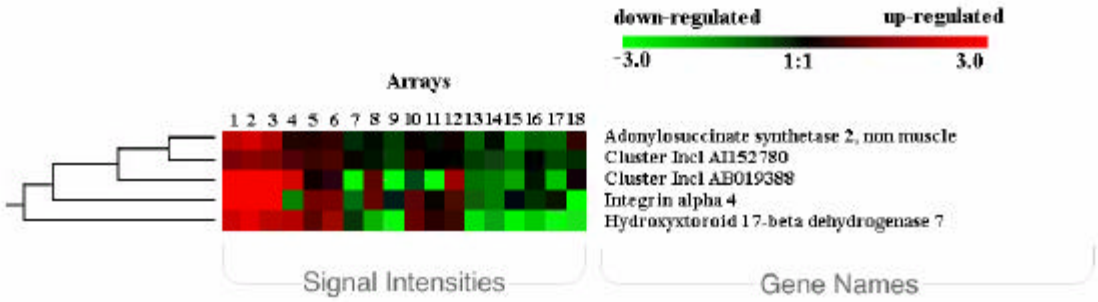
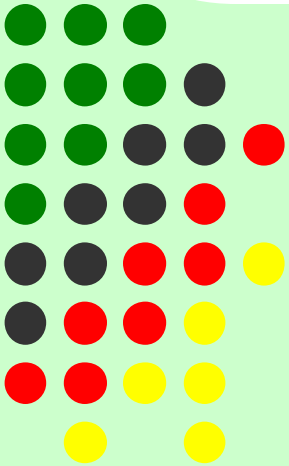


Microarray Data Analysis



國立臺灣師範大學生物資訊學程
生物資訊學導論

2004/4/28

陳君厚，吳漢銘

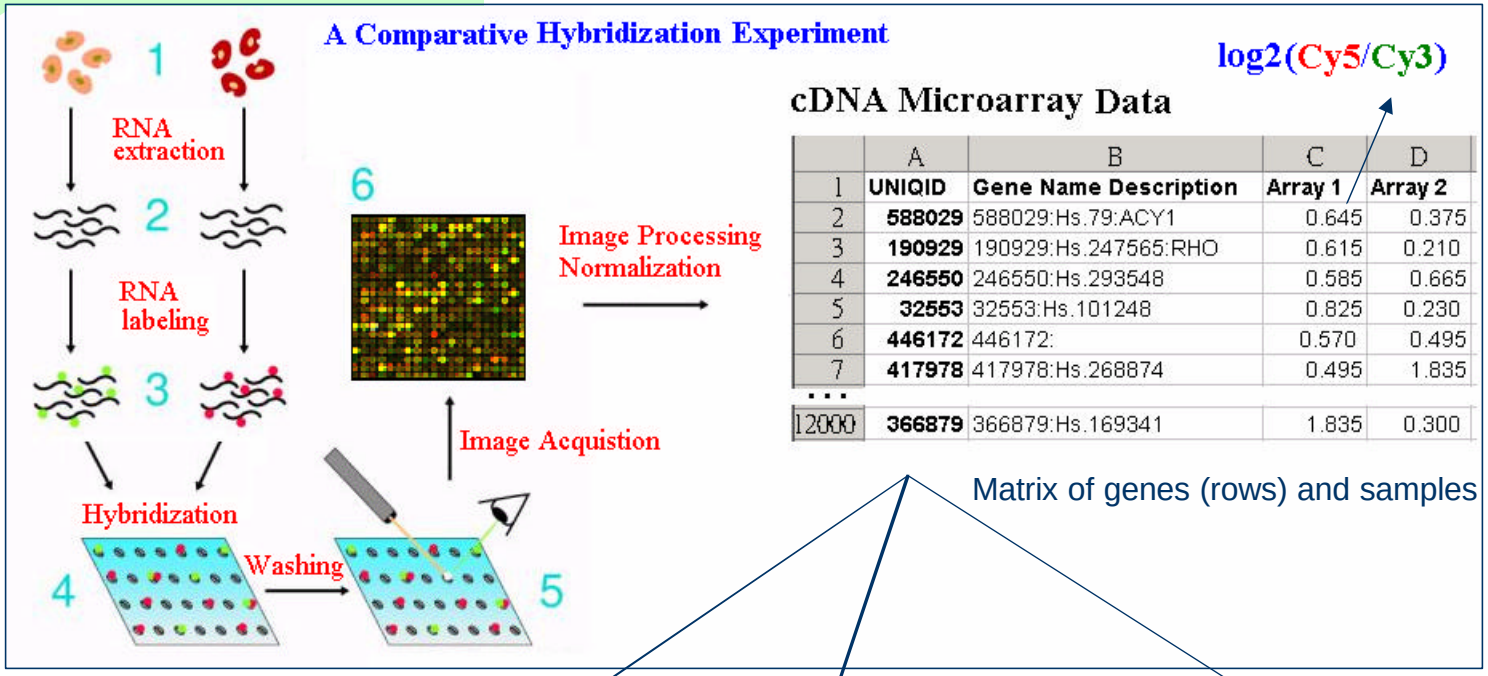
Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

Dimension Free
Data Visualization
GAP

Outlines

- Overview of Microarray Data Analysis
- *Preprocessing*: **Image Processing**, Normalization
- *Finding Differentially Expressed Genes*
  - Fold Changes Method
  - Inferential Statistics (Hypothesis Testing): t-test, two-sample t-test
  - Multiple-testing Problem
- *Visualization Methods* (Exploratory)
  - Principal Components Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - Biplot
- *Analysis of Relationship Between Genes, Tissues or Treatments*
  - Dendrogram, HeatMap and Hierarchical Clustering
  - K-Means Clustering
  - Self-Organizing Maps (SOM)
- *Classification of Genes, Tissues or Samples:*
  - Linear Discriminant Analysis (LDA)
  - Support Vector Machines (SVM)
- Annotations in Microarray Data
- *Software*: Significance Analysis of Microarray (SAM), Cluster and TreeView, Bioconductor, Matlab: Bioinformatics ToolBox

