Statistical

Microarray Data Analysis

Course：國立中正大學 分子生物研究所

生物晶片及其生醫應用

2005/05/12

吳漢銘

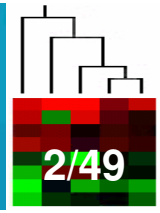
hmwu@stat.sinica.edu.tw

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

中央研究院 統計科學研究所

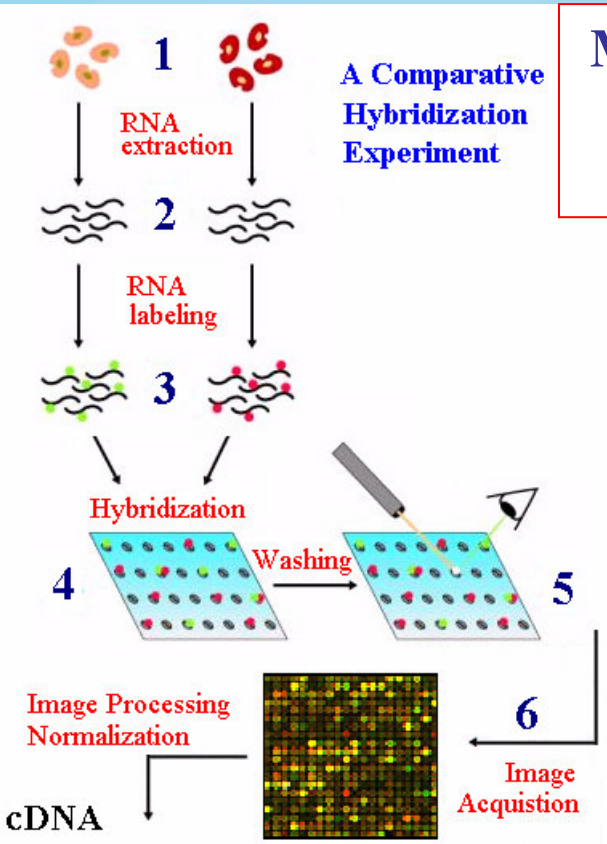
Institute of Statistical Science, Academia Sinica

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

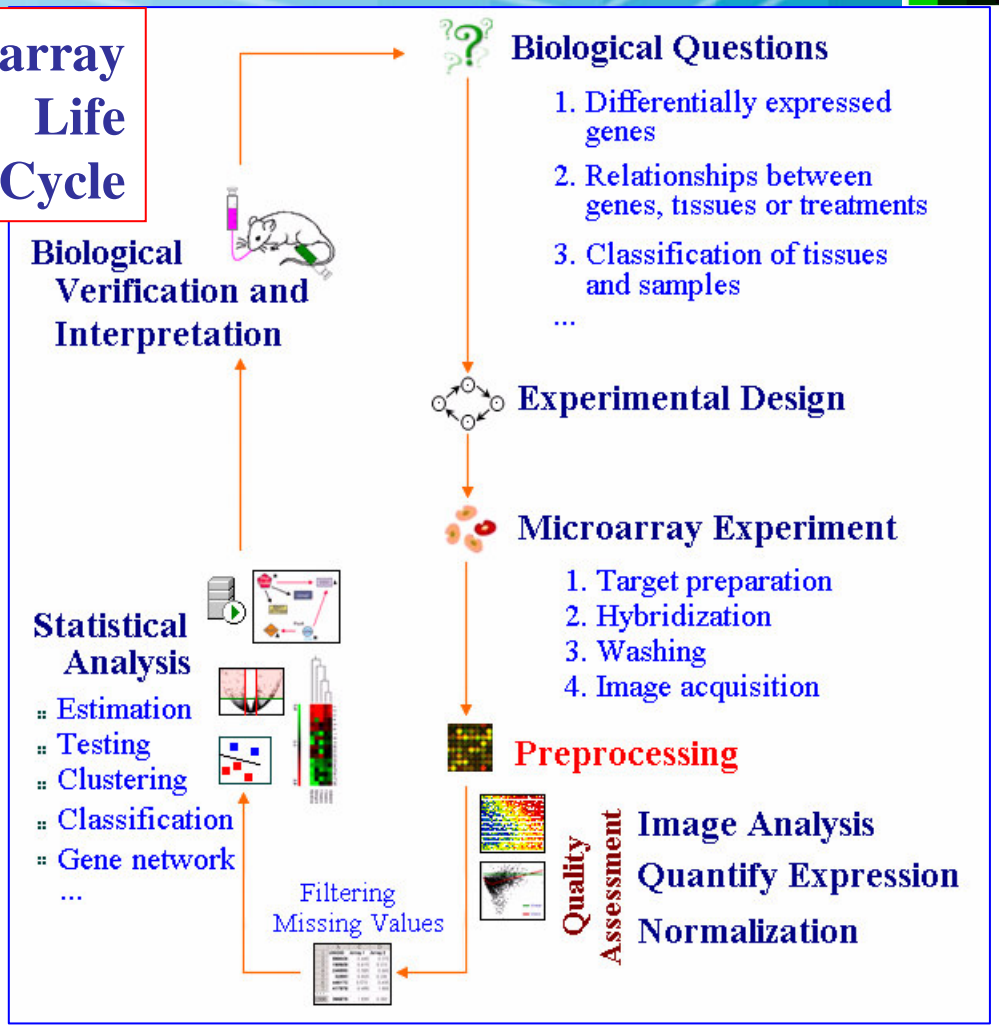


A Comparative Hybridization Experiment

	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

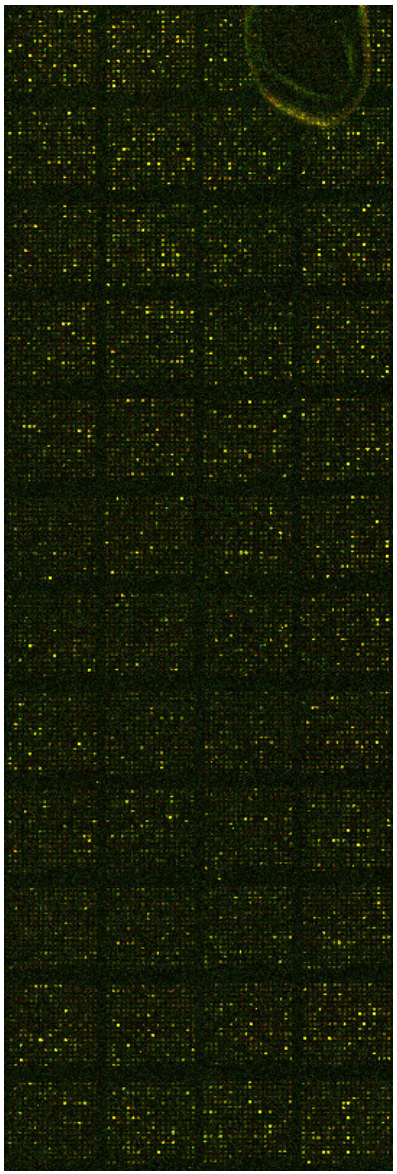


Spot color	*Signal strength*	*Gene expression*
Yellow	Control = Treated	Unchanged
Red	Control < Treated	Induced
Green	Control > Treated	Repressed

R=Rf-Rb
G=Gf-Gb

M=log2R/G
A=1/2 log2RG

Array Image



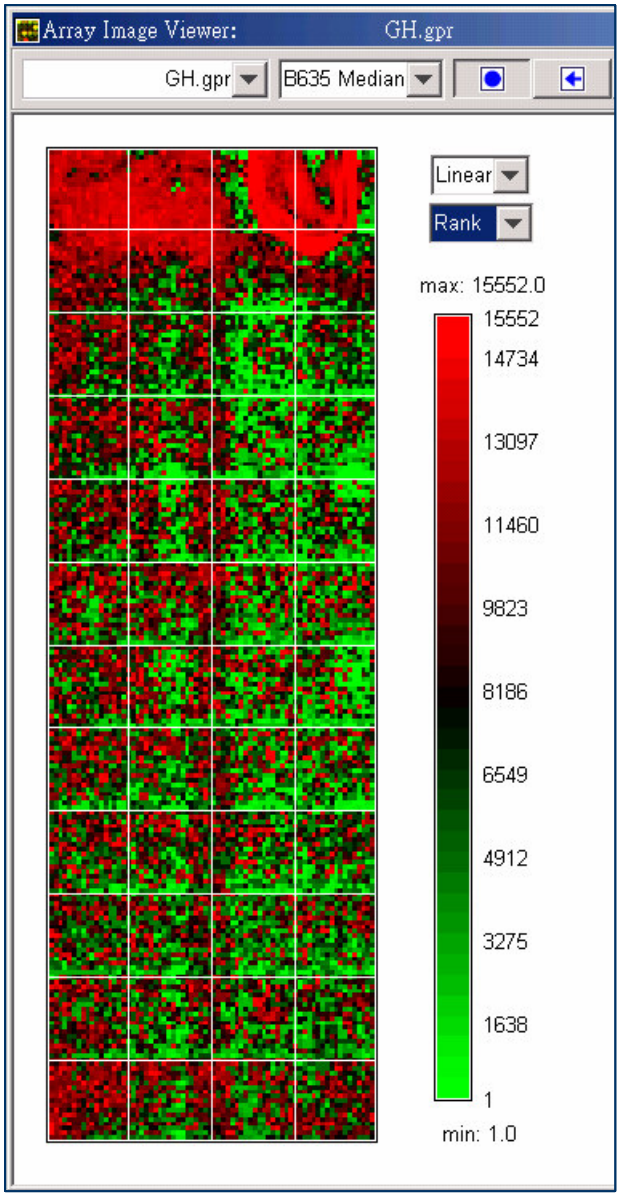
Blocks:
12 by 4

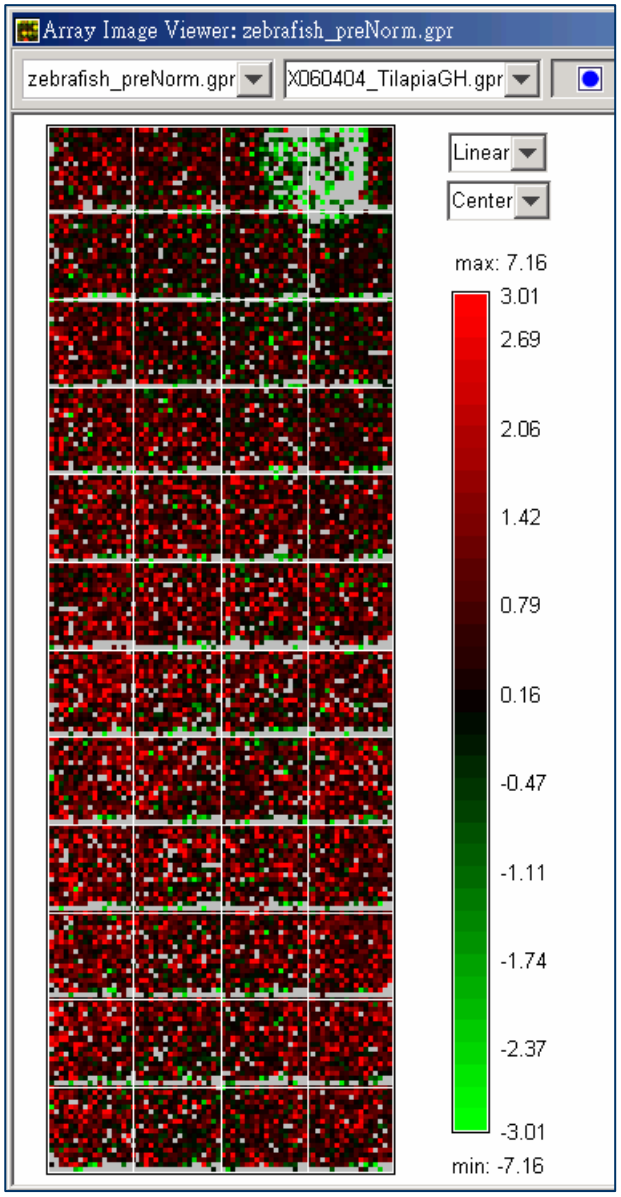
Features:
18 by 18

Signal
16-bit
0~65535

*.gpr

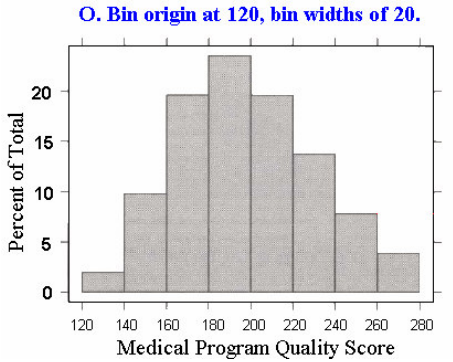
GAL





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.



Disadvantage for displaying a variable's distribution:

- selection of rigin 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.

