA (Unweighted Pair-Groups Method Average)


Average-linkage

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

Only shown for 1 cases

Ward's Method

- The Ward’s method does not compute distances between clusters.
- It forms clusters by maximizing within-clusters homogeneity.
- The within-group (i.e., within-cluster) sum of squares is used as the measure of homogeneity.
- The Ward’s method tries to minimize the total within-group or within-cluster sum of squares.
- Clusters are formed at each step such that the resulting cluster solution has the fewest within-cluster sums of squares.
- The within-cluster sums of squares that is minimized is also known as the error sums of squares (ESS).

Example:

Charles H. Romesburg (1984)

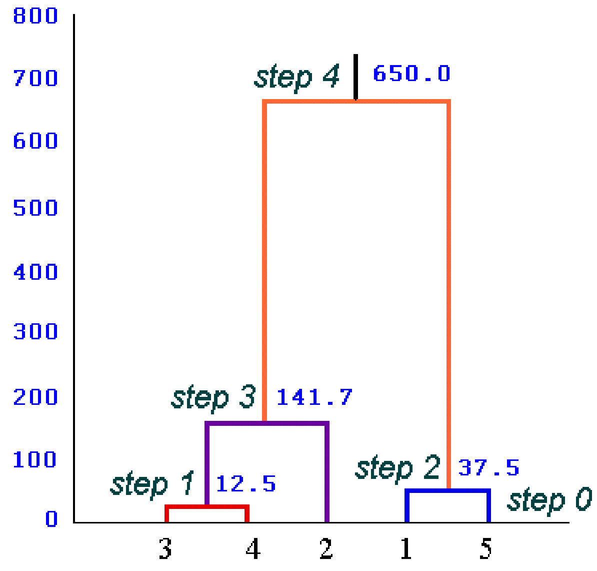
Toy Data

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

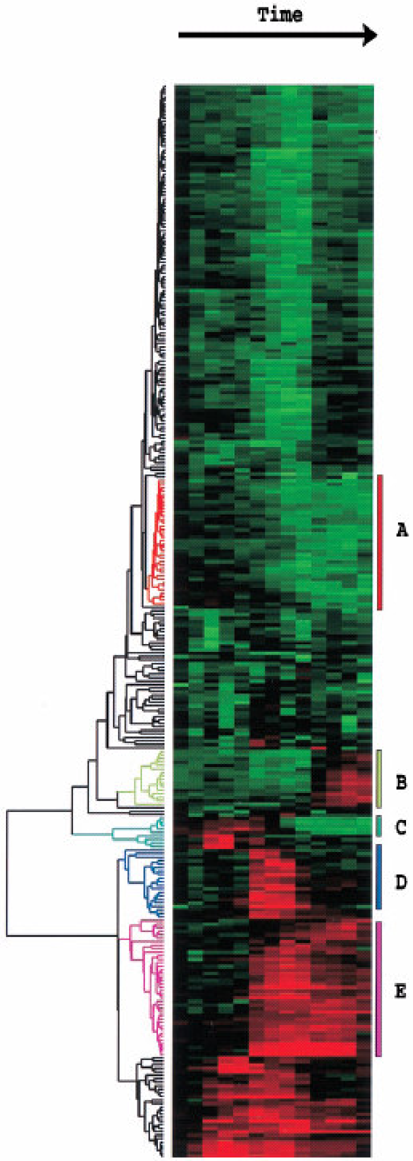
step	Possible Partitions				ESS
1	(12)	3	4	5	?

$$\{\bar{12}\} = [\ (10+20)/2, (5+20)/2\ ]$$
$$= [\ 15, 12.5]$$
$$ESS = wss\{12\} + wss\{3\} + wss\{4\} + wss\{5\}$$
$$= ss(1, \{\bar{12}\}) + ss(2, \{\bar{12}\})$$
$$= (10-15)^2 + (5-12.5)^2 + (20-15)^2 + (2-12.5)^2$$
$$= 162.5$$

step	Possible Partiti				
1	(12)	3	4	5	162.5
	(13)	2	4	5	212.5
	(14)	2	3	5	250.0
	(15)	2	3	4	25.0
	(23)	1	4	5	100.0
	(24)	1	3	5	62.5
	(25)	1	3	4	162.5
	(34)	1	2	5	12.5*
	(35)	1	2	4	312.5
	(45)	1	2	3	325.0
2	(34)	(12)	5		175.0
	(34)	(15)	2		37.5*
	(34)	(25)	1		175.0
	(134)	2	5		316.7
	(234)	1	5		116.7
	(345)	1	2		433.3
3	(234)	(15)			141.7*
	(125)	(34)			245.9
	(1345)	2			568.8
4	(12345)				650.0



Hierarchical Clustering



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

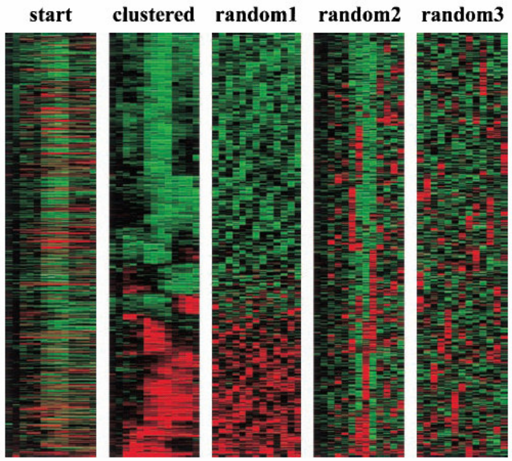
Cluster analysis and display of genome-wide expression patterns

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.

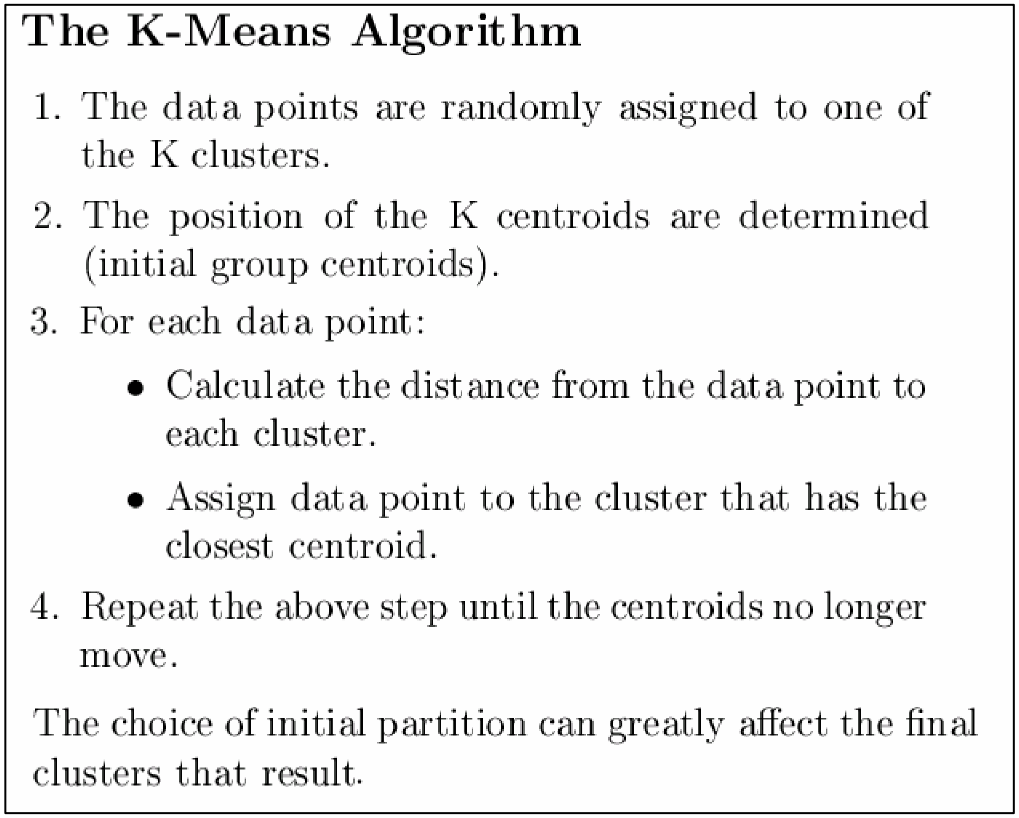
Software:
Cluster and TreeView

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.



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.

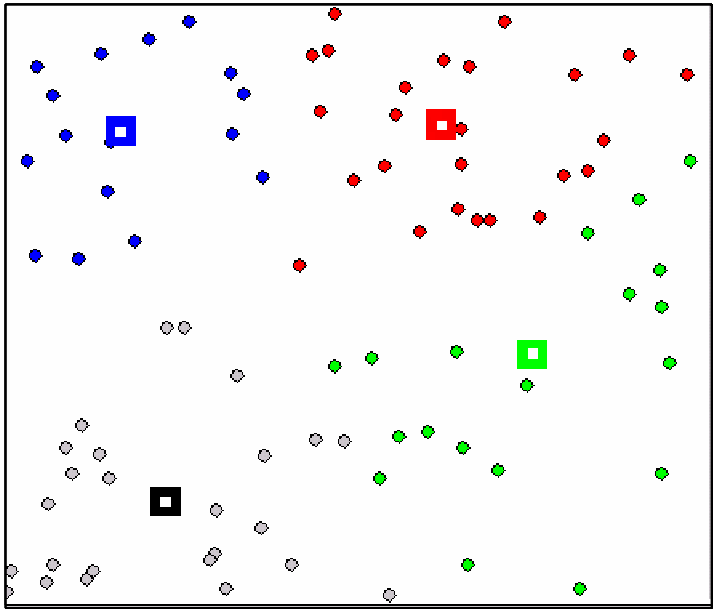
The choice of initial partition can greatly affect the final clusters that result.