Overview

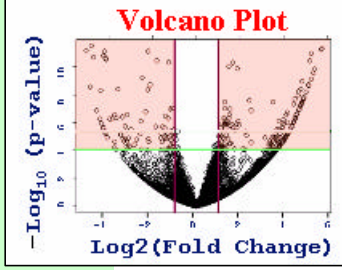


	A	B	C	D
1	UNIQID	Gene Name Description	Array 1	Array 2
2	588029	588029:Hs.79:ACY1	0.645	0.375
3	190929	190929:Hs.247565:RHO	0.615	0.210
4	246550	246550:Hs.293548	0.585	0.665
5	32553	32553:Hs.101248	0.825	0.230
6	446172	446172:	0.570	0.495
7	417978	417978:Hs.268874	0.495	1.835
...				
12000	366879	366879:Hs.169341	1.835	0.300

Discovery of differentially expressed genes

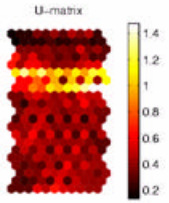
Parametric : t-test

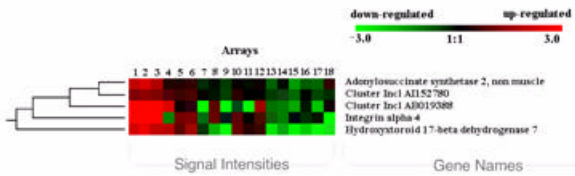
Non-parametric : Wilcoxon, Mann-Whitney test



Unsupervised: clustering

- Hierarchical clustering
- K-means clustering
- Self-organizing maps

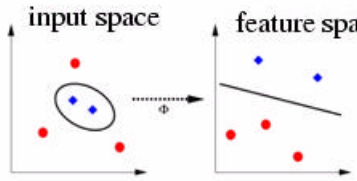




Supervised: classification

- Linear discriminants
- Decision trees
- Support vector machines

Support Vector Classifiers



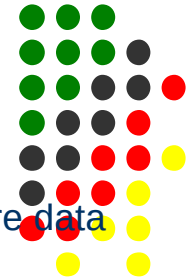
Boser, Guyon, and Vapnik (1992)

DNA Microarray Experiments

- ***Question:***
  - Which genes were most dramatically up or down regulated in the experiment? (Inferential Statistics)
  - What signatures/patterns/profiles of gene expression can be found in all the gene expression values obtained in the experiment? (Descriptive Statistics or Exploratory Analysis)
- ***Replicates:***
  Microarray experiments involve the measurement of the expression levels of many thousands of genes in only a few biological samples.
  - few technical replicates: measuring gene expression with the same starting material on independent arrays.
  - few biological replicates: measuring gene expression from multiple cell lines, each of which has been given an experimental treatment or control treatment.
- The gene expression values are ratio or relative intensities: ratios of intensity values for the **Cy3 (green)** dye and the **Cy5 (red)** dye.
- The intensity of each signal is **assumed** to be directly proportional to the abundance of mRNA for each gene.

Normalization

Assume Microarray Image has been processed appropriately.



- **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.
- **Factors that may contribute to variation include**
  - Different labeling efficiencies of fluorescently labeled nucleotides.
  - Technical artifacts: print tips, uneven spotting.
  - Variations in the performance of a fluorescence scancer or phosphorimager.
  - Variations in RNA (or mRNA) purity.
  - Variation in the way the RNA is purified, labeled and hybridized.
  - Variation in the way the microarray is washed to remove non-specific binding.
  - Variations in the way the signal is measured.
- **Assumption**: the average gene does not change in its expression level in the biological sample being tested.

Normalization Methods: constant (Global Normalization)

- The Cy3 and Cy5 are incorporated into cDNA with different efficiencies: without normalization, it would not be possible to accurately assess the relative expression of sample that are labeled with those dyes; genes that are actually expressed at comparable levels would have a ratio different than 1.
- The background intensity signal is measured and subtracted from the signal for each gene.
- Global normalization: the average ratio for gene expression is 1.
- Housekeeping Genes: beta-actin, GAPDH.
  - Each gene expression value in a single array experiment is divided by the mean expression values of these housekeeping genes.
  - assumption: such genes do not change in their expression values between two conditions.
  - Human Gene Expression (HuGE) database: 7000 gene in 19 tissues: 451 housekeeping genes are commonly expressed across all theses tissues.

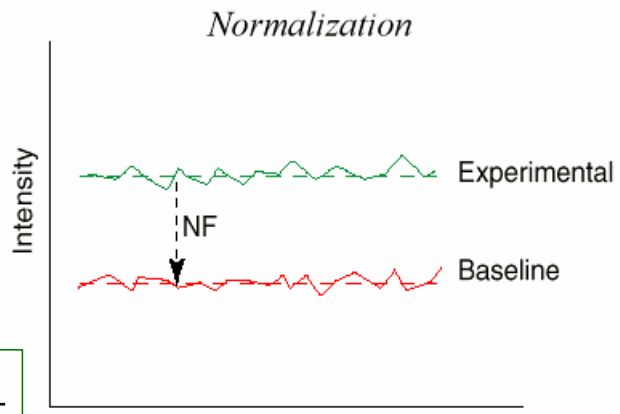
Normalization Methods: constant

Normalization and Scaling

- The data can be normalized from:
  - a limited group of probe sets.
  - all probe sets.