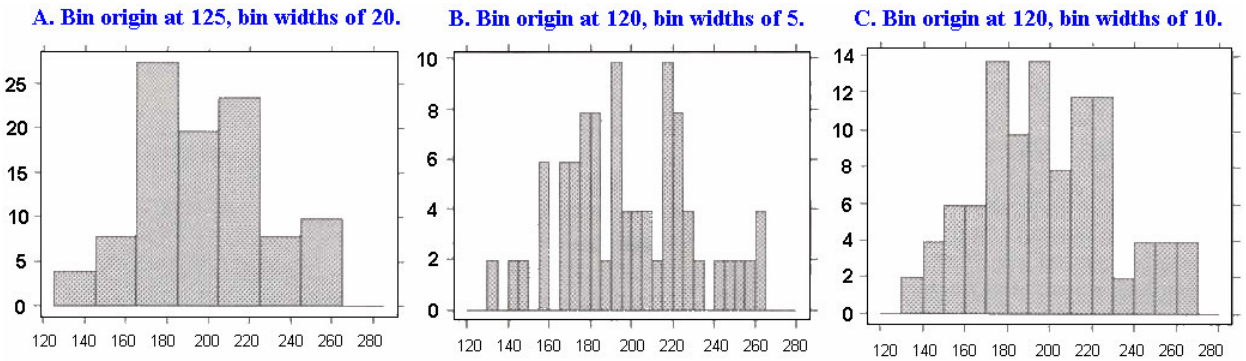
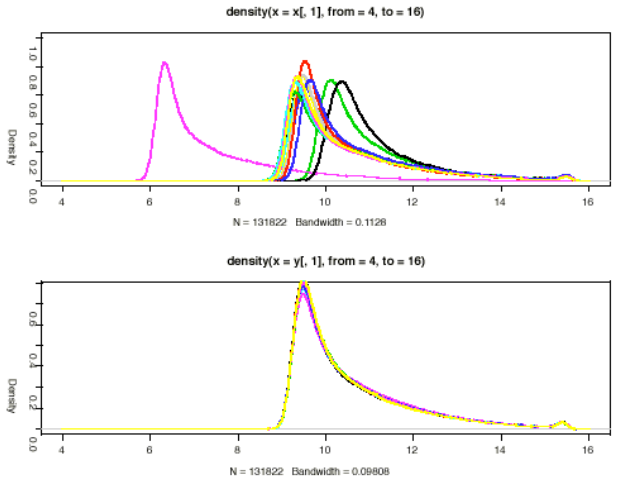


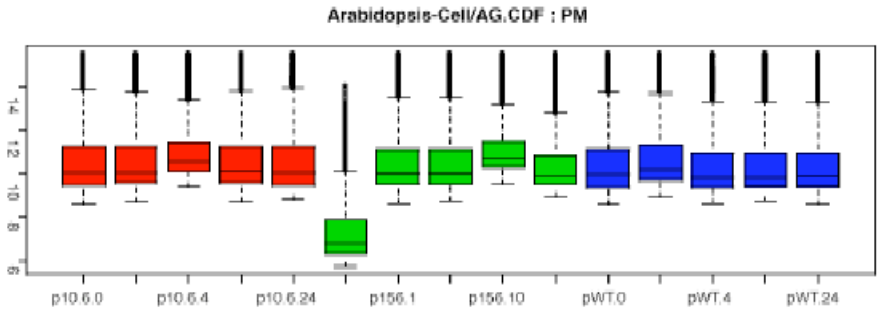
Figure Sources: Jacoby (1997).

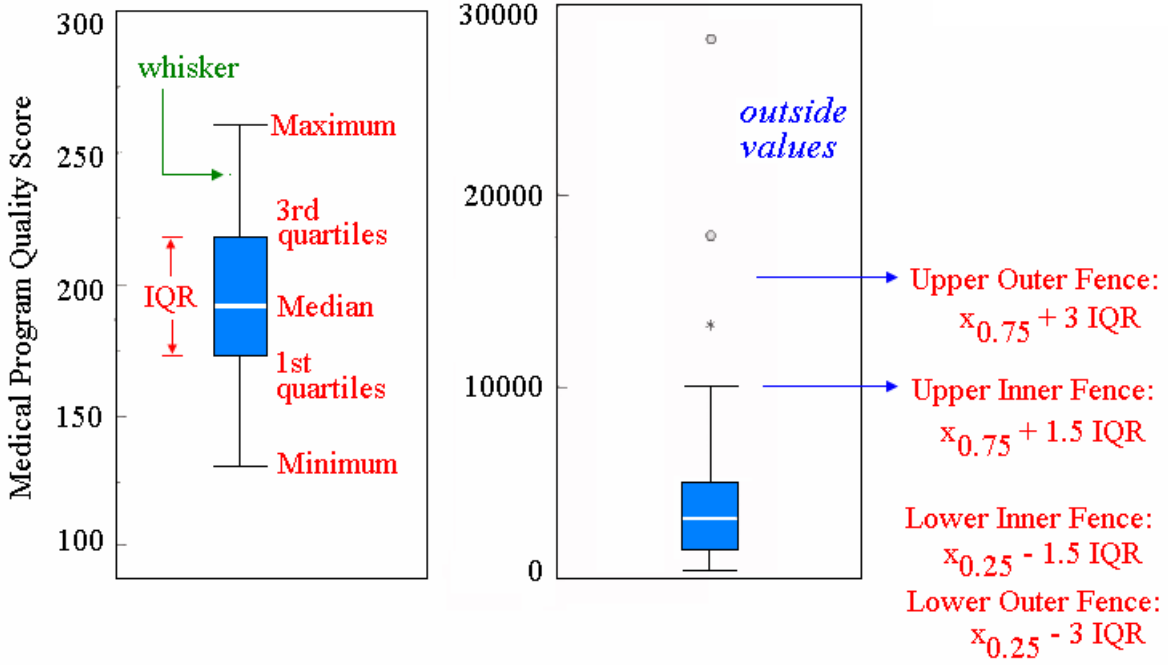
Density Plots



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.





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

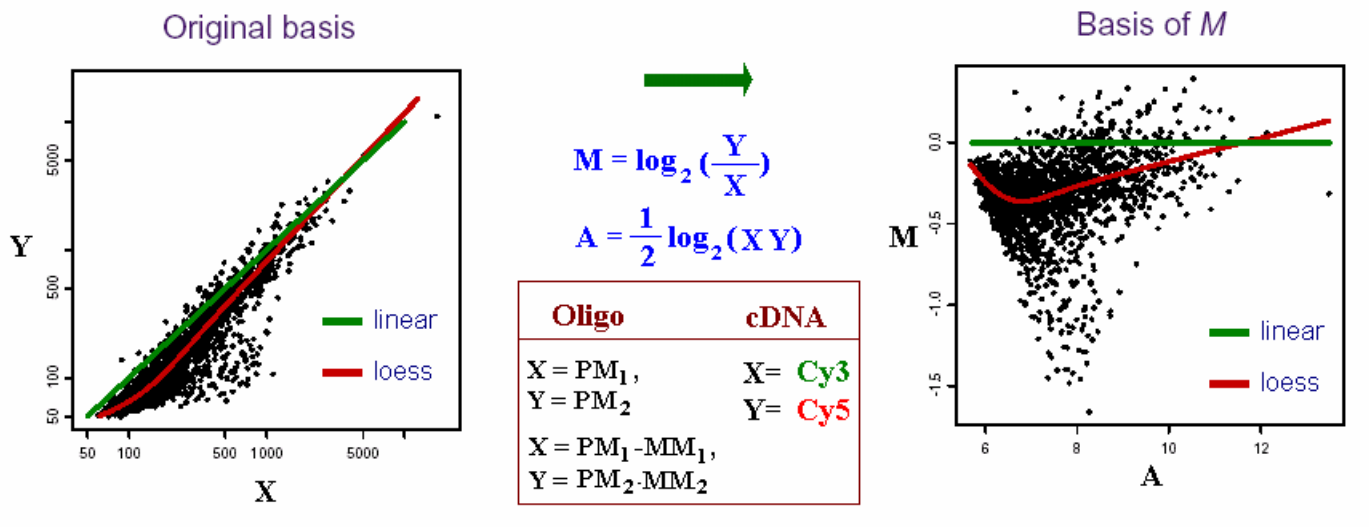
Scatterplot and MA plot

- **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.

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.



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

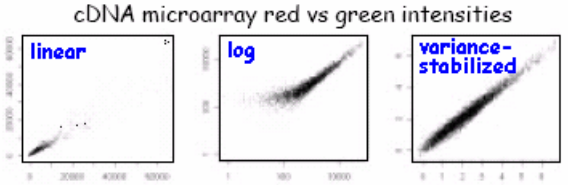
Quantification of Expression

Red intensity (Cy5) = Rfg - Rbg

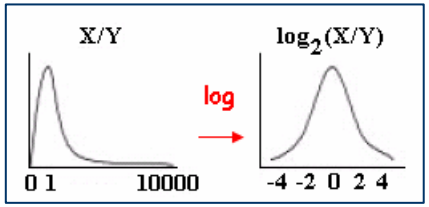
Green intensity (Cy3) = Gfg - Gbg

→ 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.
  - **Main idea:** remove the systematic bias in the data as completely possible while preserving the variation in the gene expression that occurs because of biologically relevant changes in transcription.



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.

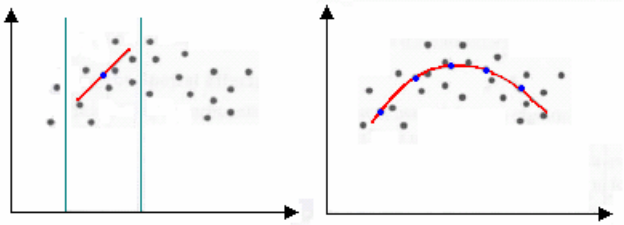
Basic Assumption

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)



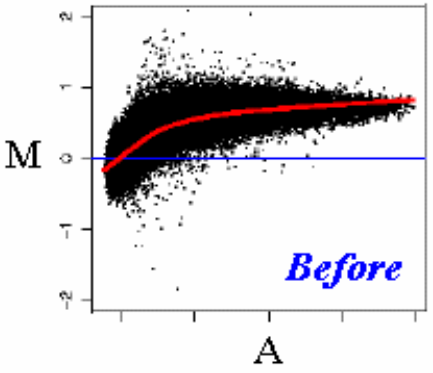
1. For any two arrays i,j with probe intensities x_{ki} and x_{kj} where $k = 1,\ldots,p$ represents the probe
2. we calculate $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \frac{1}{2}\log_2(x_{ki}x_{kj})$.
3. A normalization curve is fitted to this M versus A plot using loess.

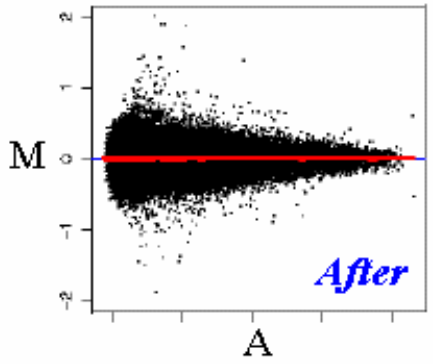
   Loess is a method of local regression (see Cleveland and Devlin (1988) for details).
4. The fits based on the normalization curve are $\hat{M}_k$
5. the normalization adjustment is $M'_k = M_k - \hat{M}_k$.
6. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k + \frac{M'_K}{2}}$ and $x'_{kj} =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

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

$$A = \frac{1}{2}\log_2(XY)$$

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

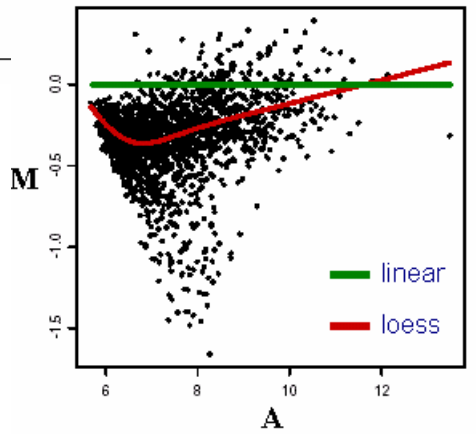
	Subject be used for estimating normalization curve			
Location Normalization				
Method	allGenes	Print-tip i	2D Location (x, y)	SelectedGenes (Controls, Housekeeping, MSP, Invariant set)
constant	global normalization $N = M - c$ c : mean, median	print-tip normalization $N = M - c_i$		
loess (Robust scatterplot 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 $N = M - ave(M)/std(A)$			

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 (DE) Genes

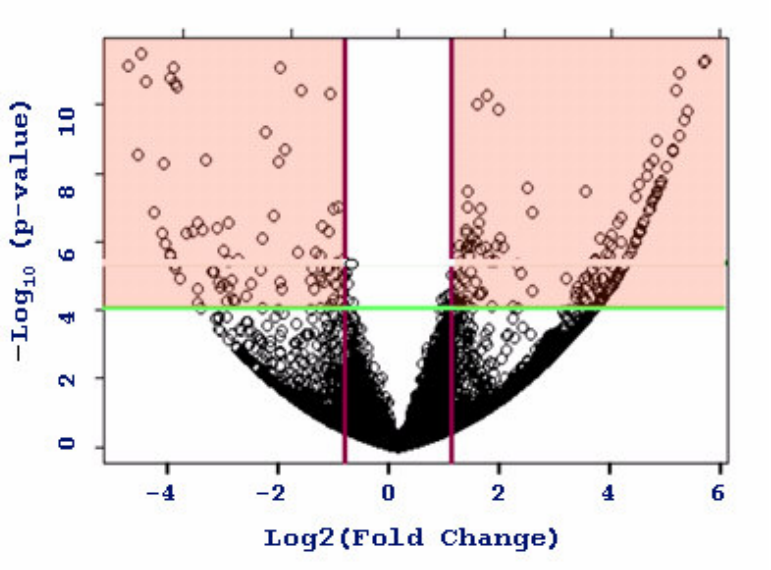
- Select a statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence.