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

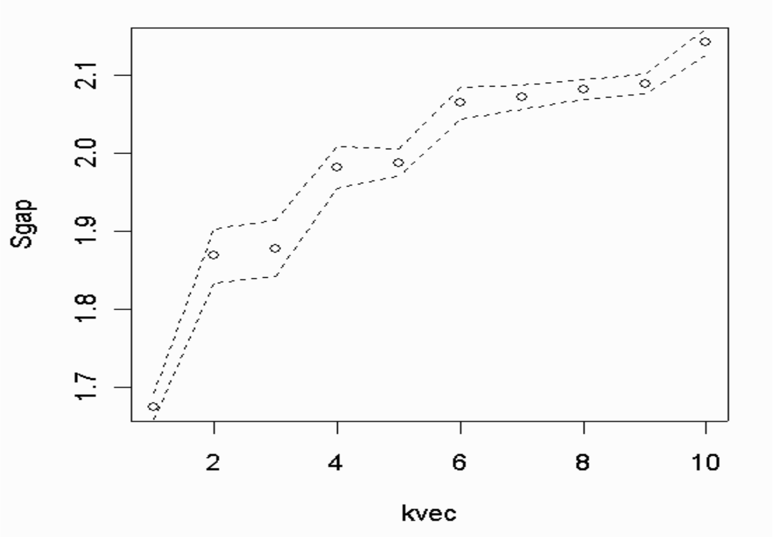


K-Means Clustering

- 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.

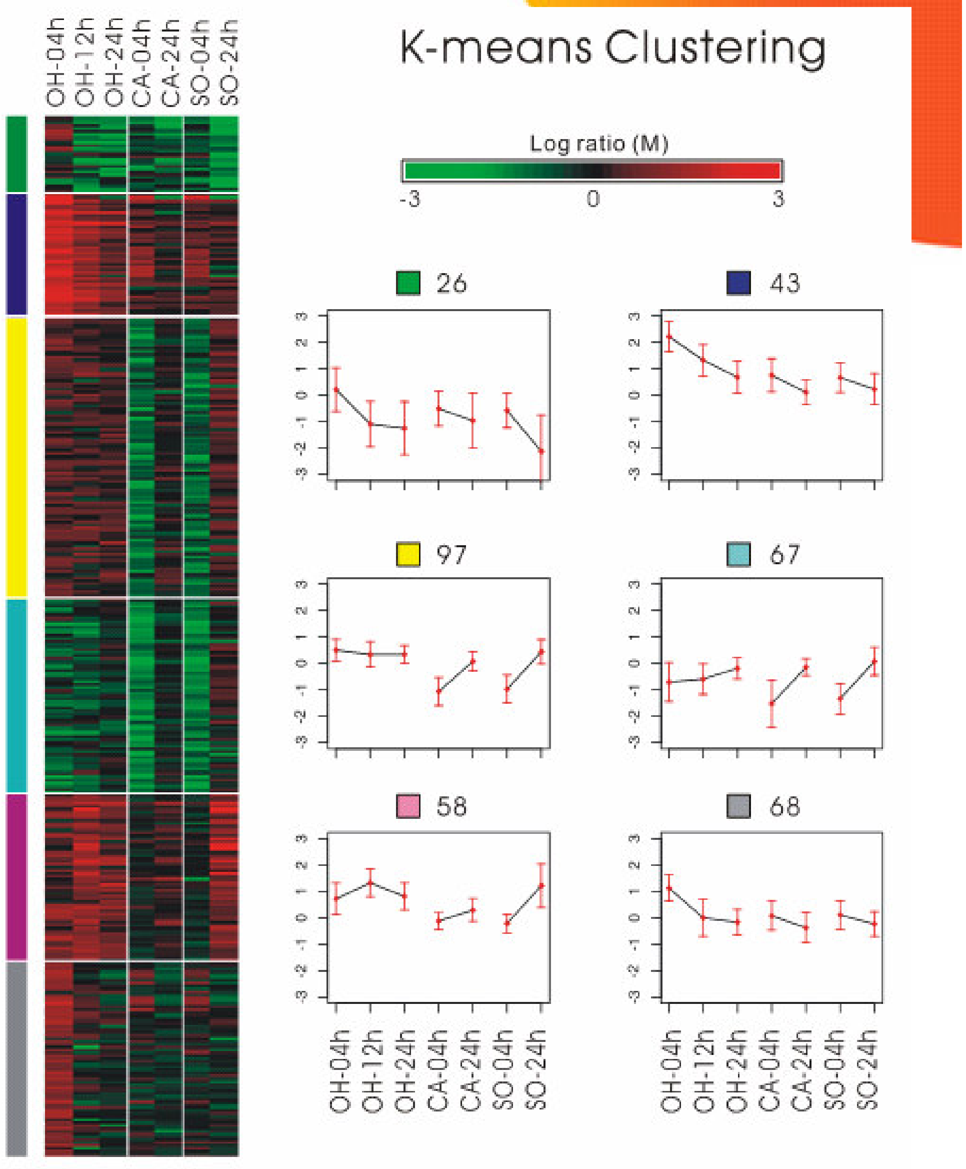


J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

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

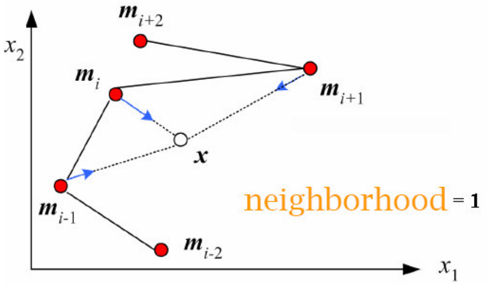
Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA



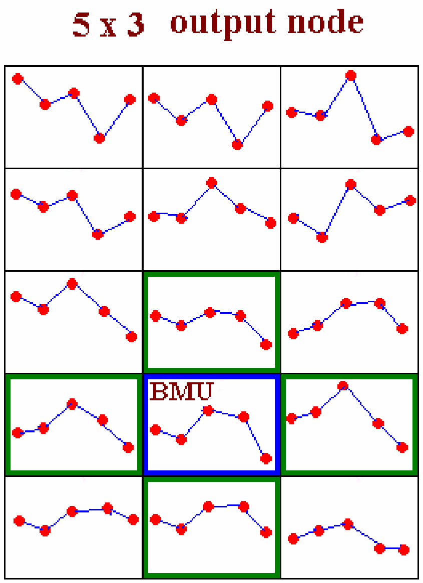
Self-Organizing Maps (SOM)

- ■ 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 $h_{ci}(t)$.

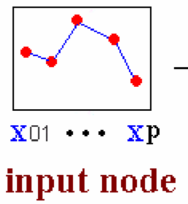


Learning process:

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

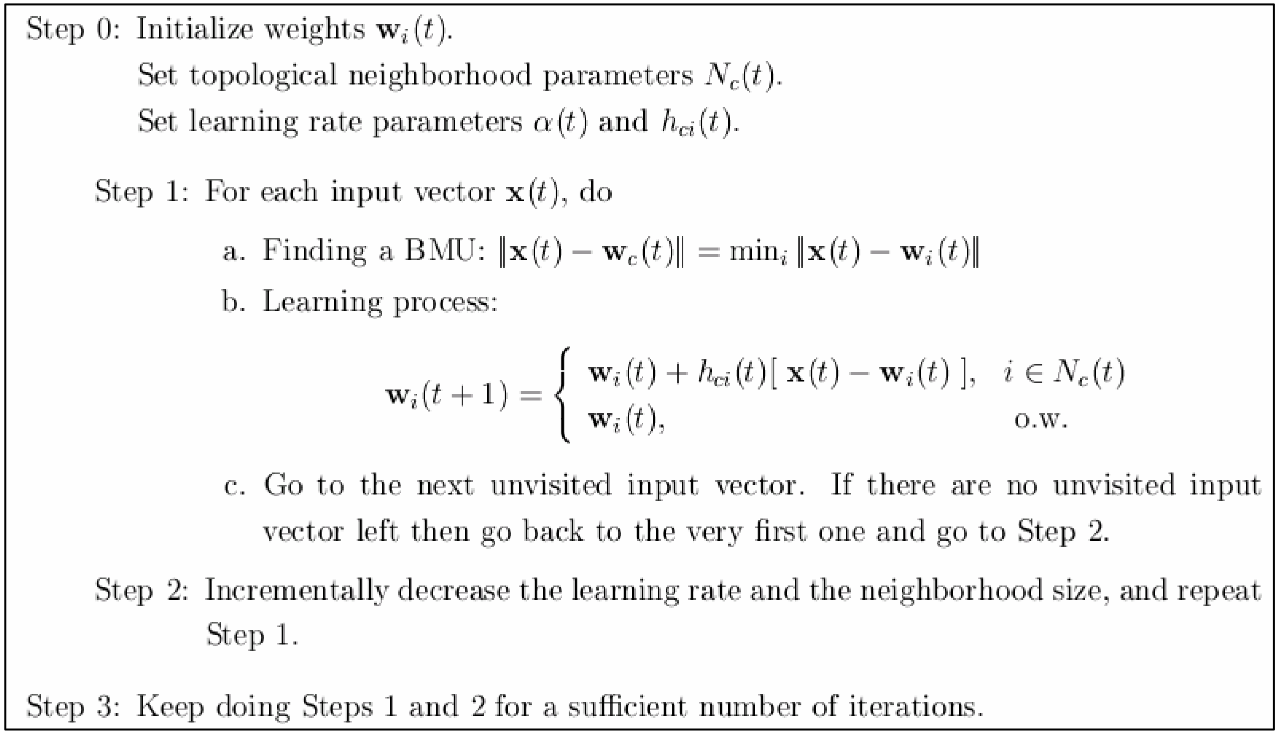
Data Matrix

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

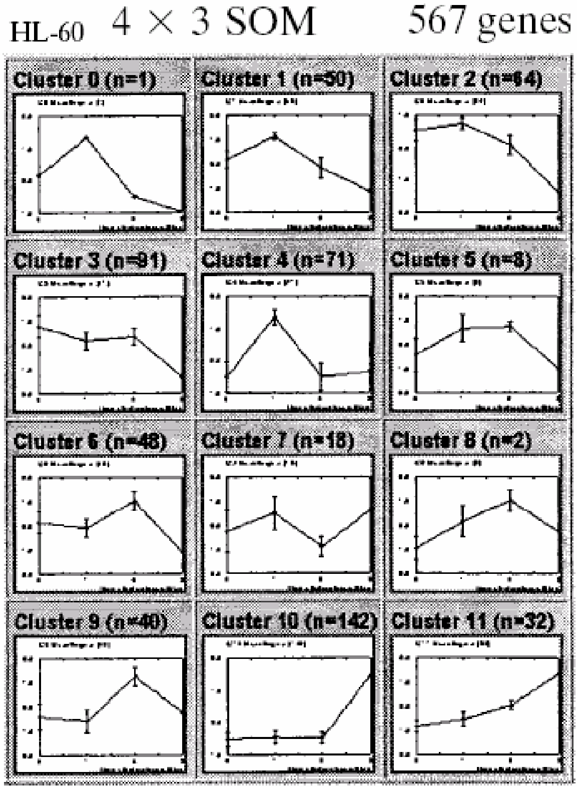


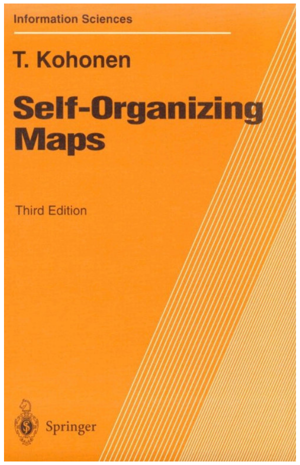
Incrementally decrease the learning rate and the neighborhood size, and repeat