$$A \times SF = TGT$$

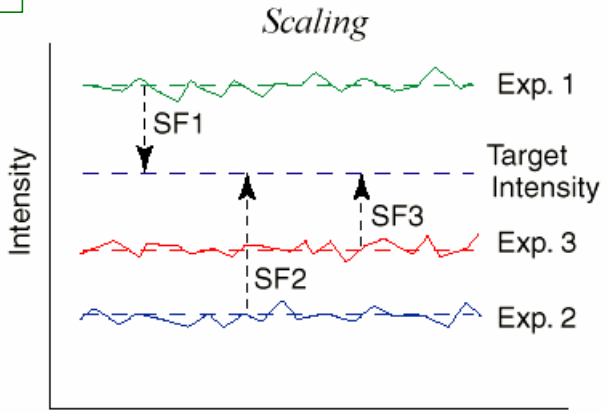
$$\Rightarrow SF = \frac{TGT}{A}$$



- **Global Scaling**
  the average intensities of all the arrays that are going to be compared are multiplied by scaling factors so that all average intensities are made to be numerically equivalent to a preset amount (termed target intensity).

$$SF = \frac{TGT}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

- **Global Normalization**
  the normalization of the array is multiplied by a Normalization Factor (NF) to make its Average Intensity equivalent to the Average Intensity of the baseline array.



$$A_{exp} \times NF = A_{base}$$

$$\Rightarrow NF = \frac{A_{base}}{A_{exp}}$$

$$nf = \frac{TrimMean\left(SPVb_{i,}\ 0.02, 0.98\right)}{TrimMean\left(SPVe_{i,}\ 0.02, 0.98\right)}$$

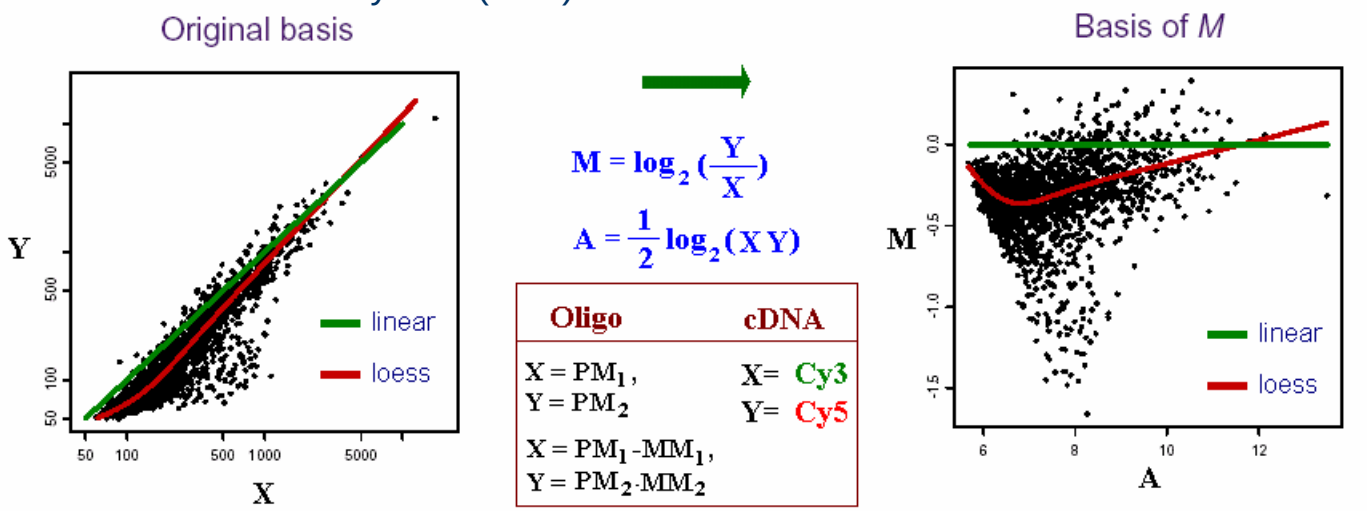
Average intensity of an array is calculated by averaging all the Average Difference values of every probe set on the array, excluding the highest 2% and lowest 2% of the values.

Scatter plot 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.

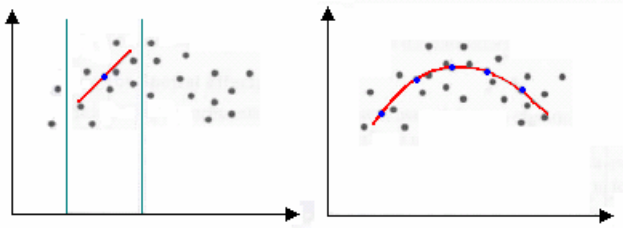
- x-axis (mean log2 intensity): average intensity of a particular element across the control and experimental conditions.
- y-axis (ratio): ratio of the two intensities.



Normalization Methods: loess

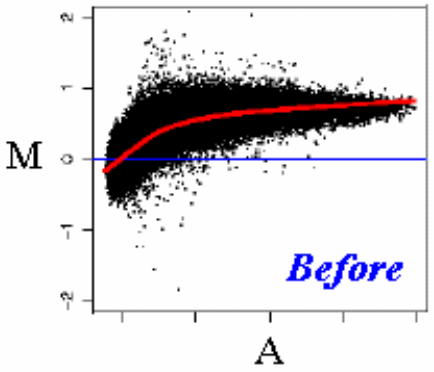
- Loess normalization (Bolstad *et al*., 2003) is based on M versus A plots. Two arrays are normalized by using a lowess smoother.
- **Skewing** reflects experimental artifacts such as the
  - contamination of one RNA source with genomic DNA or rRNA,
  - the use of unequal amounts of radioactive or fluorescent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

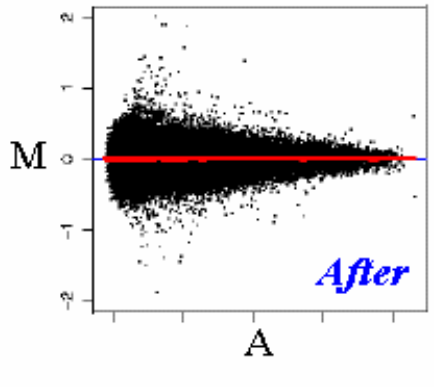
Loess regression
(locally weighted polynomial regression)



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.