(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.

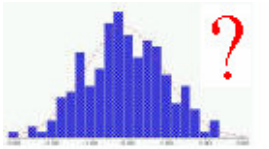
Define which genes are significantly regulated might be to choose 5% of genes that have the largest expression ratios.

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 **Dependent samples**	Unpaired data **Independent samples**	Complex data *More than two Groups*
Parametric Hypothesis Testing	■ z-test ■ ***t-test***	■ ***two-sample 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

An Example

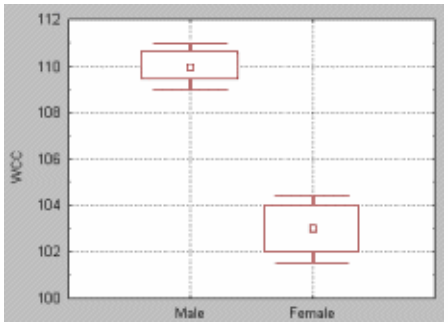
- A *hypothesis test* is a procedure for determining if an assertion about a characteristic of a population is reasonable.
- For example, suppose that someone says that the average price of a gallon of regular unleaded gas in Massachusetts is $1.15. How would you decide whether this statement is true?
  - You could try to find out what every gas station in the state was charging and how many gallons they were selling at that price. That approach might be definitive, but it could end up costing more than the information is worth.
  - A simpler approach is to find out the price of gas at a small number of randomly chosen stations around the state and compare the average price to $1.15.
- Of course, the average price you get will probably not be exactly $1.15 due to variability in price from one station to the next.
- Suppose your average price was $1.18. Is this three cent difference a result of chance variability, or is the original assertion incorrect?

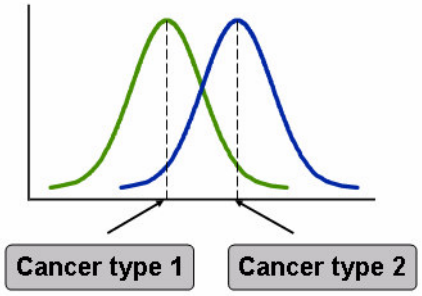
A hypothesis test can provide an answer.

Terminology in Hypothesis Testing

- The ***null hypothesis:***
  - H0: μ = 1.15. (the average price of a gallon of gas is $1.15)
- The ***alternative hypothesis:***
  - H1: μ > 1.15. (gas prices were actually higher)
  - H1: μ < 1.15.
  - H1: μ != 1.15.
- The ***significance level (alpha)*** is related to the degree of certainty you require in order to reject the null hypothesis in favor of the alternative.
  - Decide in advance to reject the null hypothesis if the probability of observing your sampled result is less than the significance level.
  - For a typical significance level of 5%, the notation is alpha = 0.05. For this significance level, the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.
  - If you need more protection from this error, then choose a lower value of alpha .
- The ***p-value*** is the probability of observing the given sample result under the assumption that the null hypothesis is true.
  - If the p-value is less than alpha, then you reject the null hypothesis.
  - For example, if alpha = 0.05 and the p-value is 0.03, then you reject the null hypothesis.
- ***Confidence intervals:*** a range of values that have a chosen probability of containing the true hypothesized quantity.
  - Suppose, in our example, 1.15 is inside a 95% confidence interval for the mean, μ. That is equivalent to being unable to reject the null hypothesis at a significance level of 0.05.
  - Conversely if the 100(1- alpha)% confidence interval does not contain 1.15, then you reject the null hypothesis at the alpha level of significance.

$$\text{Power} = 1 - \beta.$$





Hypothesis Testing		Truth	
		H0	**H1**
Decision	**Reject H0**	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H0	Right Decision	Type II Error (beta)

One Sample t-test

The One-Sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.

Assumption: the variable is normally distributed.

One sample t-test

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$ (two-tailed).
μ: population mean.
α: significant level (e.g., 0.05).
Test Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

S: sample standard deviation.

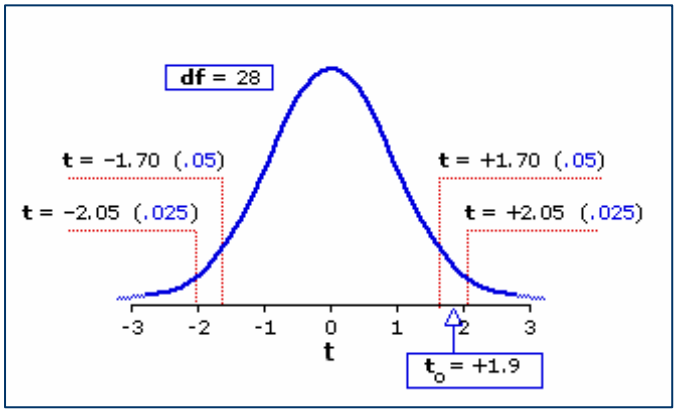
n: number of observations in the sample.

- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ: $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

Example

H0: no differential expressed.

- The test is significant
  = Reject **H0**
- False Positive
  = (Reject **H0** | **H0** true)
  = concluding that a gene is differentially expressed when in fact it is not.



Two Sample t-test

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d: $\bar{d} - t_{\alpha/2}S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

Two Sample t-test (Unpaired)

$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$

α: significant level (e.g., 0.05).

Test Statistic:

$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2,\, df}$

Paired t-test Applied to A Gene From Breast Cancer Data

- Samples are taken from 20 breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy, and analyzed using microarray. There are 9216 genes.
- **Paired data**: there are two measurements from each patient, one before treatment and one after treatment.
- These two measurements relate to one another, we are interested in the difference between the two measurements (the log ratio) to determine whether a gene has been up-regulated or down-regulated in breast cancer following doxoruicin chemotherapy.

- The samples from before and after chemotherapy have been hybridized on separate arrays, with a reference sample in the other channel.
  - Normalize the data.
  - Because this is a reference sample experiment, we calculate the log ratio of the experimental sample relative to the reference sample for before and after treatment in each patient.
  - Calculate a single log ratio for each patient that represents the difference in gene expression due to treatment by subtracting the log ratio for the gene before treatment from the log ratio of the gene after treatment.
  - Perform the t-test. t=3.22 compare to t(19).
  - The p-value for a two-tailed one sample t-test is 0.0045, which is significant at a 1% confidence level.
- Conclude: this gene has been significantly down-regulated following chemotherapy at the 1% level.

Perou et. Al, (2000), Molecular portraits of human breast tumours. Nature 406:747-752. http://genome-www.stanford.edu/breast_cancer/molecularportraits/

Unpaired t-test Applied to A Gene From Leukemia Dataset

- Bone marrow samples are taken from 27 patients suffering from acute lymphoblastic leukemia (ALL，急性淋巴細胞白血病) and 11 patients suffering from acute myeloid leukemia (AML，急性骨髓性白血病) and analyzed using Affymetrix arrays. There are 7070 genes.
- **Unpaired data**: there are two groups of patients (ALL, AML).
- We wish to identify the genes that are up- or down-regulated in ALL relative to AML. (i.e., to see if a gene is differentially expressed between the two groups.)

- The gene metallothionein IB is on the Affymetrix array used for the leukemia data.
  - To identify whether or not this gene is differentially expressed between the AML and ALL patients.
  - To identify genes which are up- or down-regulation in AML relative to ALL.
- Steps
  - the data is log transformed.
  - t=-3.4177, p=0.0016
- Conclude that the expression of metallothionein IB is significantly higher in AML than in ALL at the 1% level.

Golub et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531--537.
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Other Statistics

B-statistic

Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic $t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where a is estimated from the mean and standard deviation of the sample variances s^2.

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)
Efron et al (2001, JASA)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

Lonnstedt, I. and Speed, T.P. Replicated microarray data. *Statistica Sinica* , 12: 31-46, 2002

General Penalized t-statistic

(Lonnstedt et al 2001)

$$t = \frac{b}{s^* \times SE}$$

multiple regression model

Penalized two-sample t-statistic

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \quad \text{where } s^* = \sqrt{a + s^2}$$

Robust General Penalized t-statistic

Assigning Significance and Multiplicity of Testing

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiplicity of p-values.

- Since every sample hybridized to the arrays is the same reference sample, we know that no genes are differentially expressed: all measured differences in expression are experimental error.
  - By the very definition of a p-value, each gene would have a 1% chance of having a p-value of less than 0.01, and thus be significant at the 1% level.
  - Because there are 10000 genes on this imaginary microarray, we would expect to find 100 significant genes at this level.
  - Similarly, we would expect to find 10 genes with a p-value less than 0.001, and 1 gene with p-value less than 0.0001.

- **Bonferroni Correction**: the alpha level for statistical significance is divided by the number of measurements taken.

- ***Example:*** In Breast Cancer Dataset with 9216 genes, even if the chemotherapy had no effect whatsoever, we expect to find 92 differentially expressed genes with p-values less than 0.01, simple because of the large number of genes being analyzed.

Multiple Testing

- Controlling the Family-wise Error Rate
- Controlling the False Discovery Rate

Multiple Testing

	# Reject H_0	# not Reject H_0	
# true H_{0j}	V	U	m_0
# true H_{1j}	S	T	m_1
	R	m - R	m

V : *false positives = Type I errors*

T : *false negatives = Type II errors*

Type One Errors Rates

$$\text{PCER} = \frac{E[V]}{m}$$

$$\text{PFER} = E[V]$$

$$\text{FWER} = p(V \geq 1)$$

$$\text{FDR} = E[\frac{V}{R}] \quad \text{if } R > 0$$

Power = Reject the false null hypothesis

$$\text{Any-pair Power} = p(S \geq 1)$$

$$\text{Per-pair Power} = \frac{E[S]}{m_1}$$

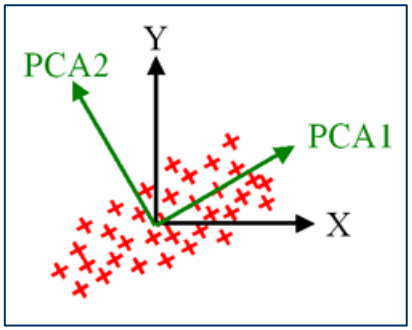
$$\text{All-pair Power} = p(S = m_1)$$

Reference for Finding Differential Expressed Genes

http://www.sinica.edu.tw/~hmwu/Talks/DEindex.htm

Principal Component Analysis (PCA)

(Pearson 1901; Hotelling 1933; Jolliffe 2002)



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

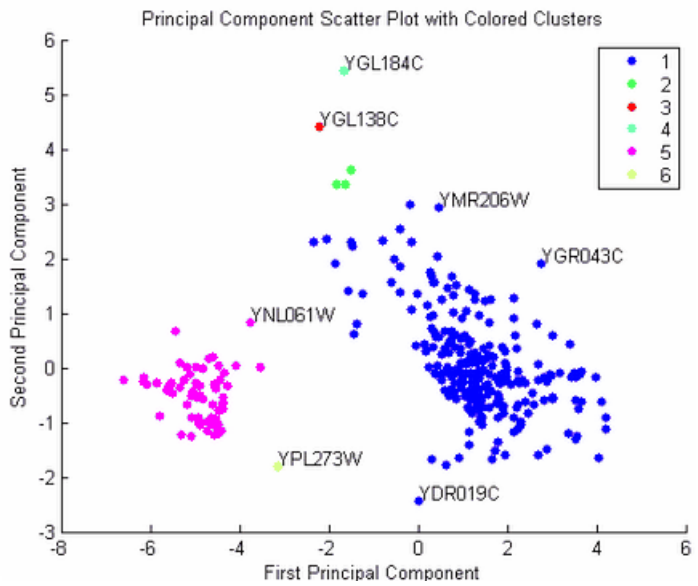
The PCA summaries the dispersion of data points as data cloud in a small number of major axes (principal components) of varition among the variables.

Goal: to reduce the dimensionality of the data matrix by finding the new variables.

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.



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)

Classical MDS

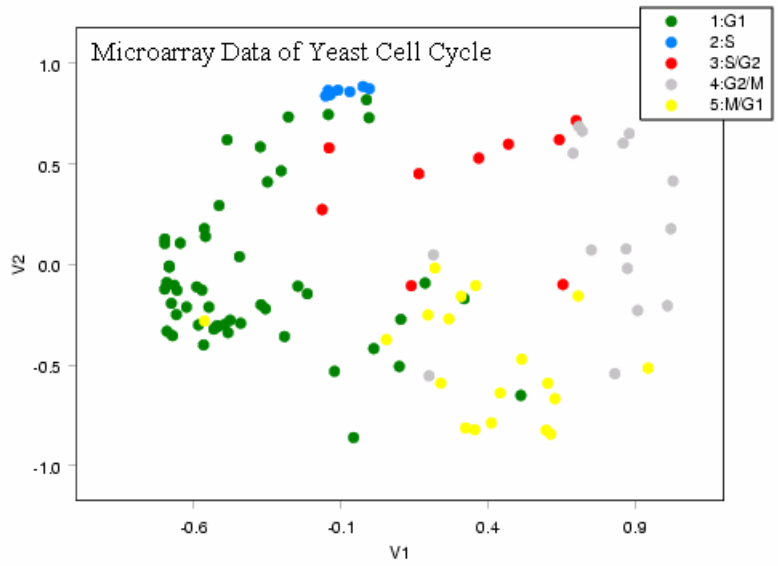
Multidimensional scaling takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.