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.
```





Information Sciences

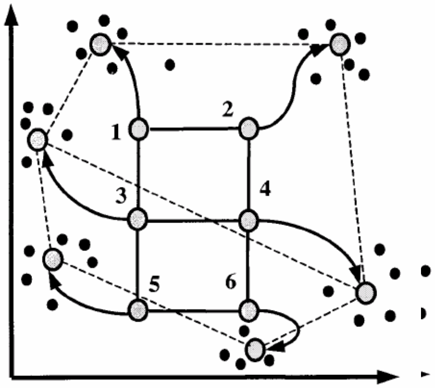
T. Kohonen

Self-Organizing Maps

Third Edition

Springer

1995, 1997, 2001



Iacrophage Differentiation in HL-60 cells

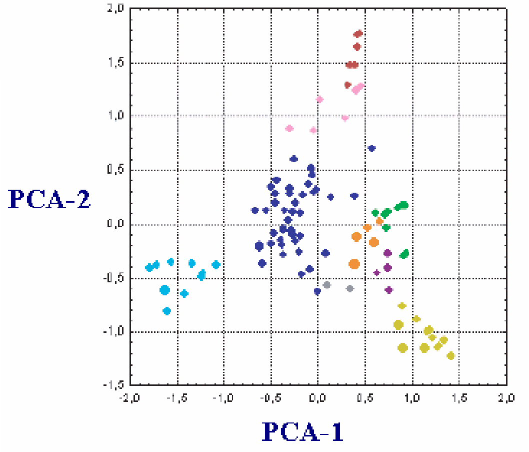
Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96:2907-2912.

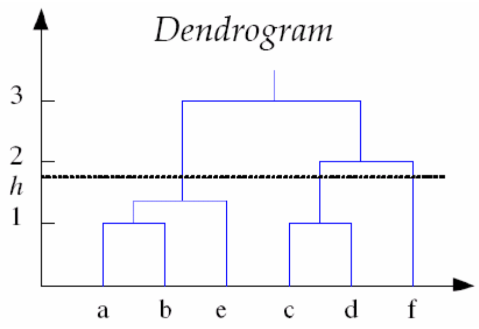
Choosing the Number of Clusters

(1) K is defined by the application.

(2) Plot the data in two PAC dimensions.

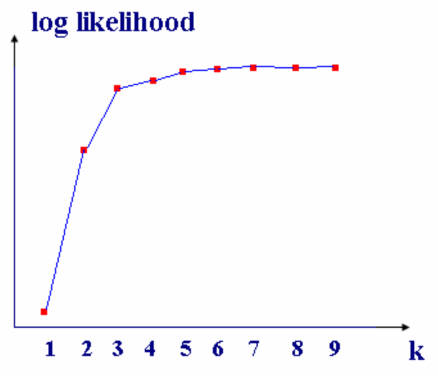
(4) Hierarchical clustering: look at the difference between levels in the tree.

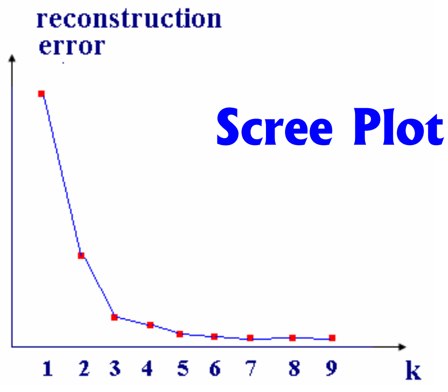




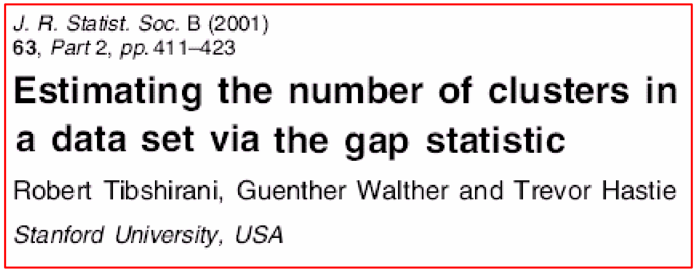
(e.g., k-means: within-cluster sum of squares)

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





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

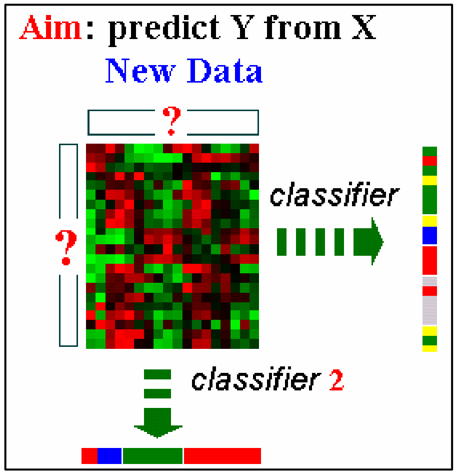
Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA

Classification of Genes, Tissues or Samples

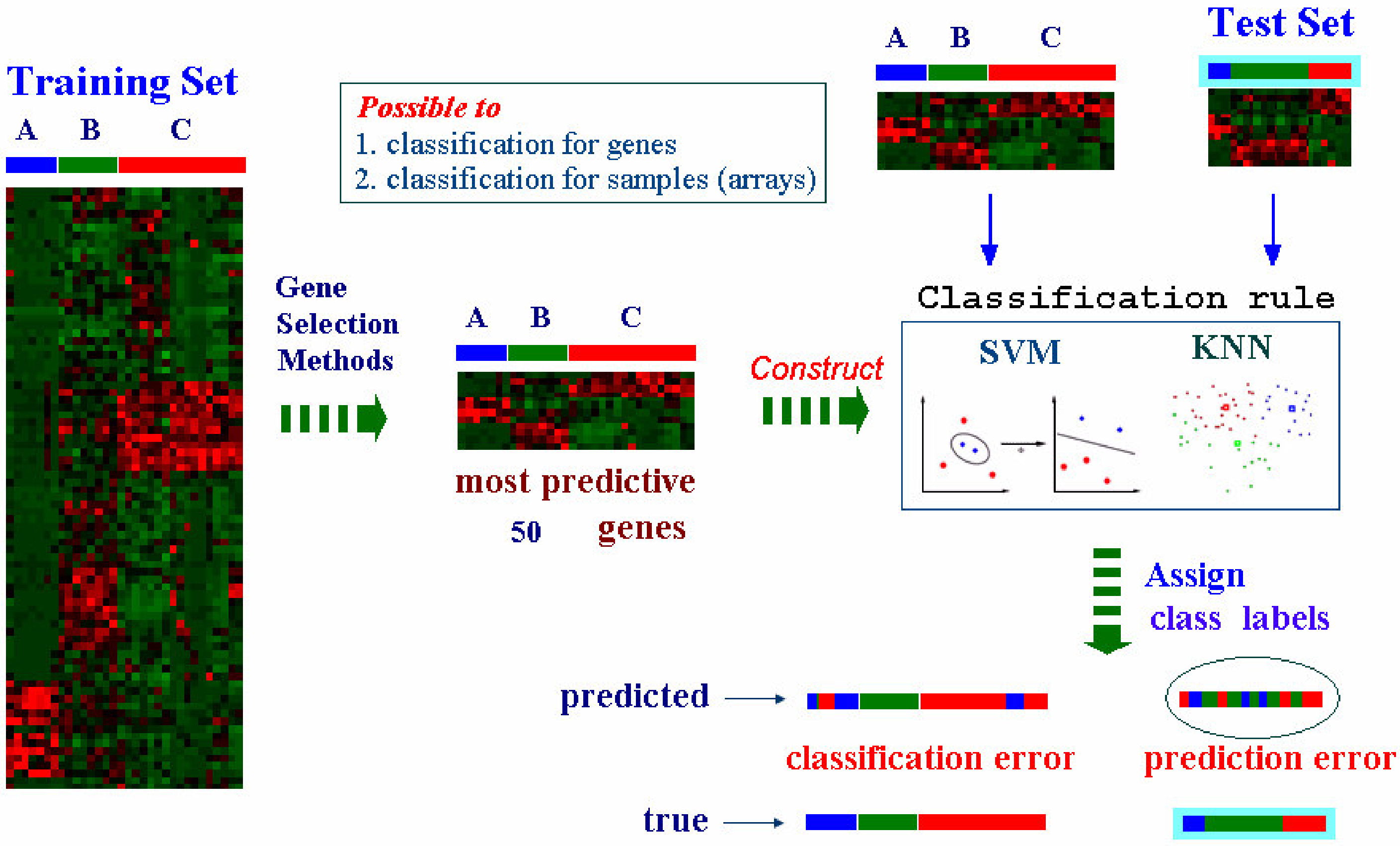
- Linear Discriminant Analysis (LDA)
- Support Vector Machines (SVM)

Classification of Genes, Tissues or Samples (Supervised Learning)



Aim: predict Y from X

New Data



Training Set

Possible to

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

Test Set

Classification rule

Assign class labels

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.

$$\begin{array}{cc} \text{Genes (variables)} & \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} & \begin{array}{c} \text{mRNA} \\ \text{samples} \\ \text{(observations)} \end{array} \end{array}$$

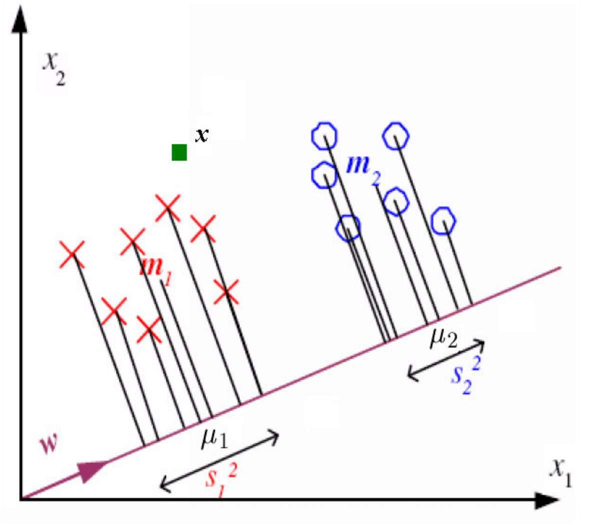
$X_{[n \times p]}$: data matrix. *Aim :* $\underset{\mathbf{a}}{\text{Max}}\,(\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a})$

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.



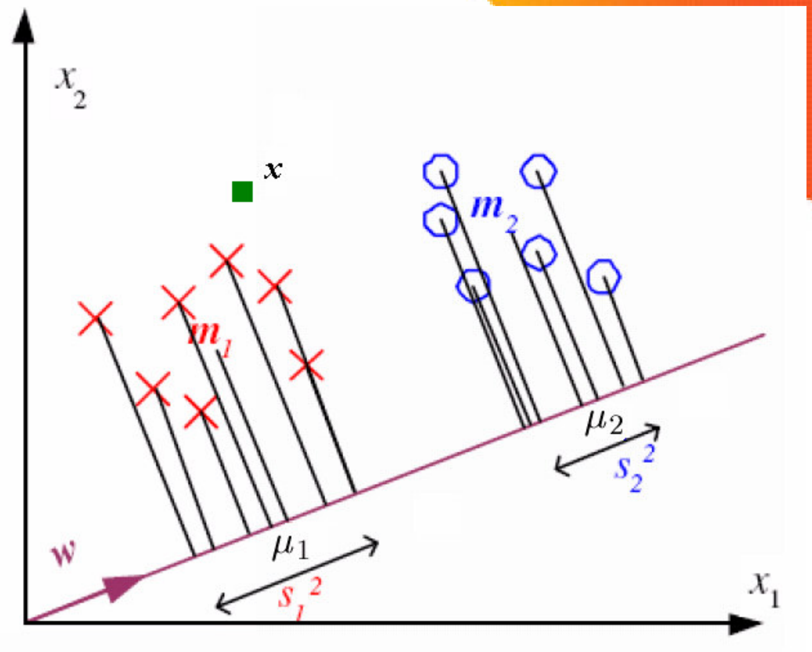
Solution:

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 region as possible.



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

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

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

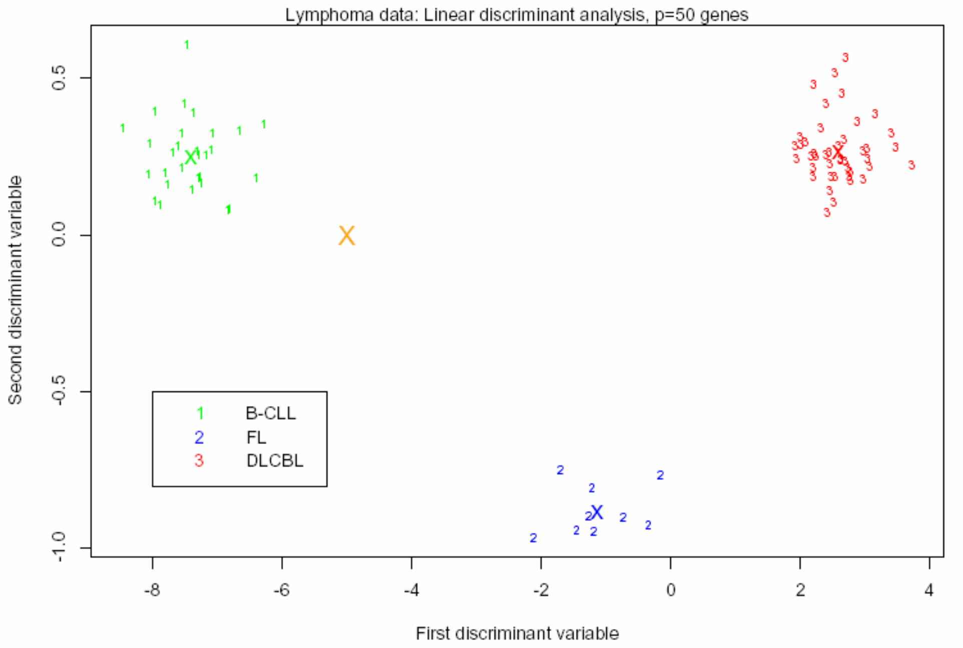
$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

$\bar{x}_{.j}$ 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

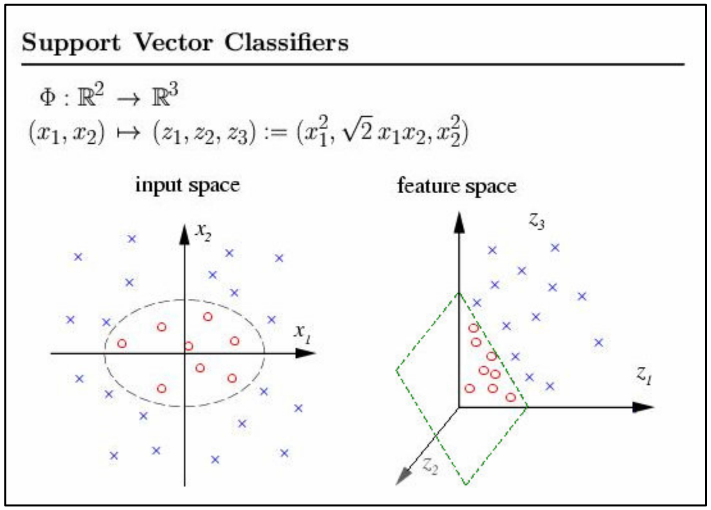
the p genes with the largest BSS/WSS ratios.



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)

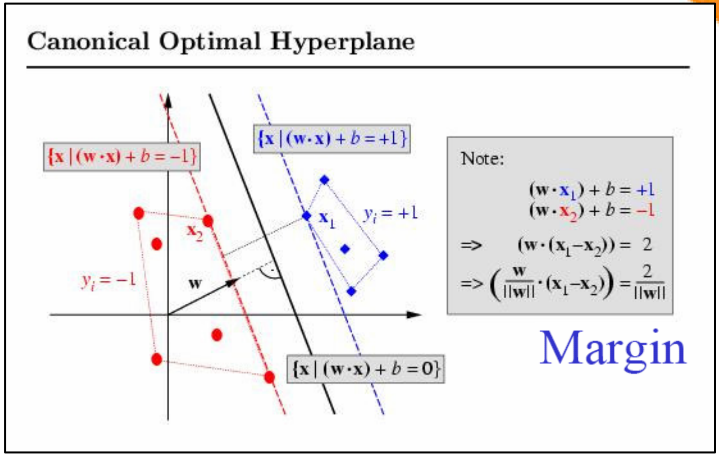
SVMs (Vapnik, 1995) map the data (input space) into high dimensional space (feature space) through a kernel function ϕ and then find a hyperplane $\mathbf{w}$ to separate two groups (binary classification).



Support Vector Classifiers

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$