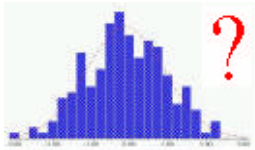
   Loess is a method of local regression (see Cleveland and Devlin (1988) for details).
4. The fits based on the normalization curve are $\hat{M}_k$
5. the normalization adjustment is $M'_k = M_k - \hat{M}_k$.
6. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k + \frac{M'_K}{2}}$ and $x'_{ki} =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





Inferential Statistics (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

Fold Change Methods

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.

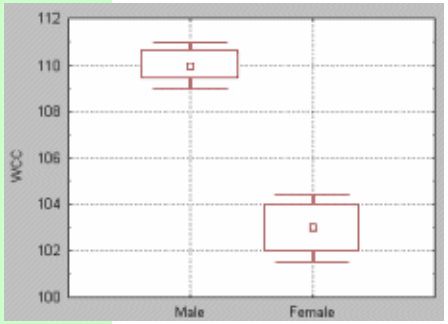
An Example

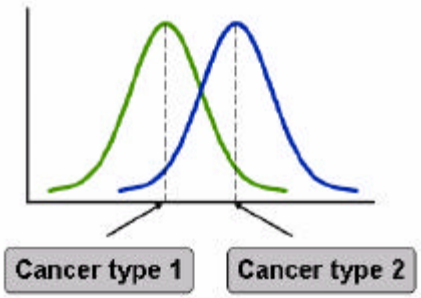
- A *hypothesis test* is a procedure for determining if an assertion about a characteristic of a population is reasonable.
- For example, suppose that someone says that the average price of a gallon of regular unleaded gas in Massachusetts is $1.15. How would you decide whether this statement is true?
  - 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)

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.

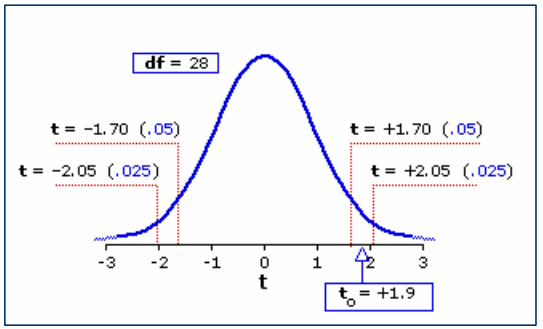
- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ: $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

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.



t-test: Testing for Differences Between Two Groups

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

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.

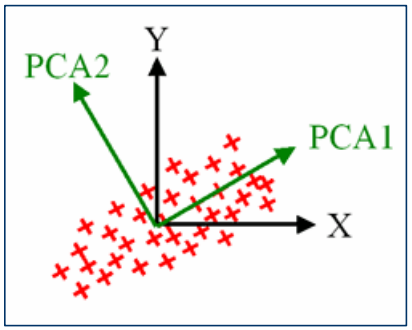
- ***Example:*** In Breast Cancer Dataset with 9216 genes, even if the chemotherapy had no effect whatsoever, we expect to find 92 differentially expressed genes with p-values less than 0.01, simple because of the large number of genes being analyzed.

- Bootstrap Analysis
- Permutation Test

How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just artifact introduced because we are analyzing a large number of genes? Is this gene truly differentially expressed, or could it be a false positive results?

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.

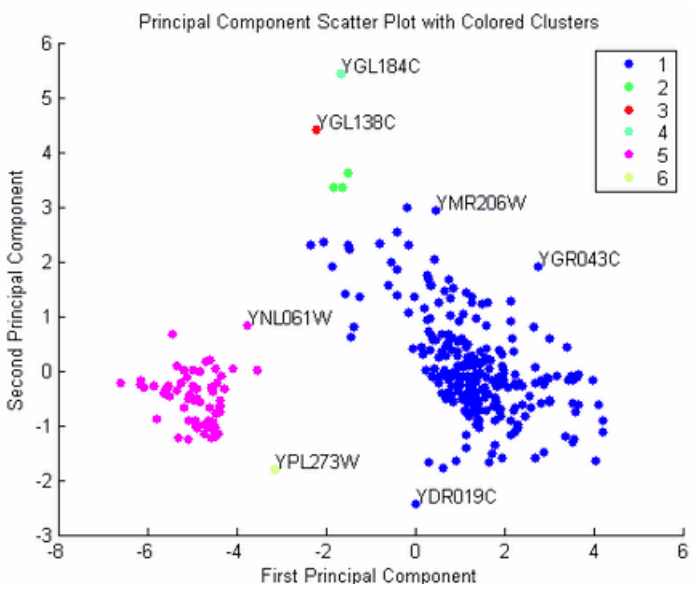
The PCA summaries the dispersion of data points as data cloud in a small number of major axes (principal components) of varition among the variables.

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

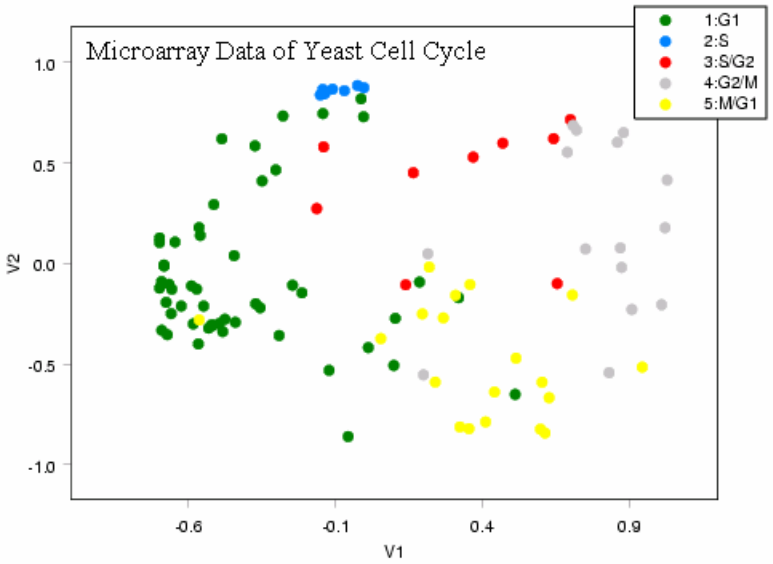
Multidimensional scaling takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.