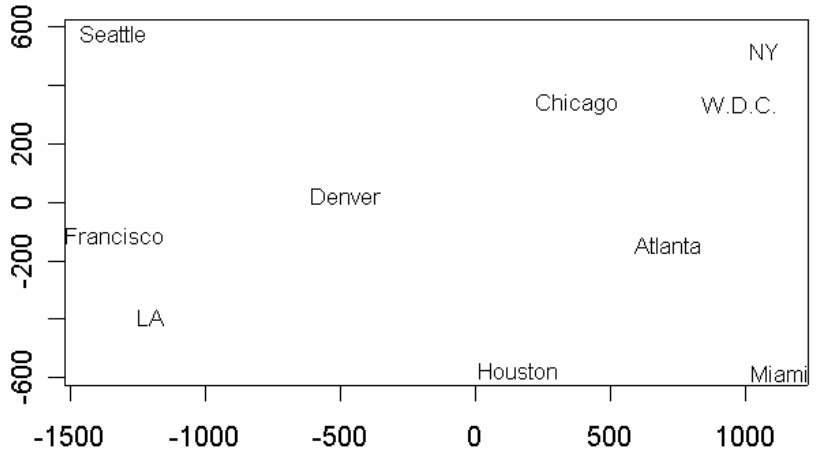
Note that if the input-space distances are Euclidean, classical MDS is equivalent to PCA. (Mardia et al. 1979)

Analysis of 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.

2D MDS configuration plot for 103 known genes

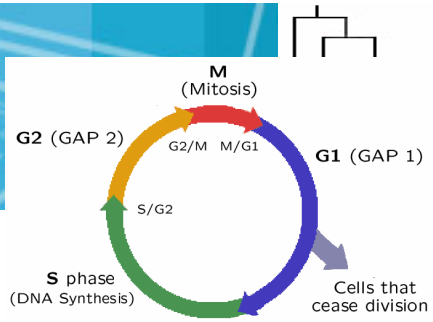




找microarray paper

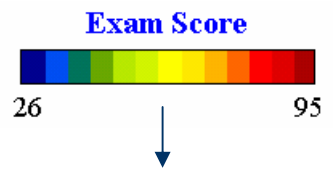
Heat Map
(Data Image, Matrix Visualization)

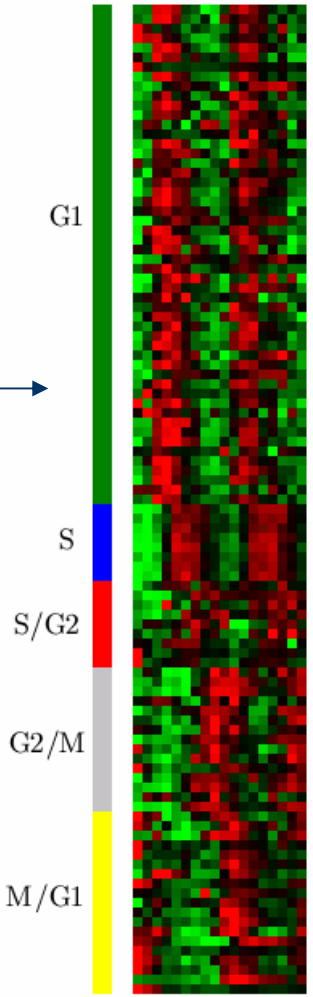
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

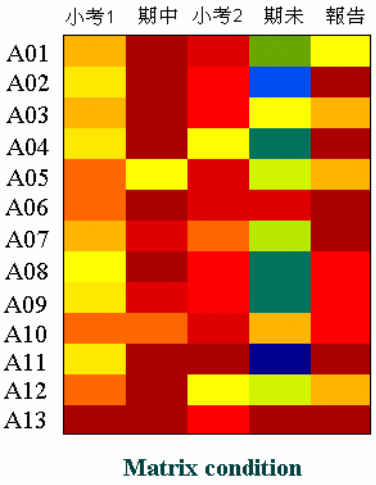


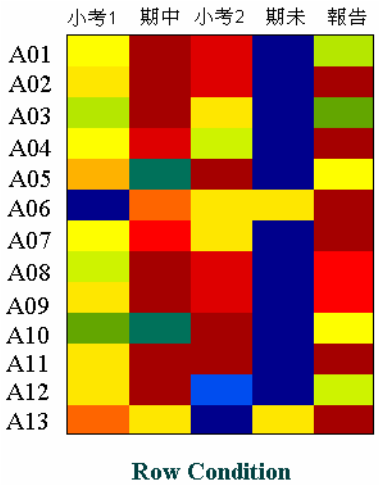
Microarray Data of Yeast Cell

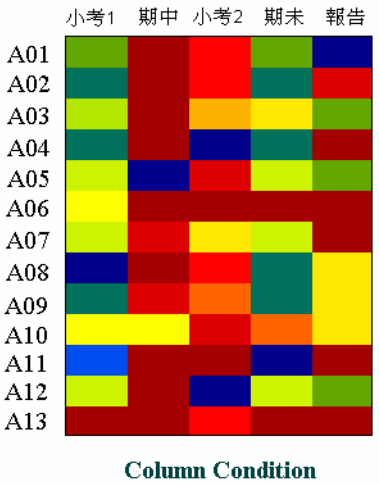
- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points.











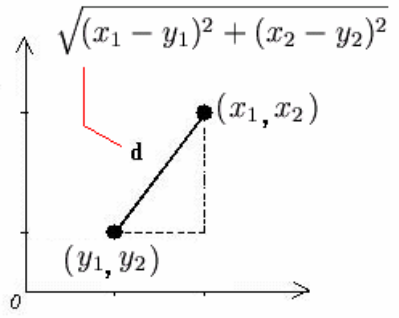
Clustering Analysis (Unsupervised Learning)

- Clustering is the representation of distance measurements between objects.
- The main goal of clustering is to use similarity or distance measurements between objects to represent them.
- Data points within a cluster are more similar, and those in separate cluster are less similar.
- ***Hierarchical clustering*** can be perform using agglomerative and divisive approaches. 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, in each subsequent step a cluster is split off, until there are n clusters.
  - **Agglomerative clustering:** all the objects start apart. There are n clusters at step 0, each object forms a separate cluster. In each subsequent step two clusters are merged, until only cluster is left.
- ***Non-Hierarchical clustering***

Distance and Similarity Measure

- The ***Euclidean distance*** of two points $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$ in Euclidean n-space is computed as

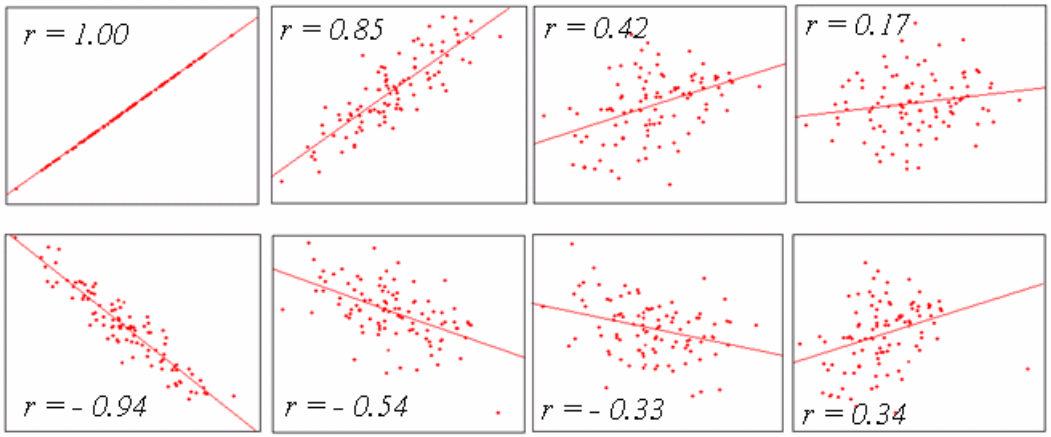
$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



- ***Pearson Correlation Coefficient***

the distance between two mRNA samples, with gene expression profiles $\mathbf{x} = (x_1, \ldots, x_p)$ and $\mathbf{x}' = (x'_1, \ldots, x'_p)$, is based on the correlation between their two gene expression profiles:

$$r_{\mathbf{x},\mathbf{x}'} = \frac{\sum_{j=1}^{p}(x_j - \bar{x})(x'_j - \bar{x}')}{\sqrt{\sum_{j=1}^{p}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{p}(x'_j - \bar{x}')^2}}.$$



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)

Dendrogram (Kaufman and Rousseeuw, 1990)

Hierarchical Clustering

Example:

Agglomerative algorithm + Average linkage clustering

	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

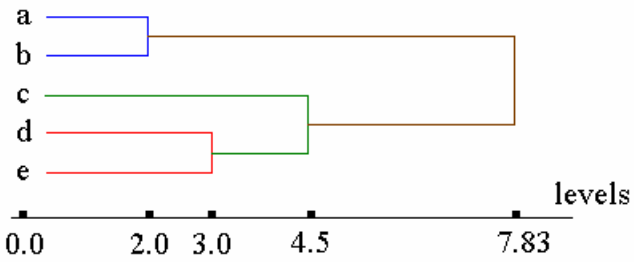
$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)] = \frac{1}{2}(6+5) = 5.5$$

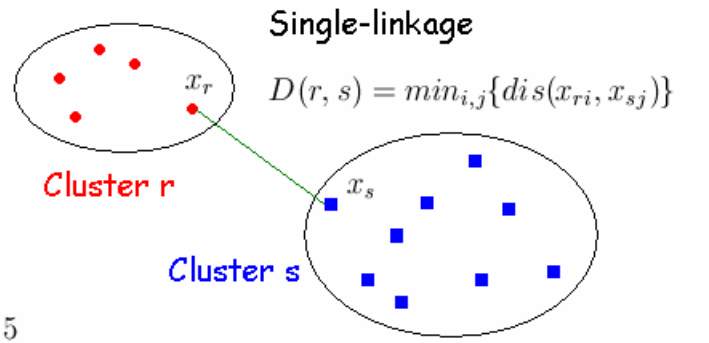
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

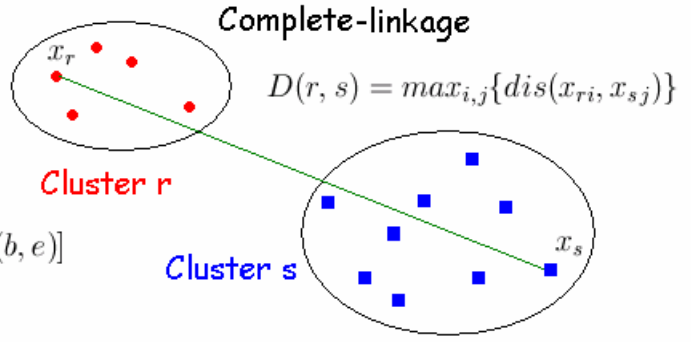
$$D(\{a,b\},\{d,e\}) = \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$
$$= \frac{1}{4}(10+9+9+8) = 9$$

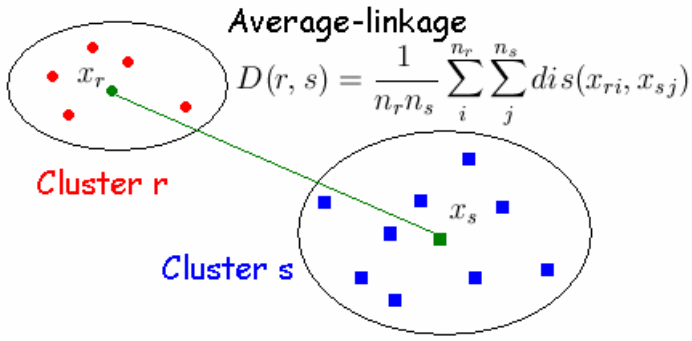
	{a,b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

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

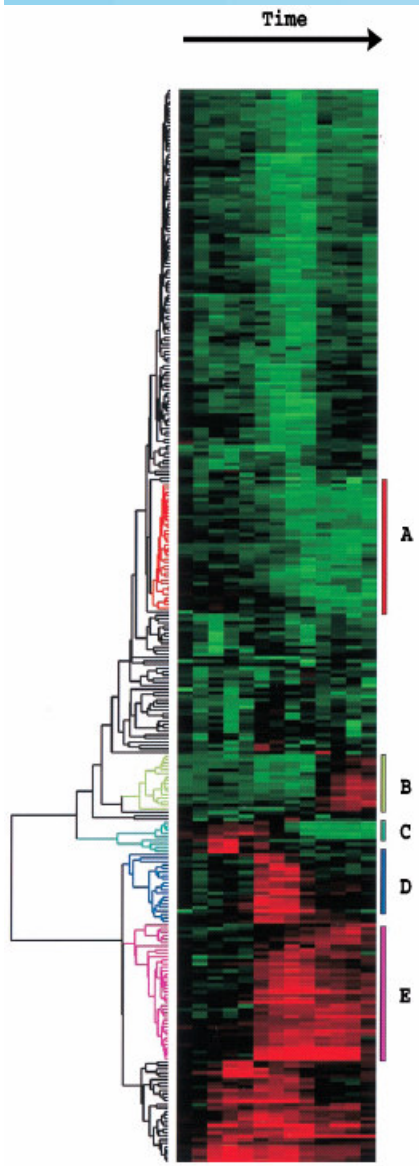








Hierarchical Clustering



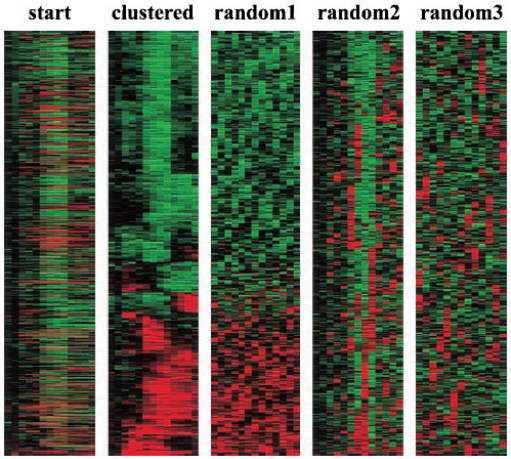
Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

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.