$$\mathbf{K} := [\mathbf{K}_{ij} =< \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) >]$$



Canonical Optimal Hyperplane

$\{\mathbf{x} \mid (\mathbf{w} \cdot \mathbf{x}) + b = -1\}$
$\{\mathbf{x} \mid (\mathbf{w} \cdot \mathbf{x}) + b = +1\}$
$\{\mathbf{x} \mid (\mathbf{w} \cdot \mathbf{x}) + b = 0\}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$y_i = +1$
$y_i = -1$
$\mathbf{w}$
Note:
$(\mathbf{w} \cdot \mathbf{x}_1) + b = +1$
$(\mathbf{w} \cdot \mathbf{x}_2) + b = -1$
$\Rightarrow \quad (\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2)) = 2$
$\Rightarrow \quad \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1 - \mathbf{x}_2)\right) = \frac{2}{\|\mathbf{w}\|}$
Margin

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

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

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

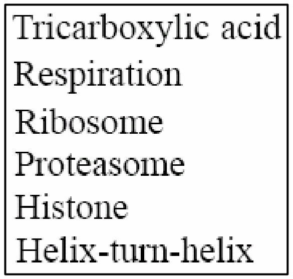
SVM (conti.)

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

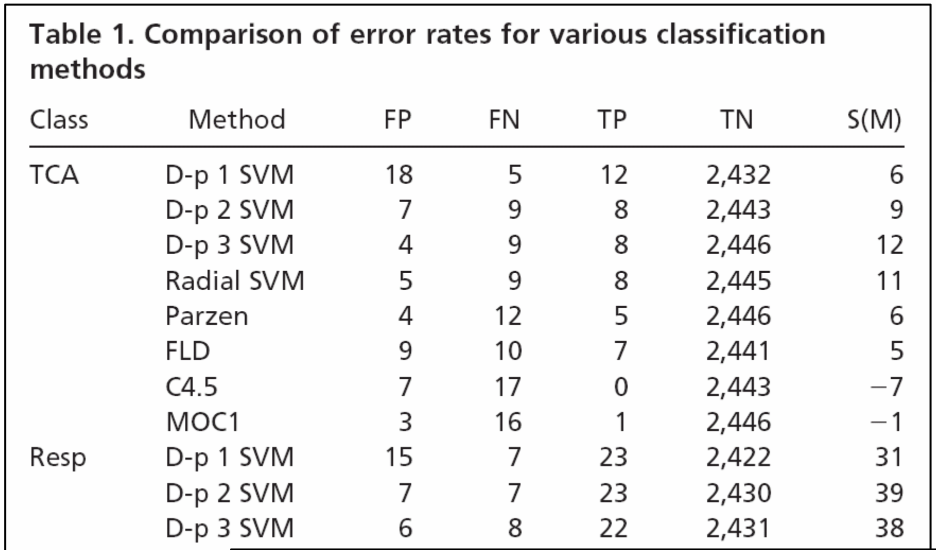


Table 1. Comparison of error rates for various classification methods

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

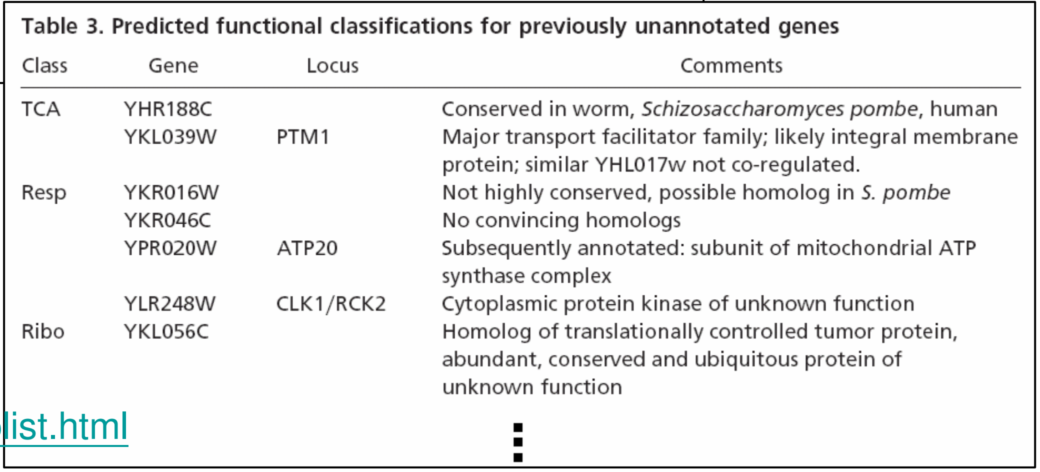


Table 3. Predicted functional classifications for previously unannotated genes

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 for Statistical Analysis and Visualization

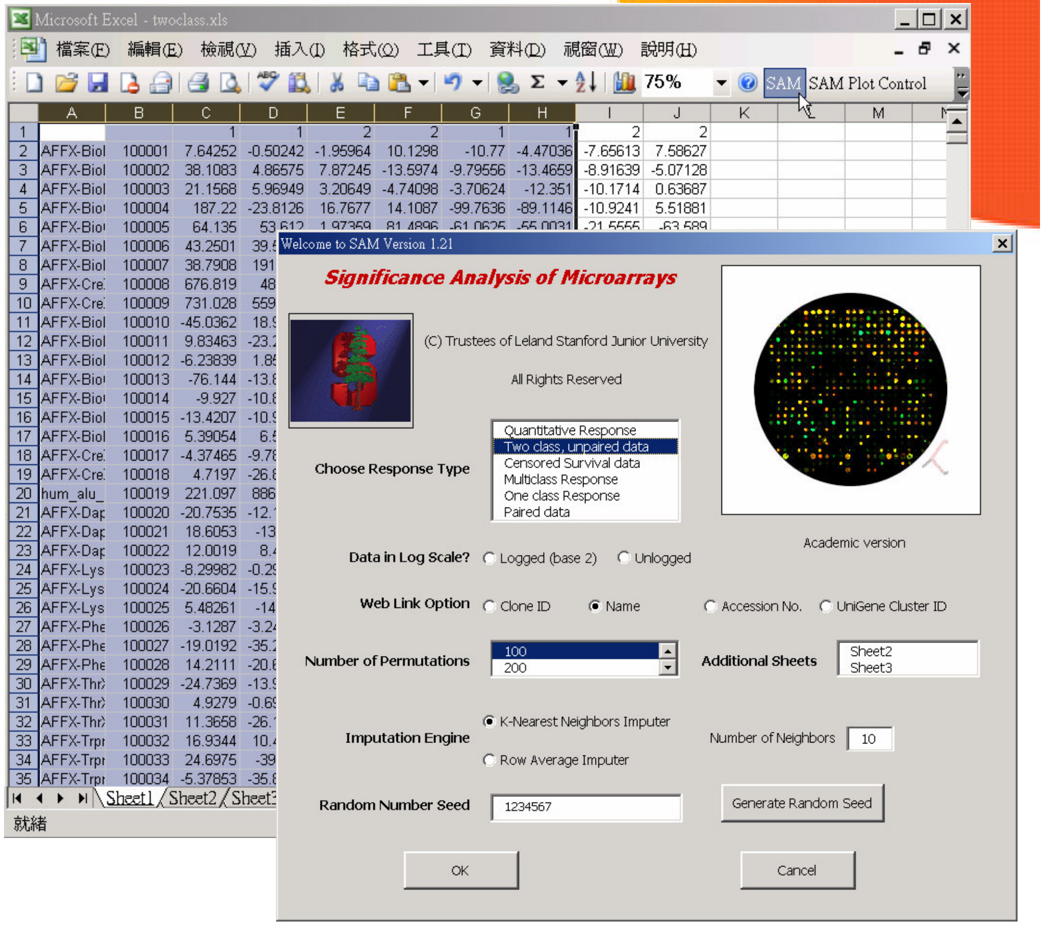
- *Freeware/Shareware*
  - Significance Analysis of Microarray (SAM)
  - Cluster and TreeView
  - The Bioconductor: limma, LImmaGUI
- *Commercial*
  - Matlab: Bioinformatics ToolBox
  - GeneSpring

Significance Analysis of Microarray

SAM assigns a score to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

- ***False discovery rate***: is the percent of genes that are expected to be identified by chance.
- ***q-value***: the lowest false discovery rate at which a gene is described as significantly regulated.
- ***Output plot***: the number of observed genes versus the expected number. This visualizes the outlier genes that are most dramatically regulated.

SAM does not do any normalization!


Microsoft Excel - twoclass.xls
檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 工具(T) 資料(D) 視窗(W) 說明(H)
75%
SAM
SAM Plot Control
Sheet1
Sheet2
Sheet3
就緒

	A	B	C	D	E	F	G	H	I	J
1			1	1	2	2	1	1	2	2
2	AFFX-Biol	100001	7.64252	-0.50242	-1.95964	10.1298	-10.77	-4.47036	-7.65613	7.58627
3	AFFX-Biol	100002	38.1083	4.86575	7.87245	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128
4	AFFX-Biol	100003	21.1568	5.96949	3.20649	-4.74098	-3.70624	-12.351	-10.1714	0.63687
5	AFFX-Biol	100004	187.22	-23.8126	16.7677	14.1087	-99.7636	-89.1146	-10.9241	5.51881
6	AFFX-Biol	100005	64.135	53.612	1.97359	81.4896	-61.0625	-55.0031	-21.5555	-63.589
7	AFFX-Biol	100006	43.2501	39.5						