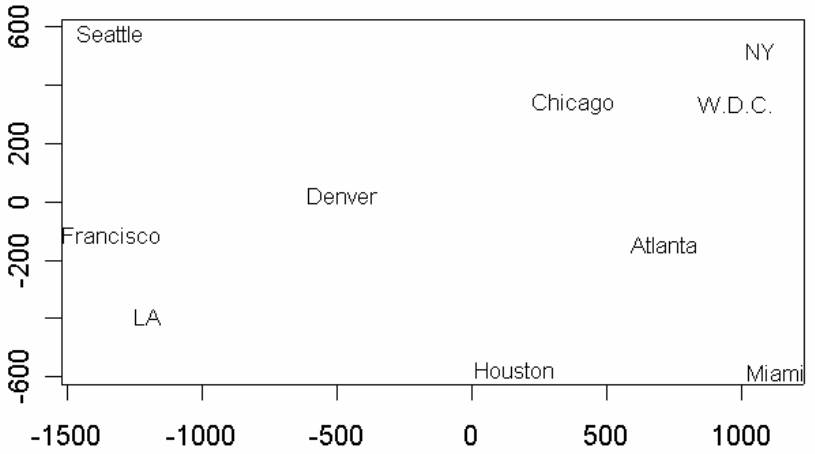
Note that if the input-space distances are Euclidean, classical MDS is equivalent to PCA. (Mardia et al. 1979)

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





The Biplot (Gabriel 1971, 1981; Gower & Hand, 1996)

The data matrix can be factored:

$$\mathbf{X} = \mathbf{AB}'$$

$\mathbf{X}_{n\times p}$: data matrix.
$\mathbf{A}_{n\times k}$: the coordinates for the n observations points along k rectangular axes.
$\mathbf{B}_{p\times k}$: the coordinates for the p variables along the same k axes.

To obtain $\mathbf{A}$ and $\mathbf{B}$, using Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{UDV}'$$

$\mathbf{A}_{[2]}$: the $n \times 2$ matrix of biplot coordinates for the observation points.
$\mathbf{B}_{[2]}$: the $p \times 2$ matrix of biplot coordinates for the variables.

$$\mathbf{A}_{[2]} = \mathbf{U}_{[2]}\mathbf{D}^{c}_{[2]}$$

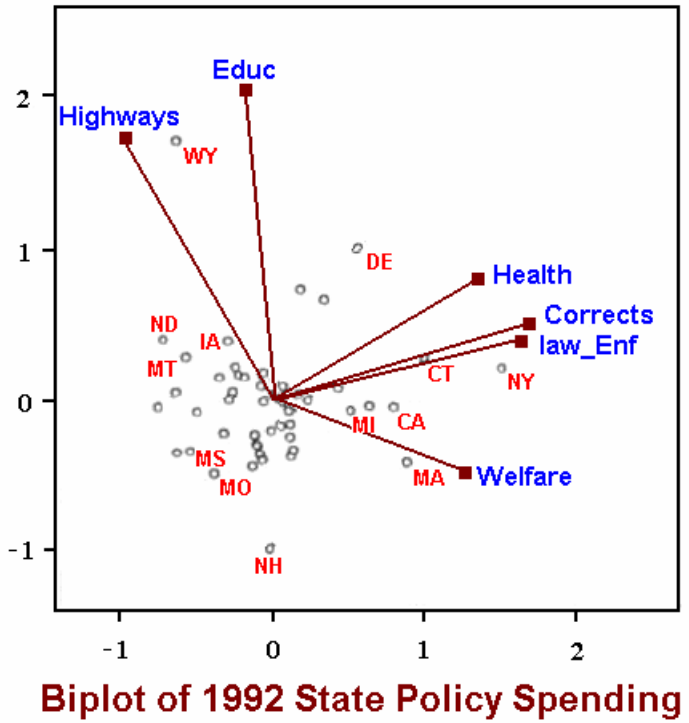
$$\mathbf{B}_{[2]} = \mathbf{V}_{[2]}\mathbf{D}^{1-c}_{[2]}$$

$\mathbf{U}_{[2]}$: the first two columns of $\mathbf{U}$.
$\mathbf{V}_{[2]}$: the first two columns of $\mathbf{V}$.
$\mathbf{D}_{[2]}$: the diagonal matrix formed by the first two singular values.

$$\mathbf{X}_{[2]} = \mathbf{A}_{[2]}\mathbf{B}'_{[2]}$$

Each row of $\mathbf{A}_{[2]}$ is plotted as a point in a two-axis coordinate system.
The rows of $\mathbf{B}_{[2]}$ are also plotted within the same space.
Goodness of fit measure R (s_r : singular values)

$$R = \frac{s_1^2 + s_2^2}{\sum_{r=1}^{p} s_s^2}$$



Biplot of 1992 State Policy Spending

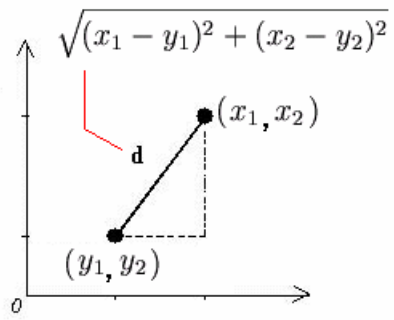
Clustering Analysis

- 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.