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-meansis 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

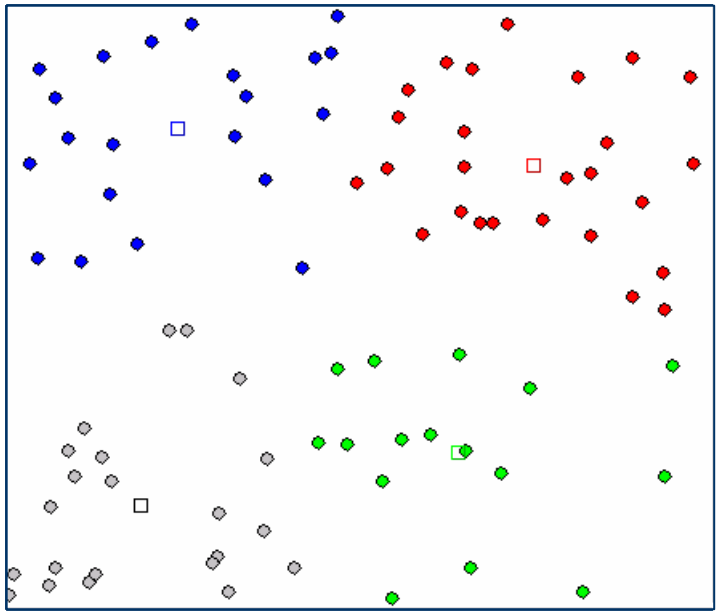
$$W(C) = \frac{1}{2}\sum_{k=1}^{K} \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

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.

Converged

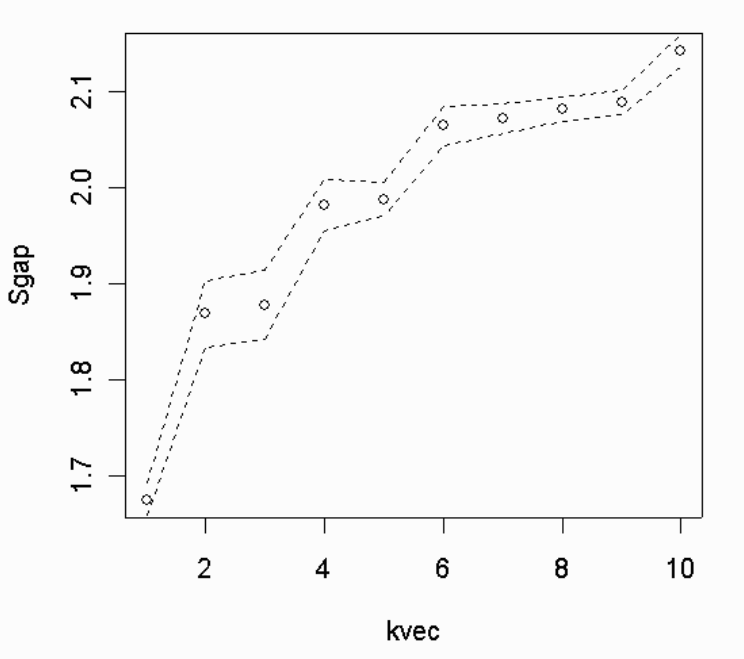


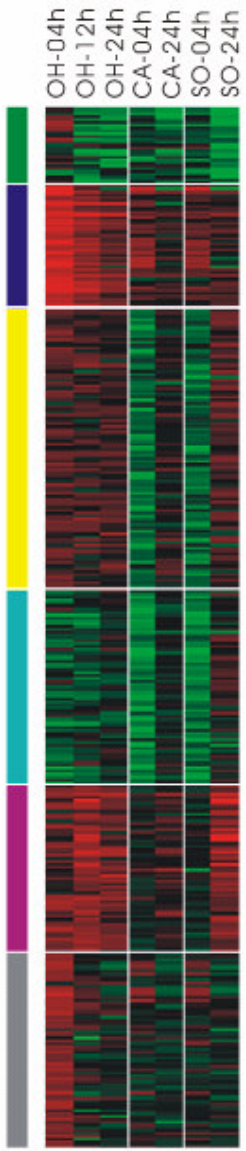
K-Means Clustering

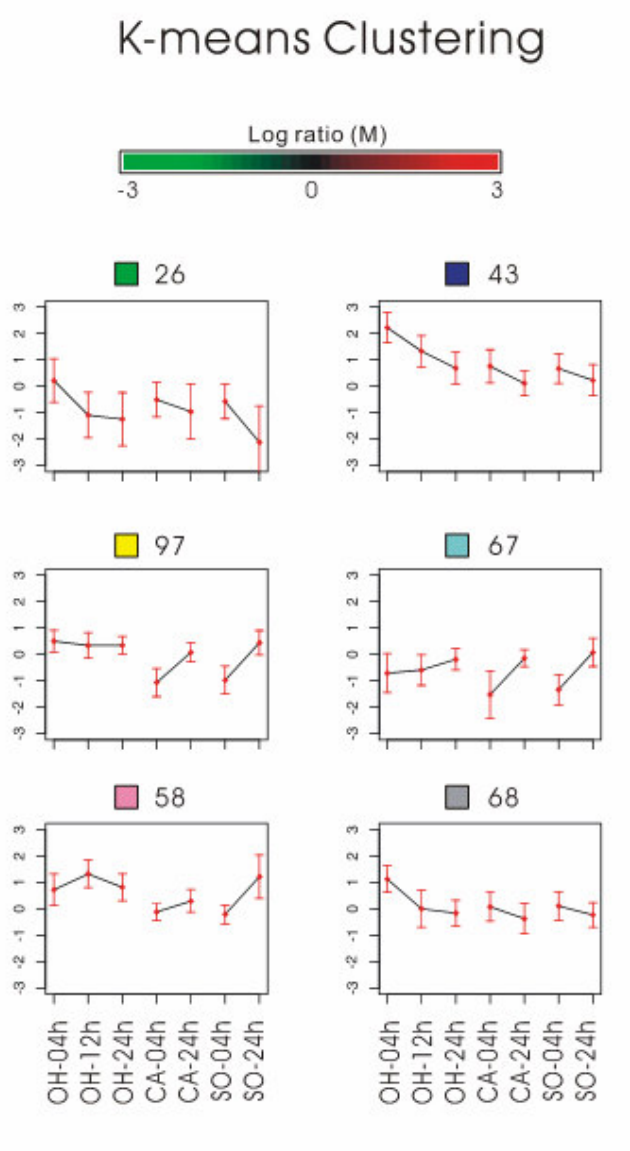
- Data

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.







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.

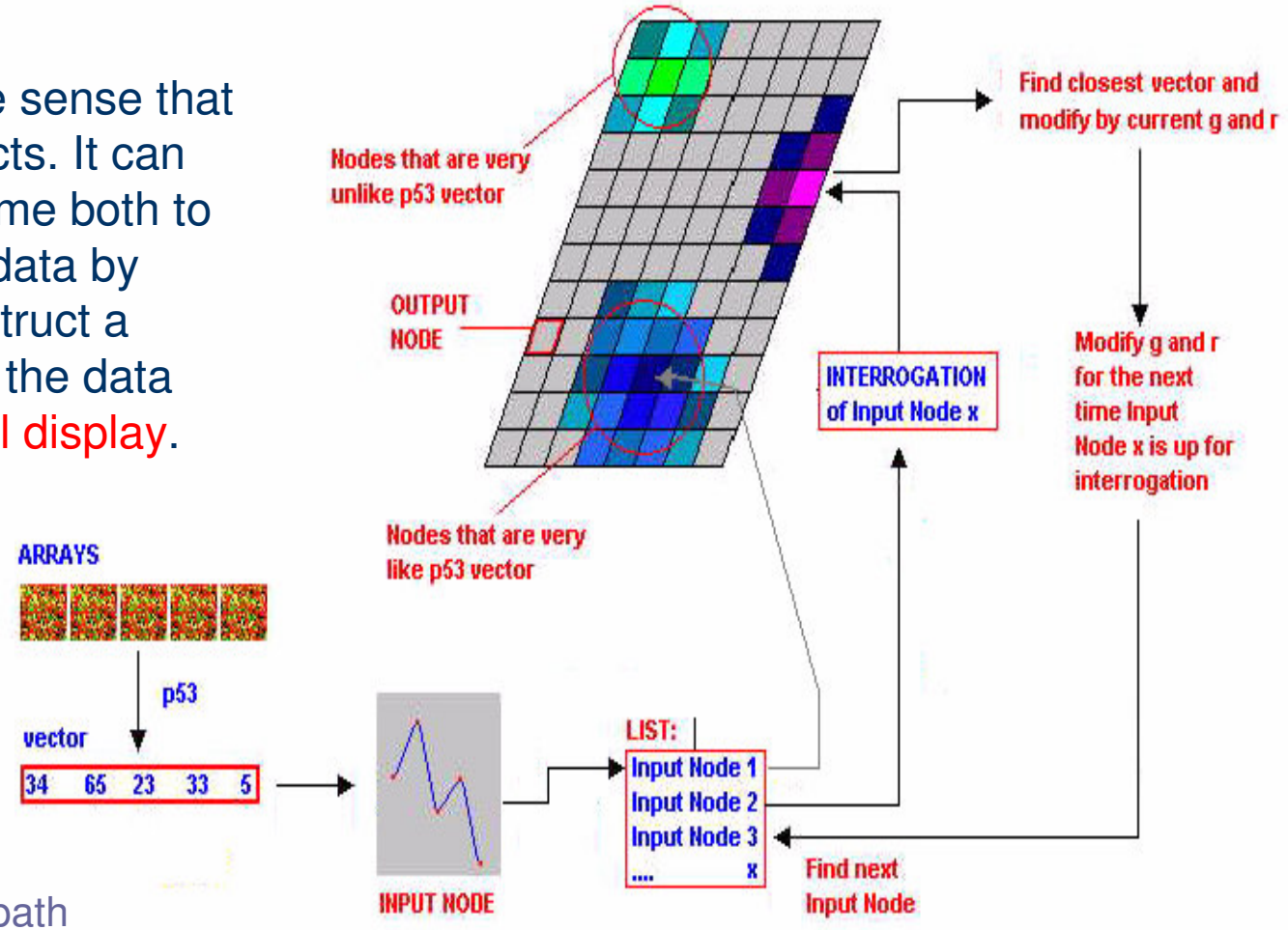


Figure modified from: *SCI*path

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

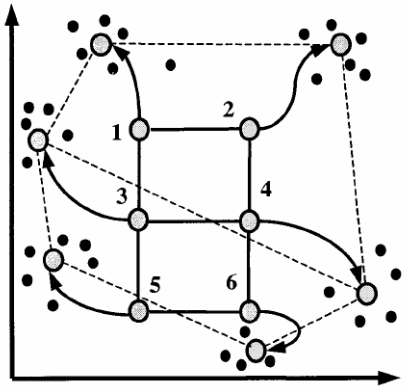
b. Learning process:

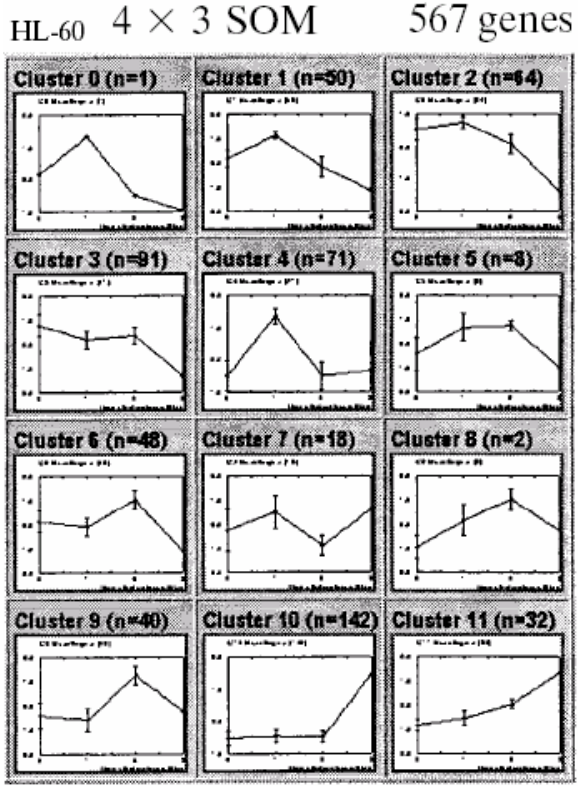
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

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.