8	AFFX-Biol	100007	38.7908	191						
9	AFFX-Cre	100008	676.819	48						
10	AFFX-Cre	100009	731.028	559						
11	AFFX-Biol	100010	-45.0362	18.9						
12	AFFX-Biol	100011	9.83463	-23.2						
13	AFFX-Biol	100012	-6.23839	1.85						
14	AFFX-Biol	100013	-76.144	-13.8						
15	AFFX-Biol	100014	-9.927	-10.8						
16	AFFX-Biol	100015	-13.4207	-10.9						
17	AFFX-Biol	100016	5.39054	6.5						
18	AFFX-Cre	100017	-4.37465	-9.78						
19	AFFX-Cre	100018	4.7197	-26.8						
20	hum_alu_	100019	221.097	886						
21	AFFX-Dap	100020	-20.7535	-12.1						
22	AFFX-Dap	100021	18.6053	-13						
23	AFFX-Dap	100022	12.0019	8.4						
24	AFFX-Lys	100023	-8.29982	-0.29						
25	AFFX-Lys	100024	-20.6604	-15.9						
26	AFFX-Lys	100025	5.48261	-14						
27	AFFX-Phe	100026	-3.1287	-3.24						
28	AFFX-Phe	100027	-19.0192	-35.2						
29	AFFX-Phe	100028	14.2111	-20.6						
30	AFFX-Thr	100029	-24.7369	-13.9						
31	AFFX-Thr	100030	4.9279	-0.69						
32	AFFX-Thr	100031	11.3658	-26.1						
33	AFFX-Trpr	100032	16.9344	10.4						
34	AFFX-Trpr	100033	24.6975	-39						
35	AFFX-Trpr	100034	-5.37853	-35.8						

Welcome to SAM Version 1.21

Significance Analysis of Microarrays

(C) Trustees of Leland Stanford Junior University

All Rights Reserved

Academic version

Choose Response Type: Quantitative Response / **Two class, unpaired data** / Censored Survival data / Multiclass Response / One class Response / Paired data

Data in Log Scale? ○ Logged (base 2) ○ Unlogged

Web Link Option ○ Clone ID ● Name ○ Accession No. ○ UniGene Cluster ID

Number of Permutations: **100** / 200

Additional Sheets: Sheet2 / Sheet3

Imputation Engine ● K-Nearest Neighbors Imputer ○ Row Average Imputer

Number of Neighbors: 10

Random Number Seed: 1234567

Generate Random Seed

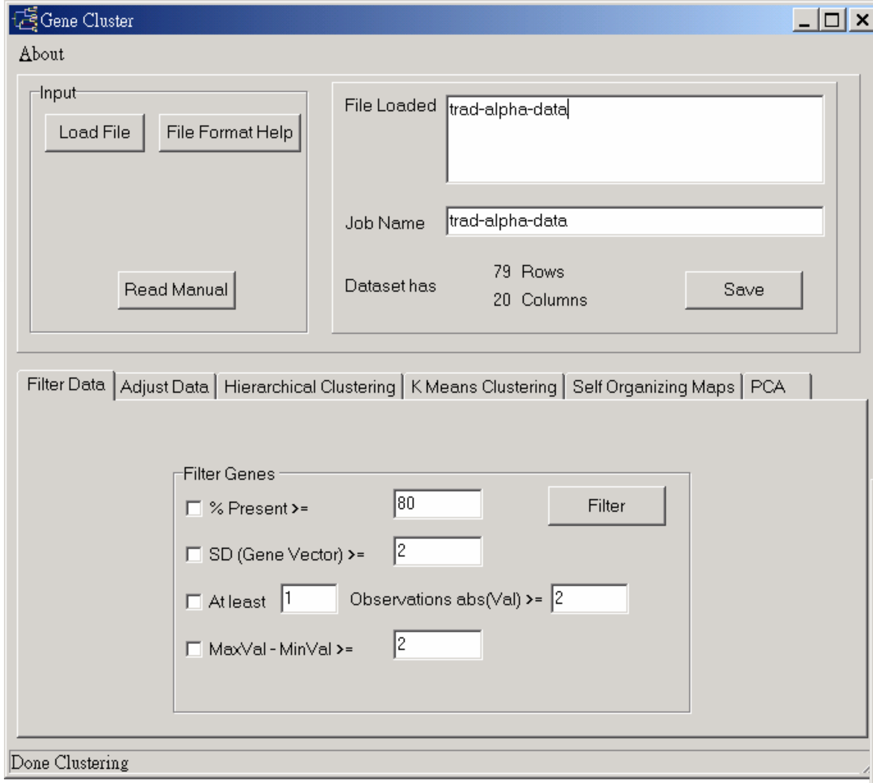
OK

Cancel

Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.

http://www-stat.stanford.edu/~tibs/SAM/

Cluster and TreeView



File Loaded: trad-alpha-data

Job Name: trad-alpha-data

Dataset has 79 Rows 20 Columns

☐ % Present >= 80

☐ SD (Gene Vector) >= 2

☐ At least 1 Observations abs(Val) >= 2

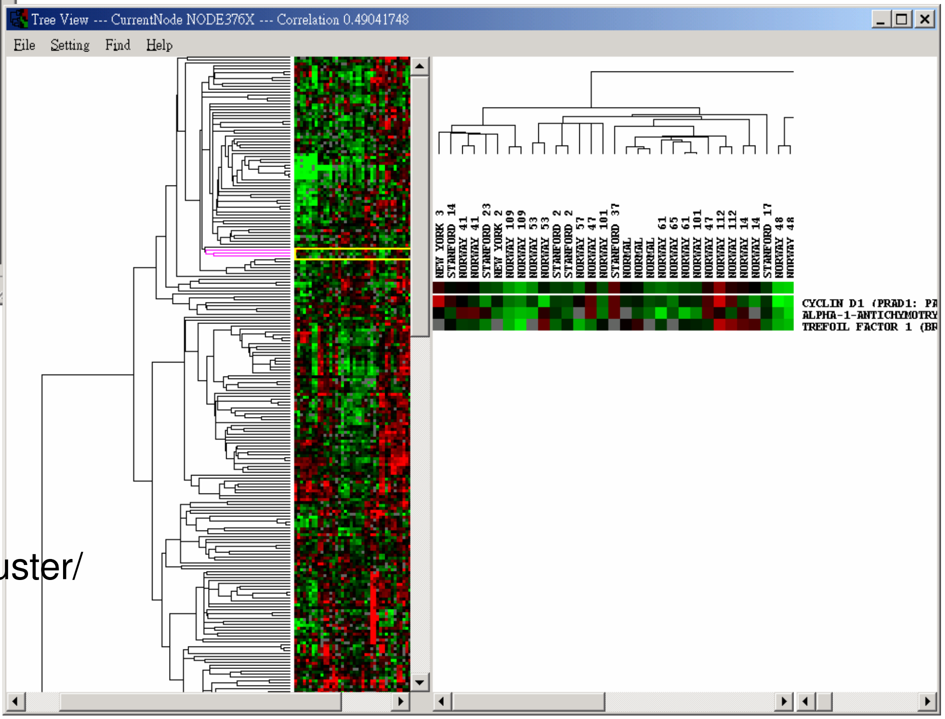
☐ MaxVal - MinVal >= 2

Done Clustering

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.



Tree View --- CurrentNode NODE376X --- Correlation 0.49041748

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

daMA

edd

externalVector

factDesign

gcrma

genefilter

geneplotter

globaltest

gpls

graph

hexbin

limma

sigg

splic

tkW

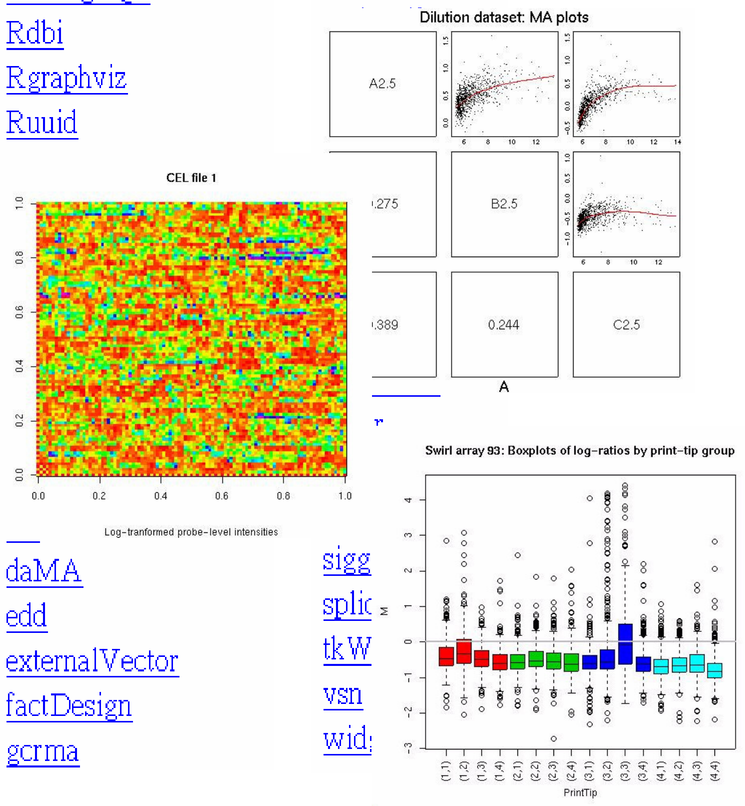
vsn

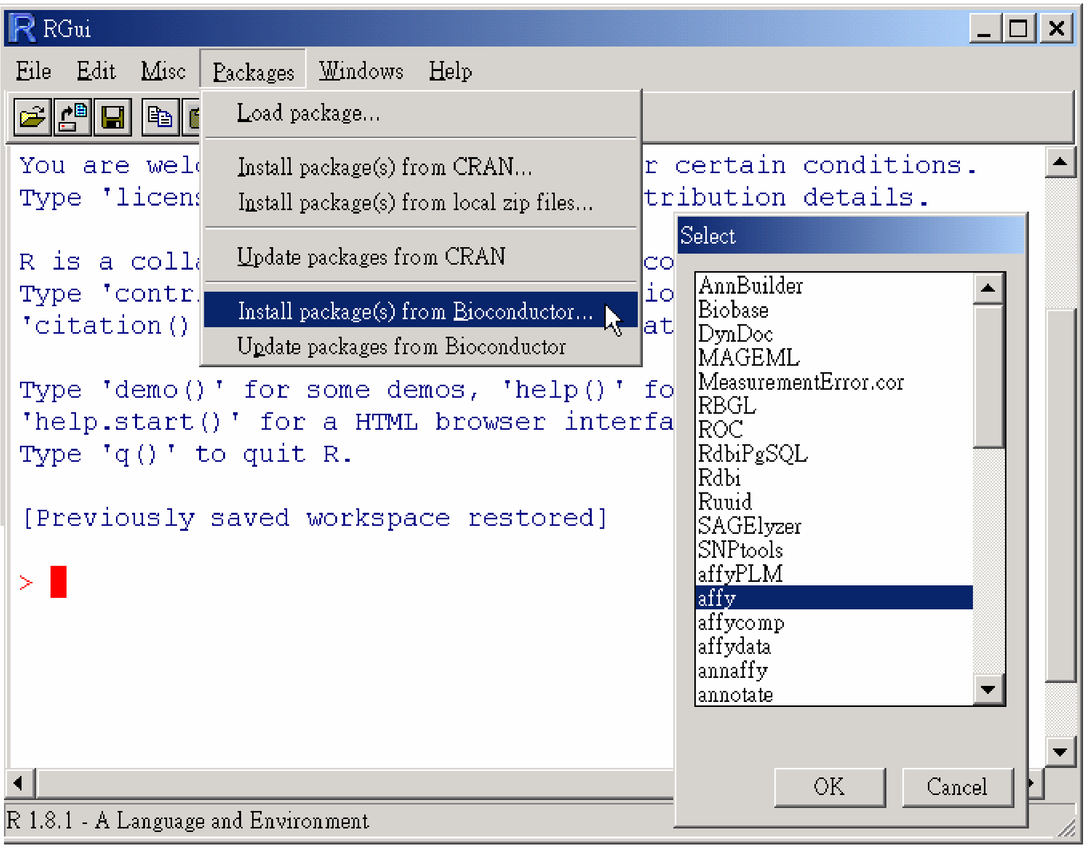
wid

The Bioconductor
version 1.7 (2005-10-14)
http://www.bioconductor.org

The R Project for Statistical Computing

R version 2.2.0 (**2005-10-6**)
http://www.r-project.org

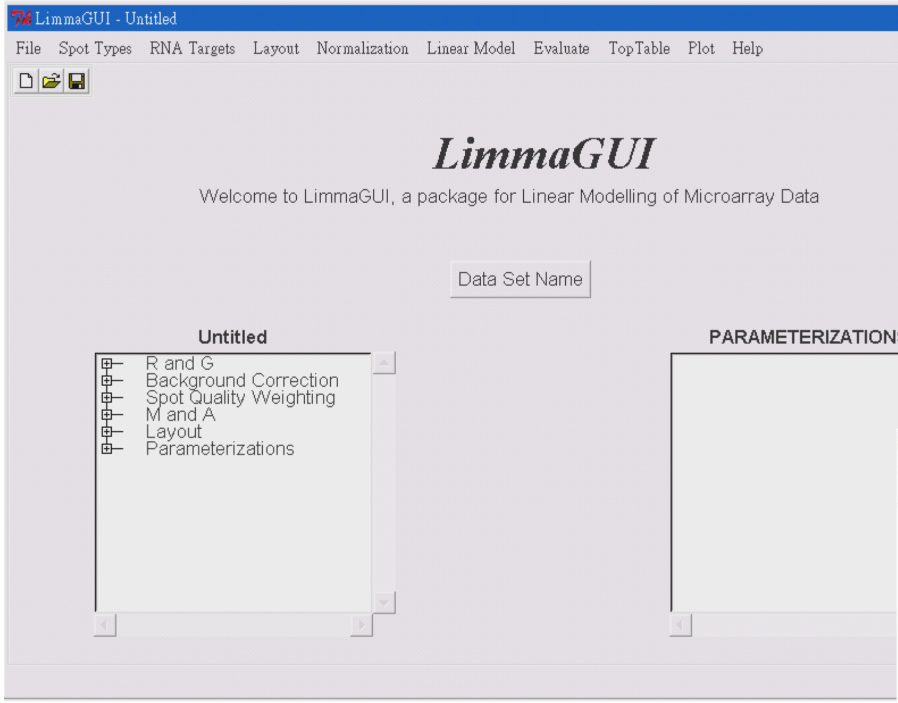


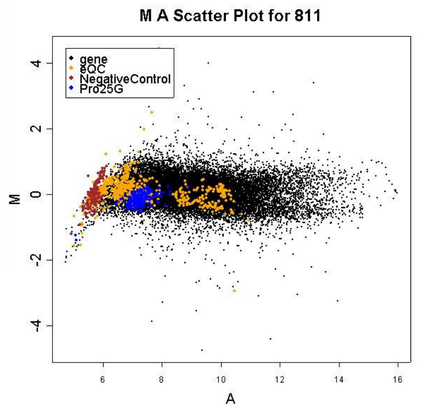