Distance and Similarity Measure

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

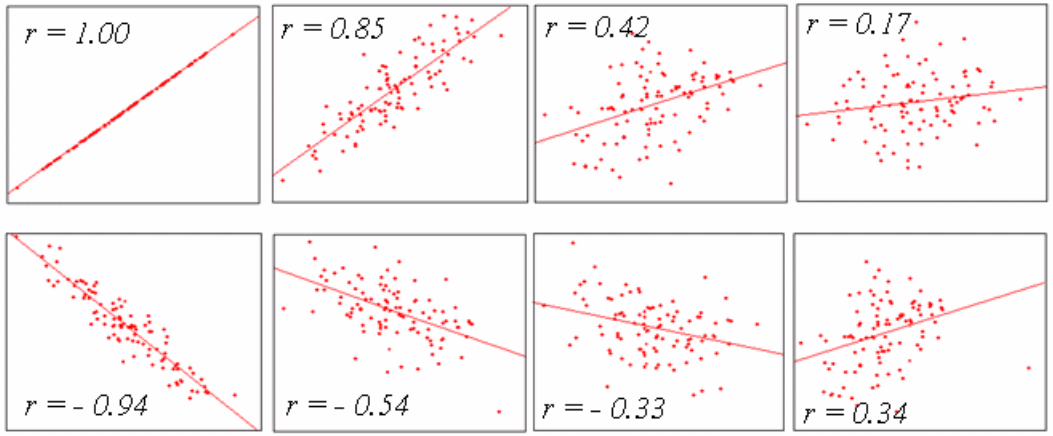
$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



- ***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:

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



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

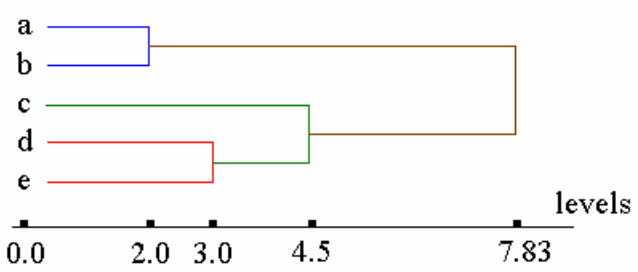
$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c)+D(b,c)] = \frac{1}{2}(6+5) = 5.5$$

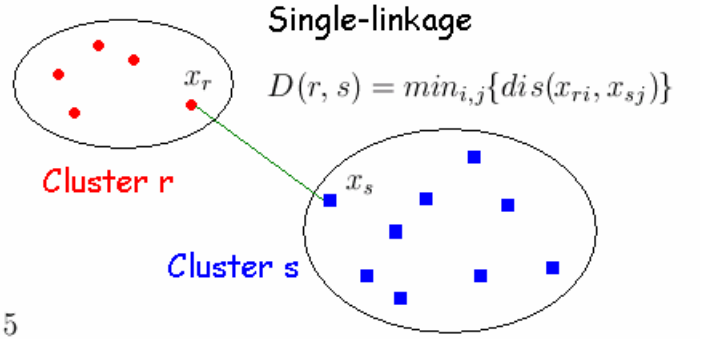
	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

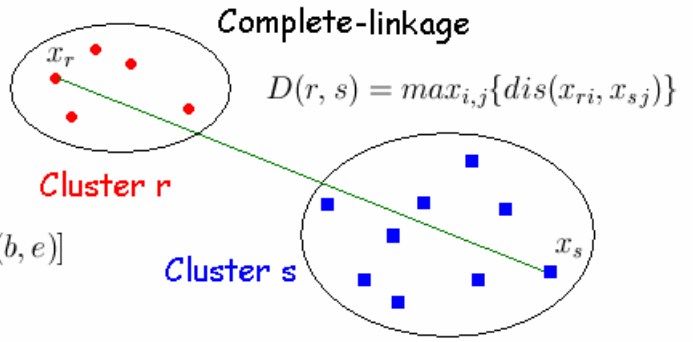
$$D(\{a,b\},\{d,e\}) = \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$
$$= \frac{1}{4}(10+9+9+8) = 9$$

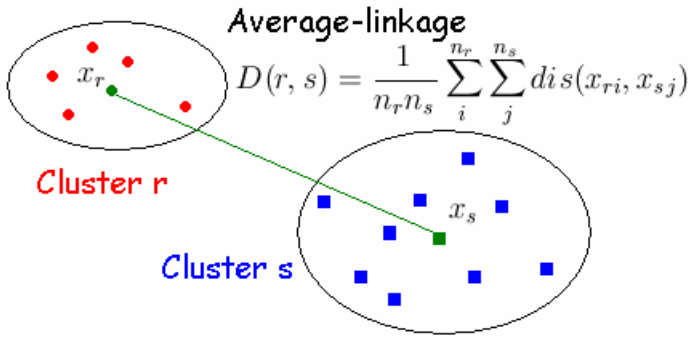
	{a,b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

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



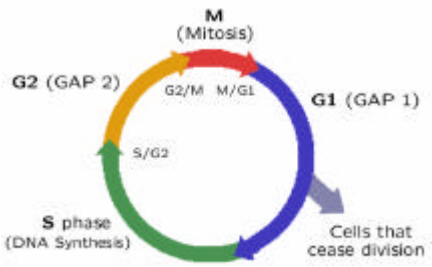






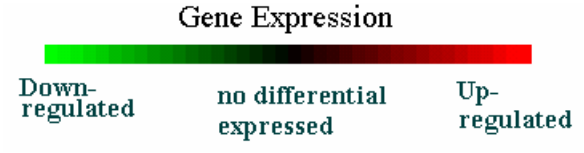
Heat Map
(Data Image, Matrix Visualization)

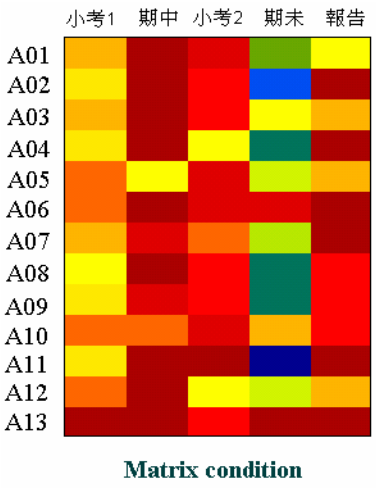
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92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