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.

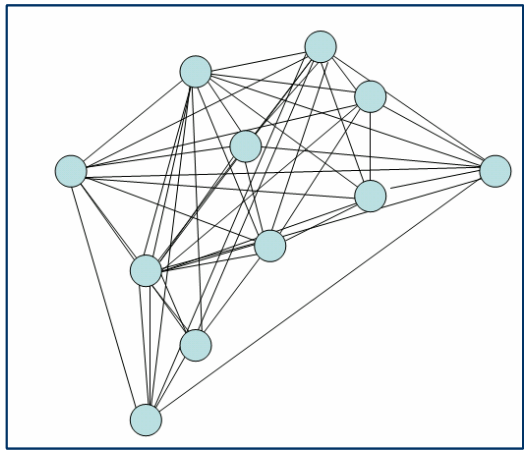
How Many Clusters?

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



Within-Cluster Sum of Squares

$$D_r = \sum_{i \in C_r} \sum_{j \in C_r} \| x_i - x_j \|^2$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

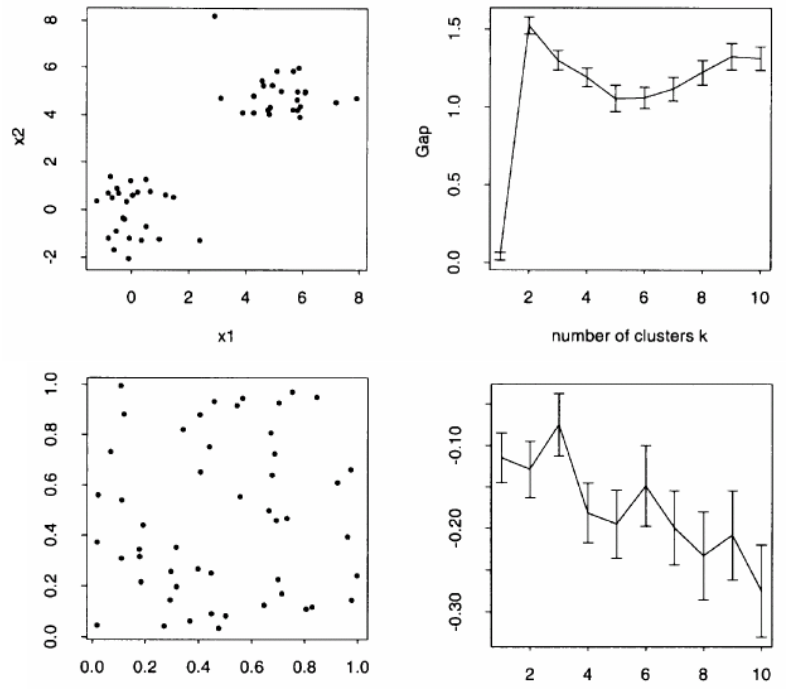
Calinski and Harabasz (1974): CH(k)
Hartigan (1975): $H(k)$
Krzanowski and Lai (1985): KL(k)
Kaufman and Rousseeuw (1990): $s(i)$

$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

Computational Implementation choose the number of clusters via

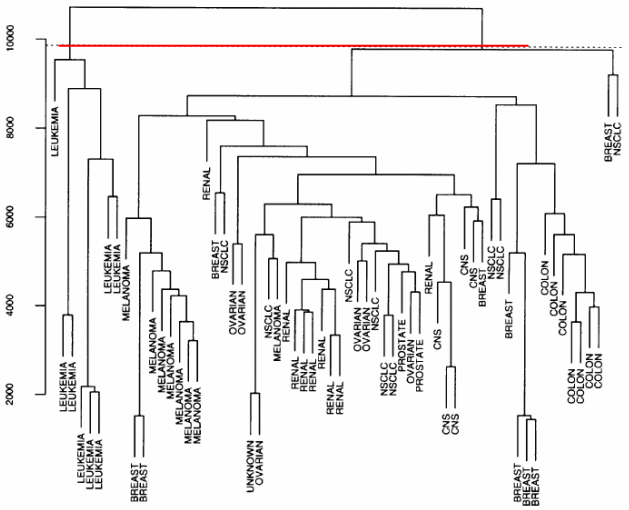
$$\hat{k} = \text{smallest } k \text{ such that}$$

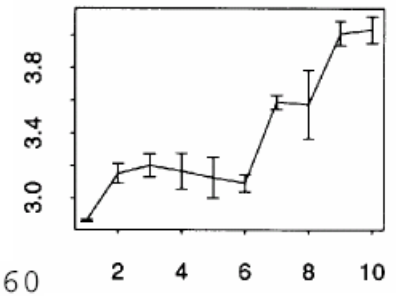
$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$$



application to hierarchical clustering and DNA microarray data

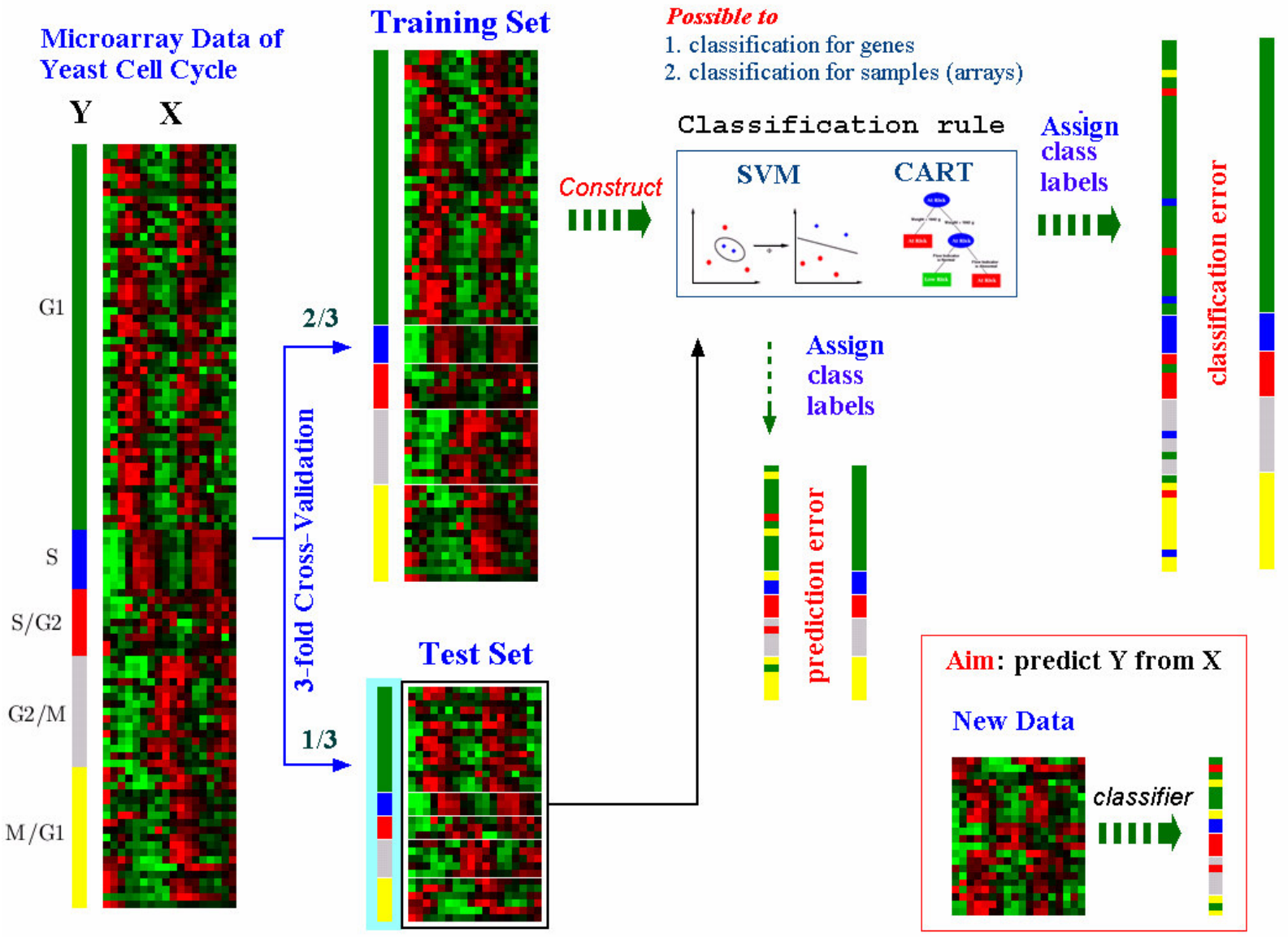
6834×64 matrix





http://www-genome.stanford.edu/nci60

Classification of Genes, Tissues or Samples (Supervised Learning)



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.

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

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.

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)

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

Classification Rules:

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

$d_k(\mathbf{x}) = \sum_{l=1}((\mathbf{x} - \bar{\mathbf{x}}_k)\mathbf{v}_l)^2$

denote its (squared) Euclidean distance, in terms of the discriminant variables, from the $1 \times p$ 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.

LDA for the Classification of Tumors

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

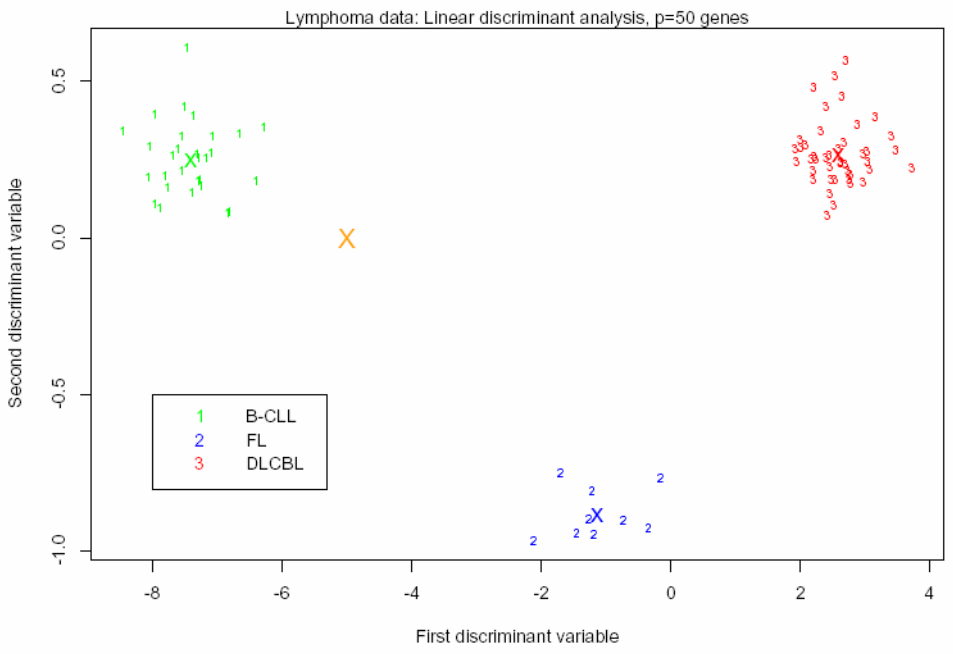
$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

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

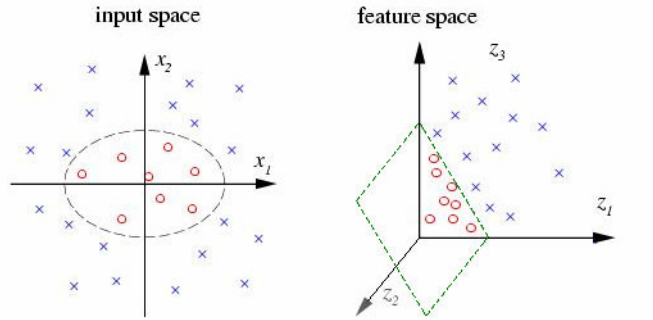
Support Vector Machine (SVM)

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).