```
You are welc                              r certain conditions.
Type 'licens                              tribution details.

R is a colla                              co
Type 'contr.                              io
'citation()                               at

Type 'demo()' for some demos, 'help()' fo
'help.start()' for a HTML browser interfa
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

R 1.8.1 - A Language and Environment

Limma, LimmaGUI, LimmaAffy





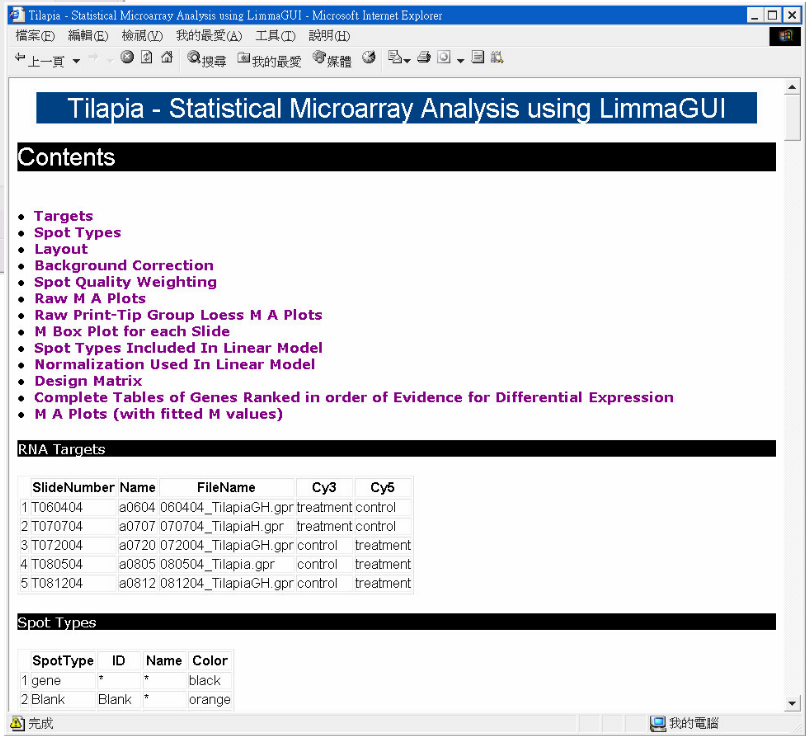
Limma: Linear Models for Microarray Data

http://bioinf.wehi.edu.au/limma/

LimmaGUI: a menu driven interface of Limma

http://bioinf.wehi.edu.au/limmaGUI

- Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Chapter 23. (To be published in 2005)
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, No. 1, Article 3.