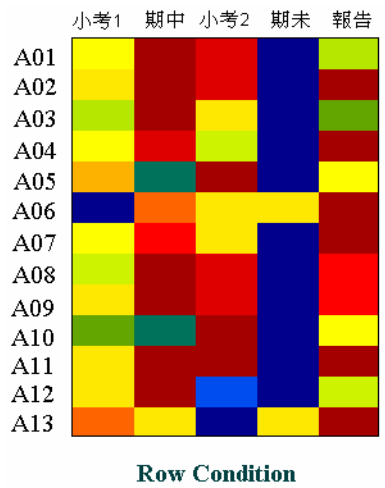


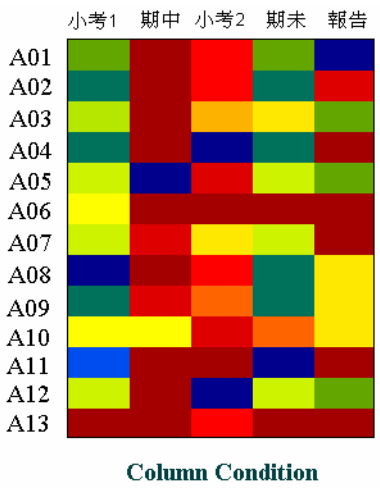
Microarray Data of Yeast Cell

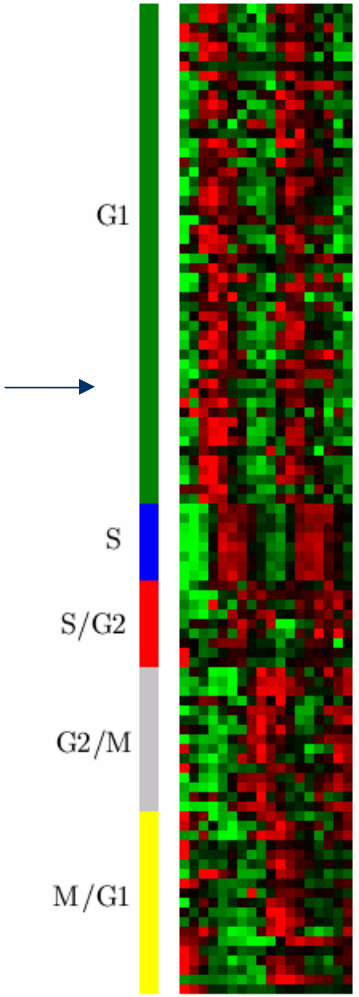
- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points.











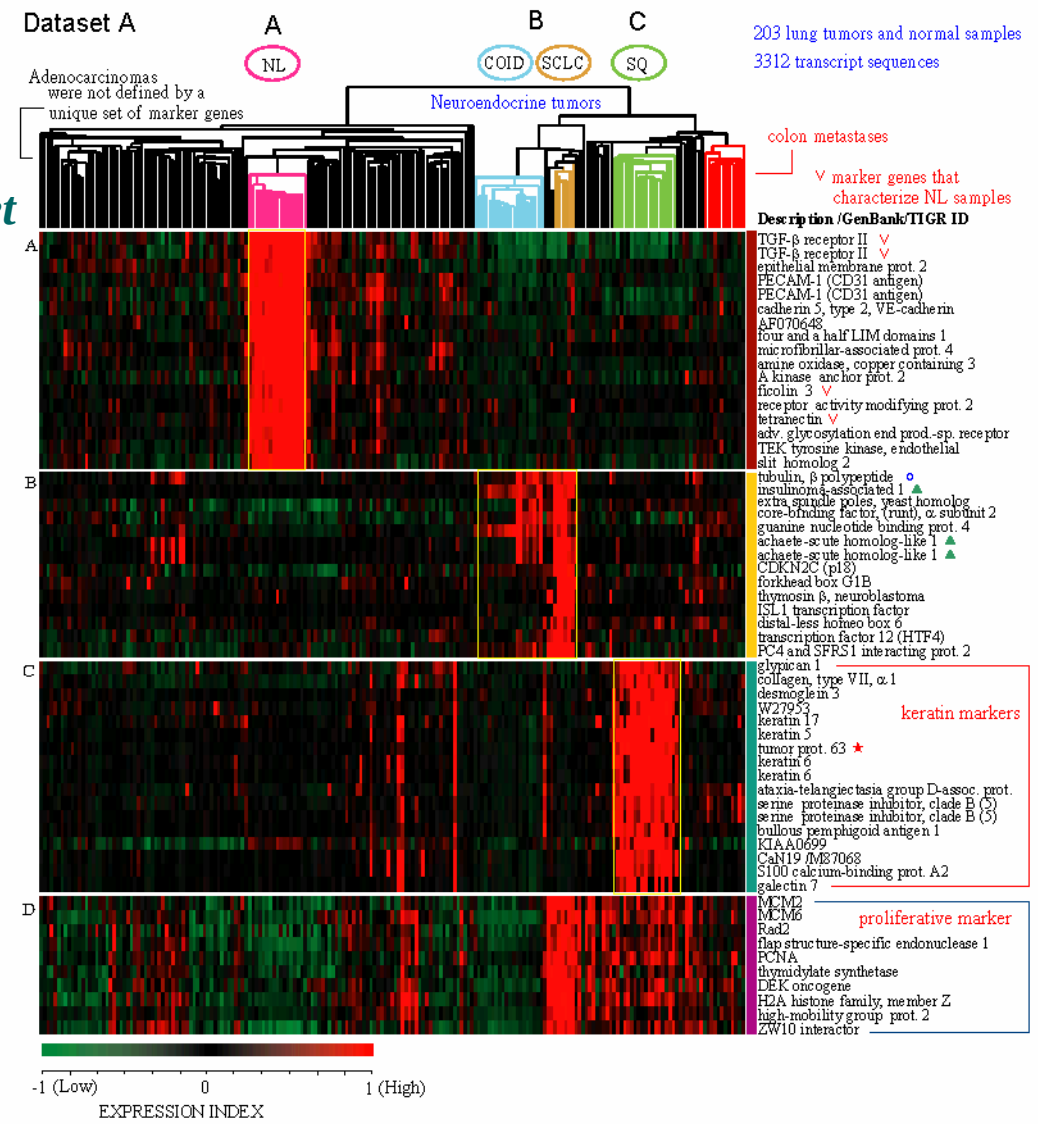
Hierarchical Clustering

Harvard Lung Cancer Dataset

#Sample: 203=186+17(NL)

Genes: 3312

SubTypes	*no.*
AD	127
SQ	21
COID	20
SCLC	6
Other	12
NL	17



Bhattacharjee *et al.*, (2001), Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. PNAS. 98 (24), 13790-13795.