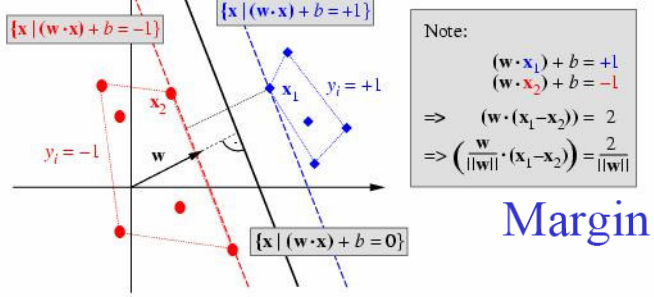
Support Vector Classifiers

$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$

$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$



Canonical Optimal Hyperplane



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

Support Vector Machine (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

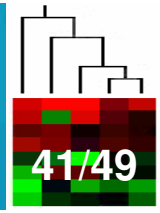
Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	Parzen	4	12	5	2,446	6
	FLD	9	10	7	2,441	5
	C4.5	7	17	0	2,443	−7
	MOC1	3	16	1	2,446	−1
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	7	7	23	2,430	39
	D-p 3 SVM	6	8	22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



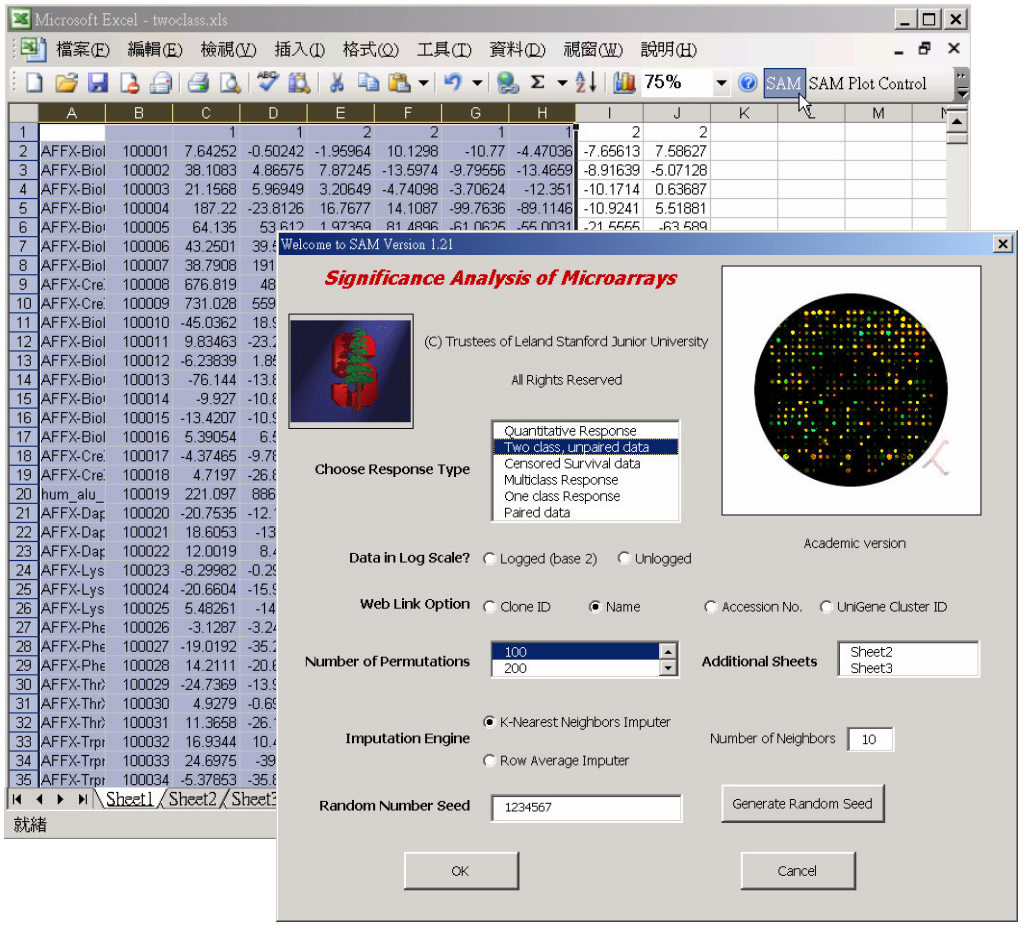
- *Freeware/Shareware*
  - Significance Analysis of Microarray (SAM)
  - Cluster and TreeView
  - The Bioconductor: limma, LImmaGUI
- *Commercial*
  - Matlab: Bioinformatics ToolBox
  - GeneSpring

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.

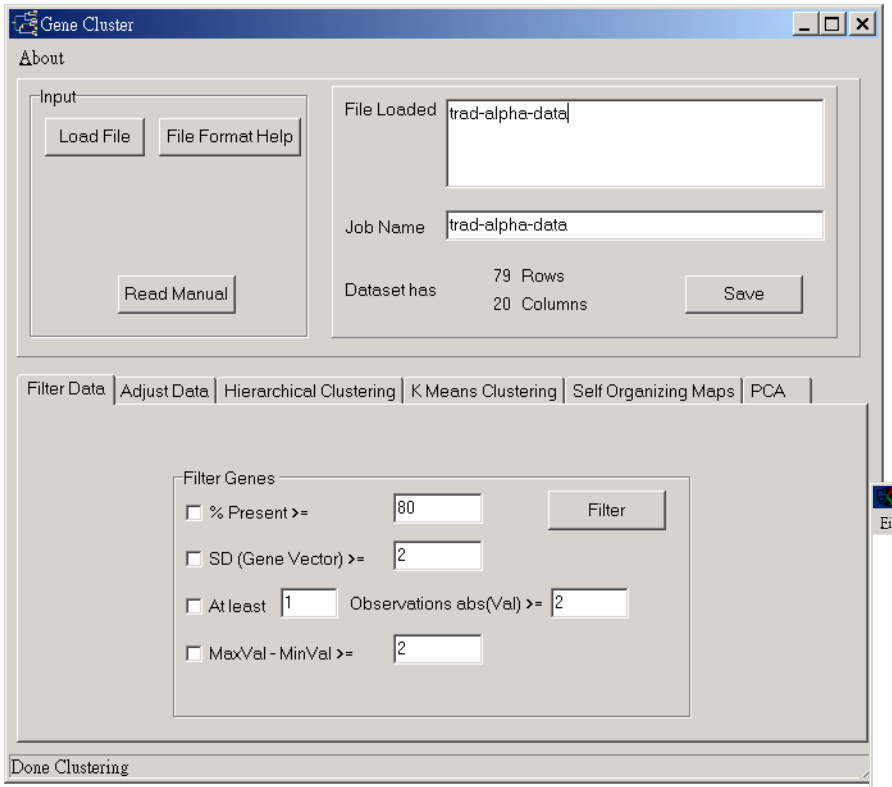
SAM does not do any normalization!



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/

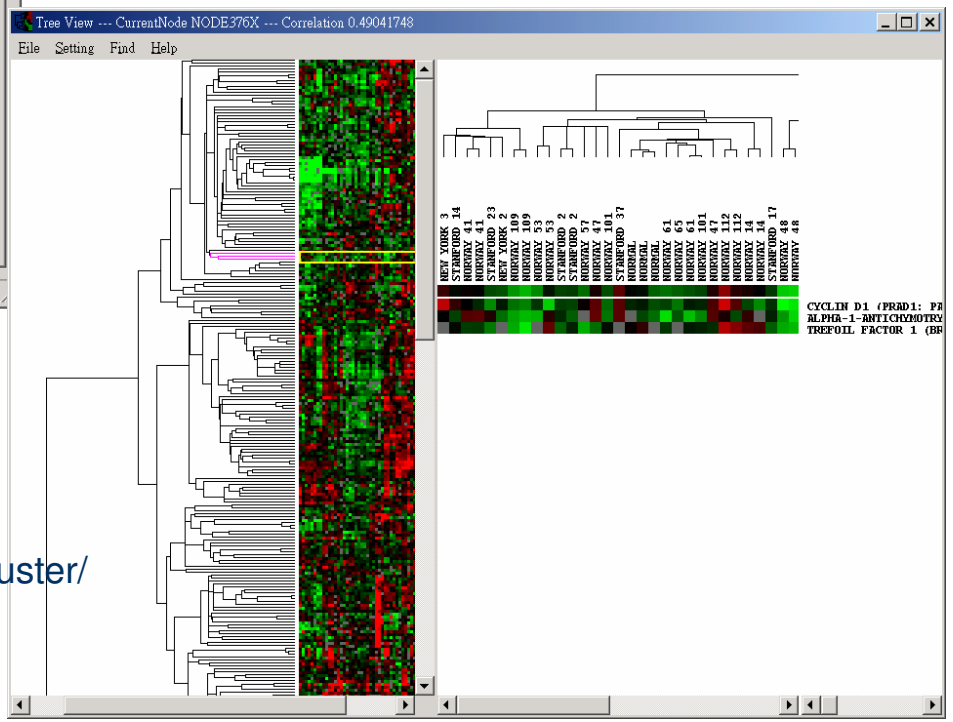
Cluster and TreeView



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)

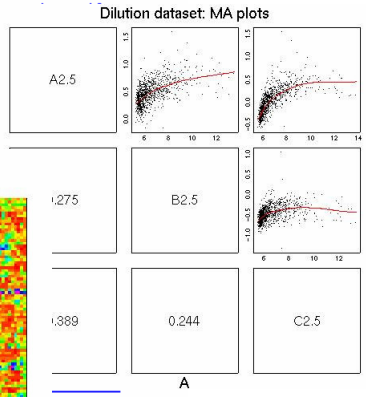
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

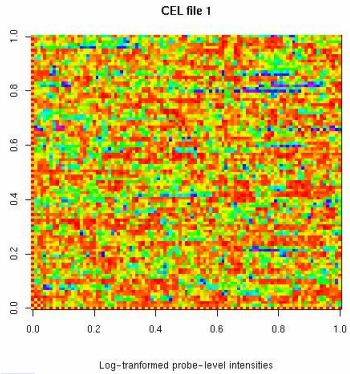
The Bioconductor

Package

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

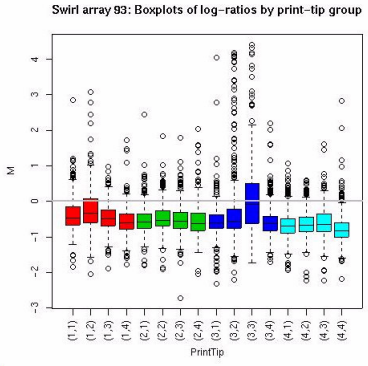
genefilter
geneplotter
globaltest
gpls
graph
hexbin
limma