Tilapia - Statistical Microarray Analysis using LimmaGUI - 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
完成
我的電腦

Tilapia - Statistical Microarray Analysis using LimmaGUI

Contents

- Targets
- Spot Types
- Layout
- Background Correction
- Spot Quality Weighting
- Raw M A Plots
- Raw Print-Tip Group Loess M A Plots
- M Box Plot for each Slide
- Spot Types Included In Linear Model
- Normalization Used In Linear Model
- Design Matrix
- Complete Tables of Genes Ranked in order of Evidence for Differential Expression
- M A Plots (with fitted M values)

RNA Targets

	SlideNumber	Name	FileName	Cy3	Cy5
1	T060404	a0604	060404_TilapiaGH.gpr	treatment	control
2	T070704	a0707	070704_TilapiaH.gpr	treatment	control
3	T072004	a0720	072004_TilapiaGH.gpr	control	treatment
4	T080504	a0805	080504_Tilapia.gpr	control	treatment
5	T081204	a0812	081204_TilapiaGH.gpr	control	treatment

Spot Types

	SpotType	ID	Name	Color
1	gene	*	*	black
2	Blank	Blank	*	orange

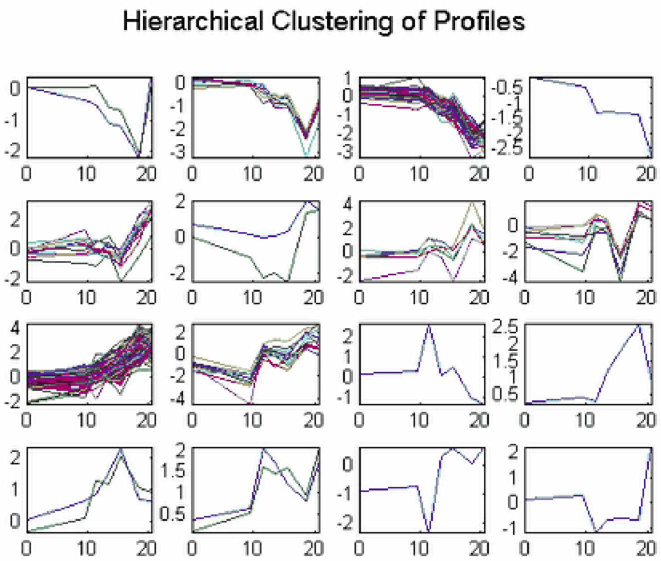
Matlab: Bioinformatics ToolBox

The MathWorks

Bioinformatics Toolbox

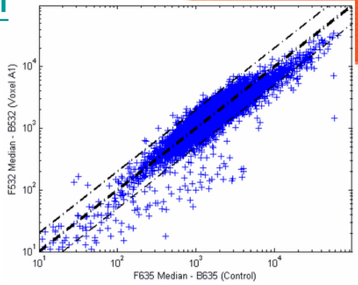
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html

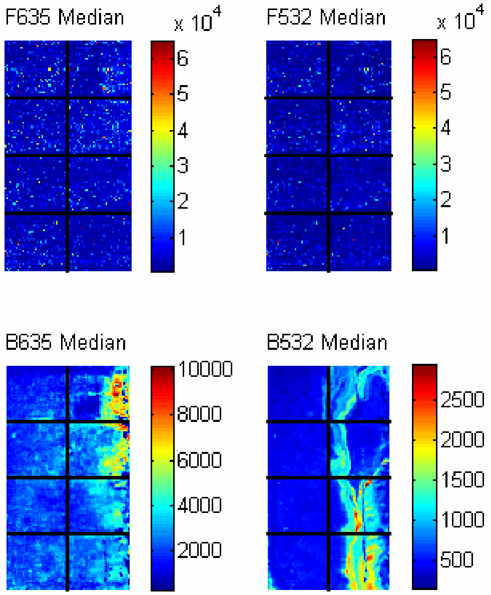
- Data Formats and Databases — Access online databases, read and write to files with standard genome and proteome formats such as FASTA and PDB.
- Sequence Alignments — Compare nucleotide or amino acid sequences using pairwise and multiple sequence alignment functions.
- Sequence Utilities and Statistics — 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.



Hierarchical Clustering of Profiles

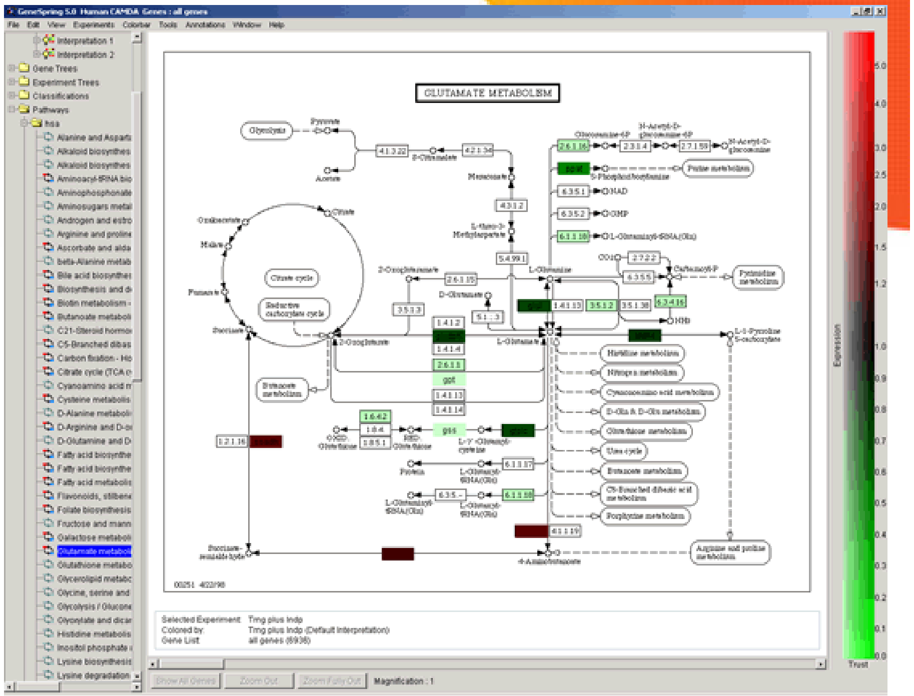
- Prototype and Development Environment — Create new algorithms, try new ideas, and compare alternatives.
- Share Algorithms and Deploy Applications — Create GUIs and stand-alone applications.



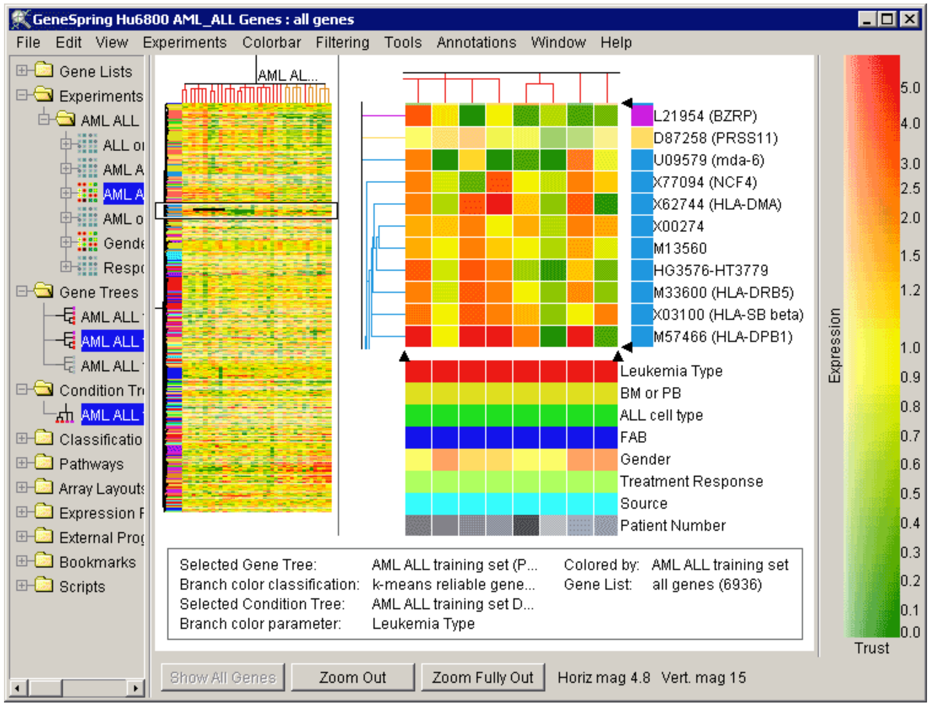


GeneSpring v7.2

- 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



Selected Experiment: Tmg plus Indp
Colored by: Tmg plus Indp (Default Interpretation)
Gene List: all genes (6936)



Selected Gene Tree: AML ALL training set (P...
Branch color classification: k-means reliable gene...
Selected Condition Tree: AML ALL training set D...
Branch color parameter: Leukemia Type
Colored by: AML ALL training set
Gene List: all genes (6936)

Horiz mag 4.8 Vert. mag 15

Images from
http://www.silicongenetics.com

Agilent Technologies

2004 Articles Citing GeneSpring®

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

More than 700 papers