daMA
edd
externalVector
factDesign
gcrma

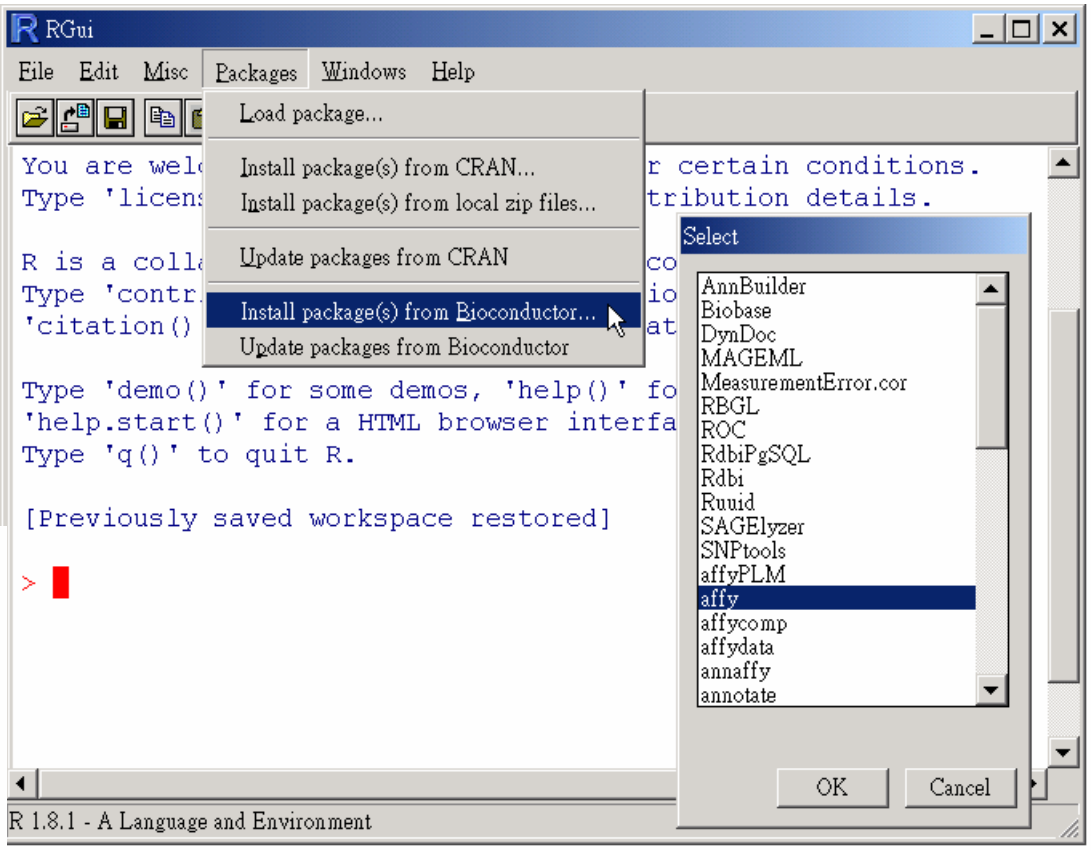
sigg
splic
tkW
vsn
wid



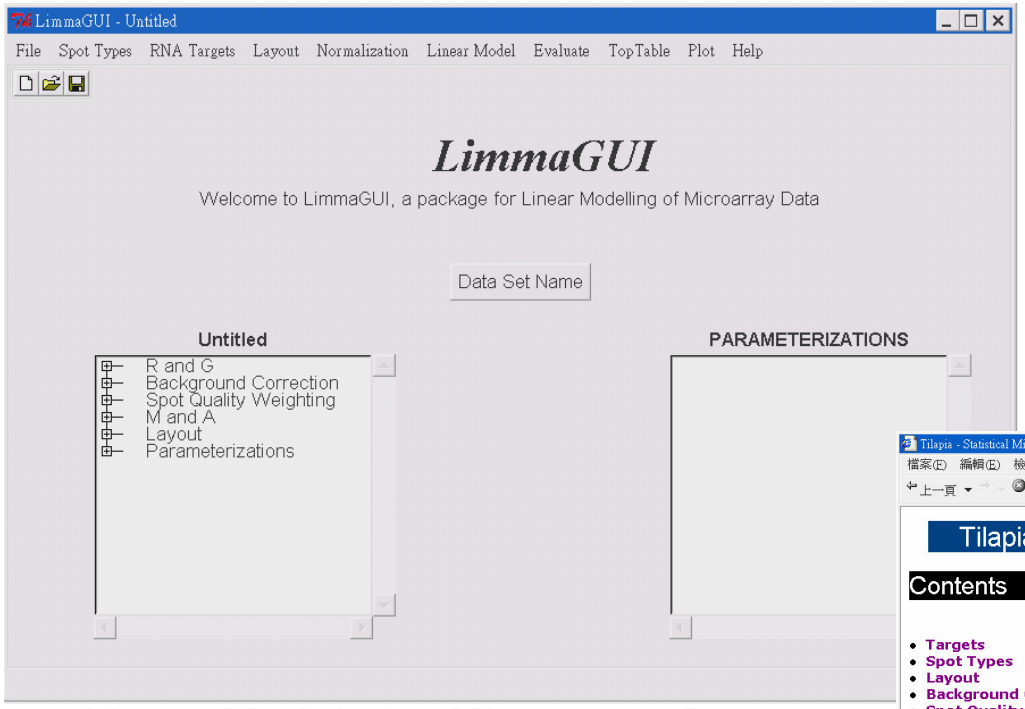
The Bioconductor
version 1.5 (2004-11-01)
http://www.bioconductor.org

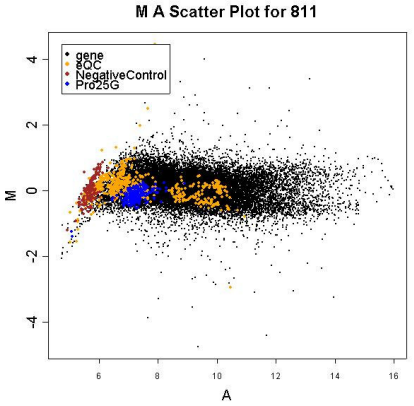
The R Project for Statistical Computing

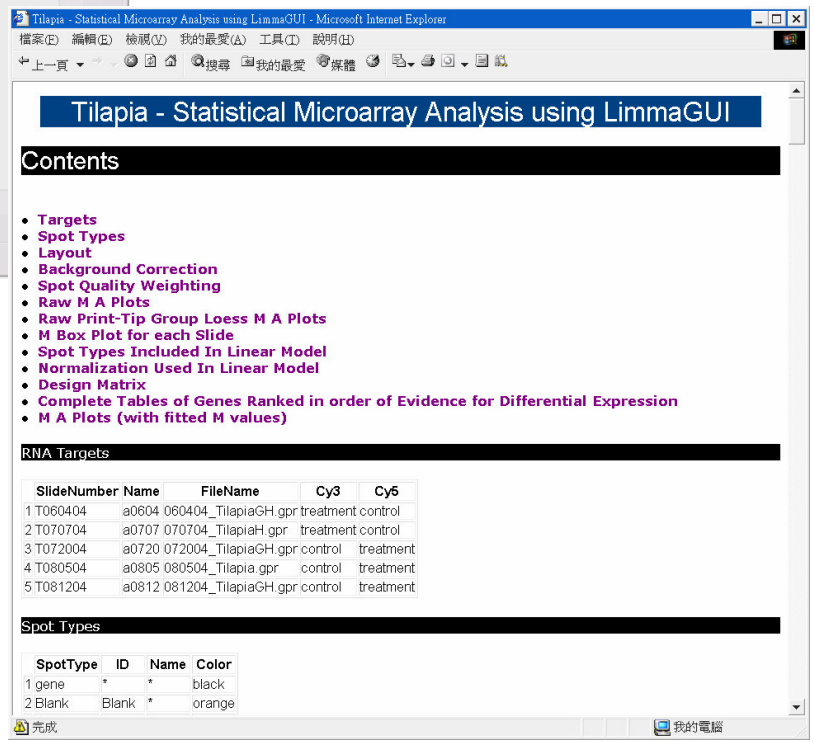
R version 2.1.0 (**2005-04-18**)
http://www.r-project.org



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

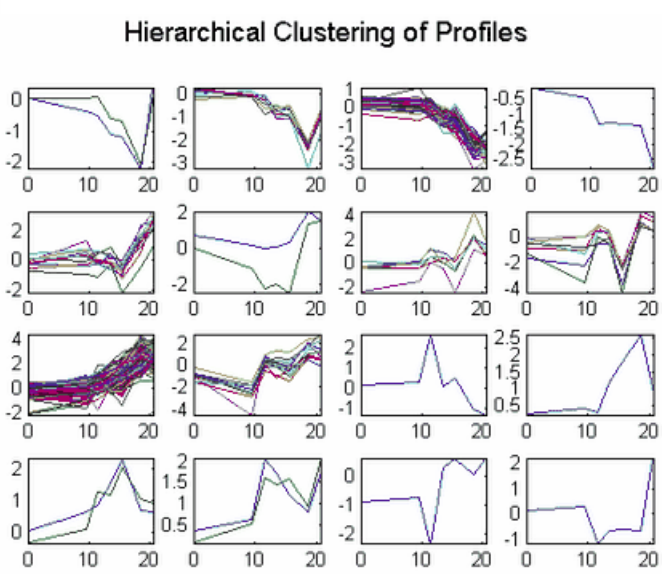
- Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Chapter 23. (To be published in 2005)
- Smyth, G. K. (2004). Linear models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

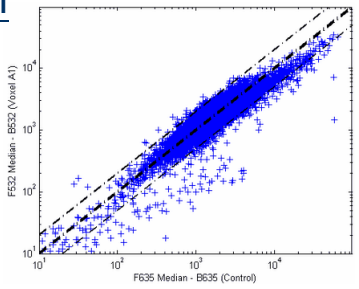
Matlab: Bioinformatics ToolBox

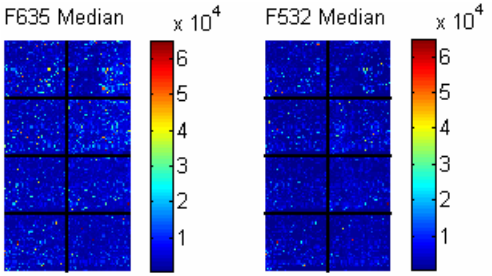
The MathWorks
Bioinformatics Toolbox

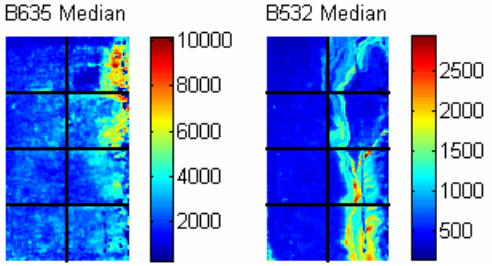
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html

- 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.





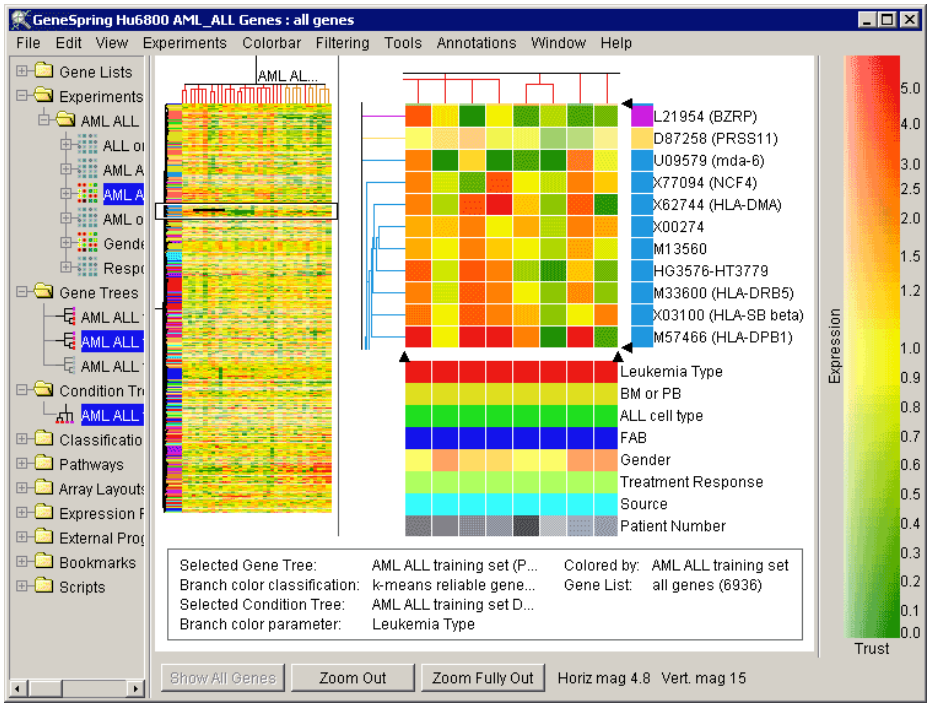


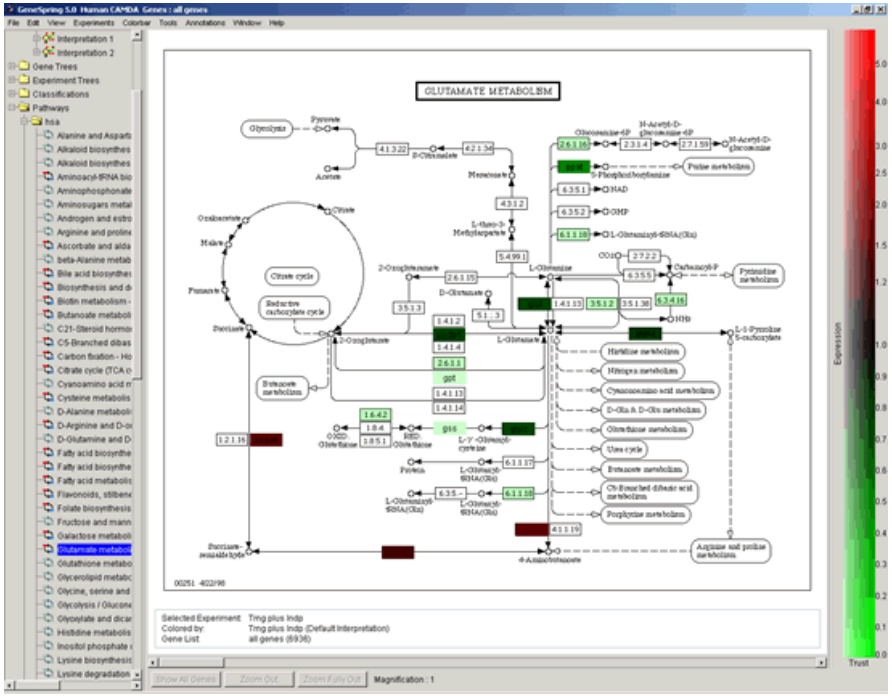


- Prototype and Development Environment — Create new algorithms, try new ideas, and compare alternatives.
- Share Algorithms and Deploy Applications — Create GUIs and stand-alone applications.

GeneSpring 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





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

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/

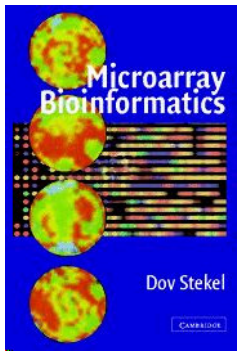
Bibliography on Microarray Data Analysis

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

BIOINFORMATICS

http://bioinformatics.oupjournals.org

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



BioConductor: open source software for bioinformatics

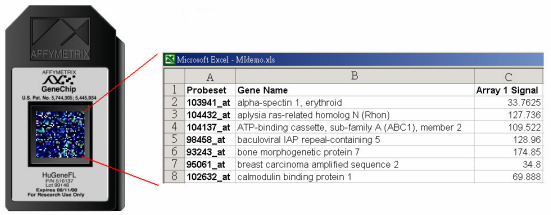
- Statistics and Genomics Short Course, Department of Biostatistics Harvard School of Public Health.
http://www.biostat.harvard.edu/~rgentlem/Wshop/harvard02.html

- Statistics for Gene Expression
http://www.biostat.jhsph.edu/~ririzarr/Teaching/688/

- Bioconductor Short Courses
http://www.bioconductor.org/workshop.htm

Other Related Issues

- Image Analysis
- Quality Measure (Array Quality, Spot Quality)
- Analysis of Replicates Arrays
- Time Series Samples
- Experimental Design
- Missing Values Imputation
- Analysis of Oligonucleotide Microarrays
- …



吳漢銘
E-mail: 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