K-Means Clustering

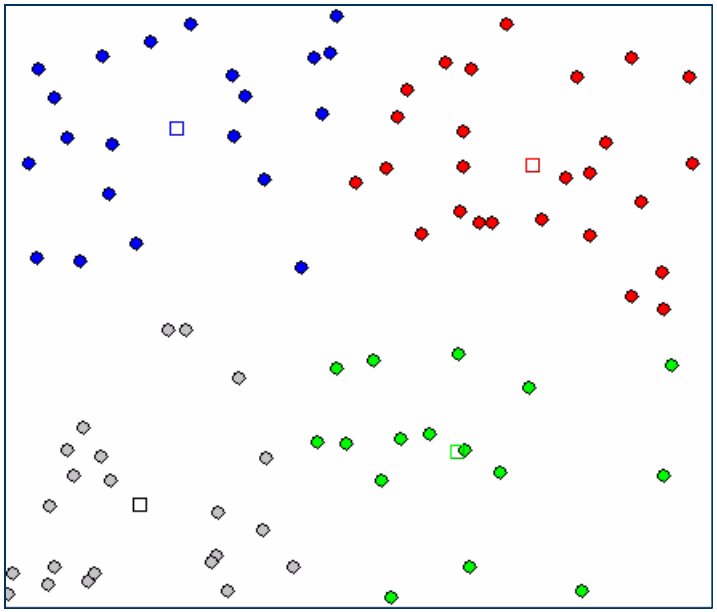
- K-meansis a partition methods for clustering.
- Data are classified into k groups as specified by the user.
- Two different clusters cannot have any objects in common, and the k groups together constitute the full data set.

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

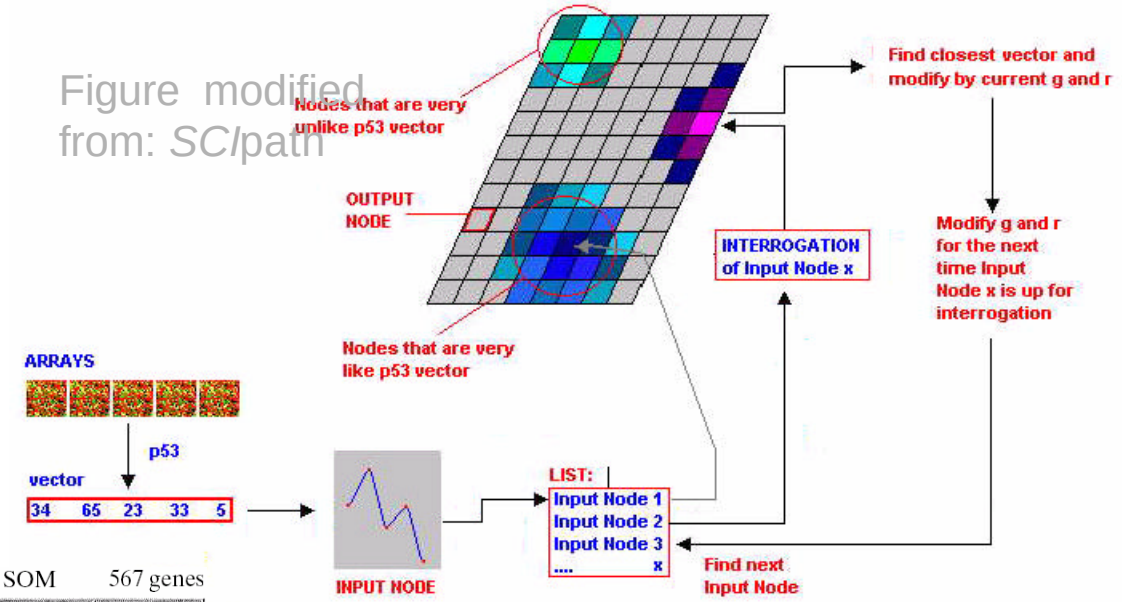


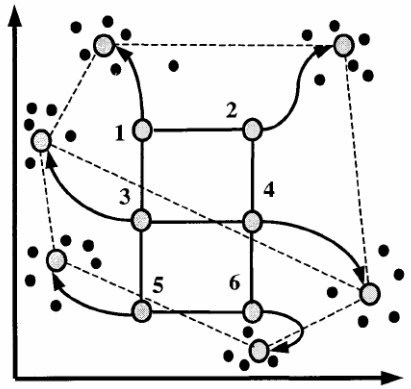
Self-Organizing Maps (SOM)

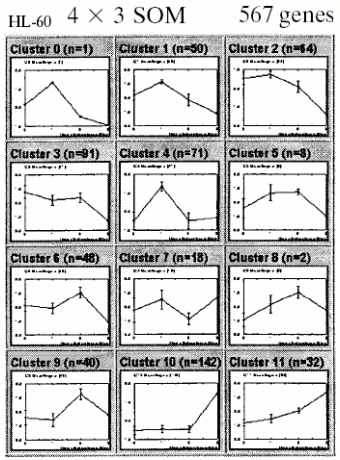
- SOMs were developed by Kohonen in the early 1980's, original area was in the area of speech recognition.
- ***Idea:*** Organise data on the basis of similarity by putting entities geometrically close to each other.

- 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







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.

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:

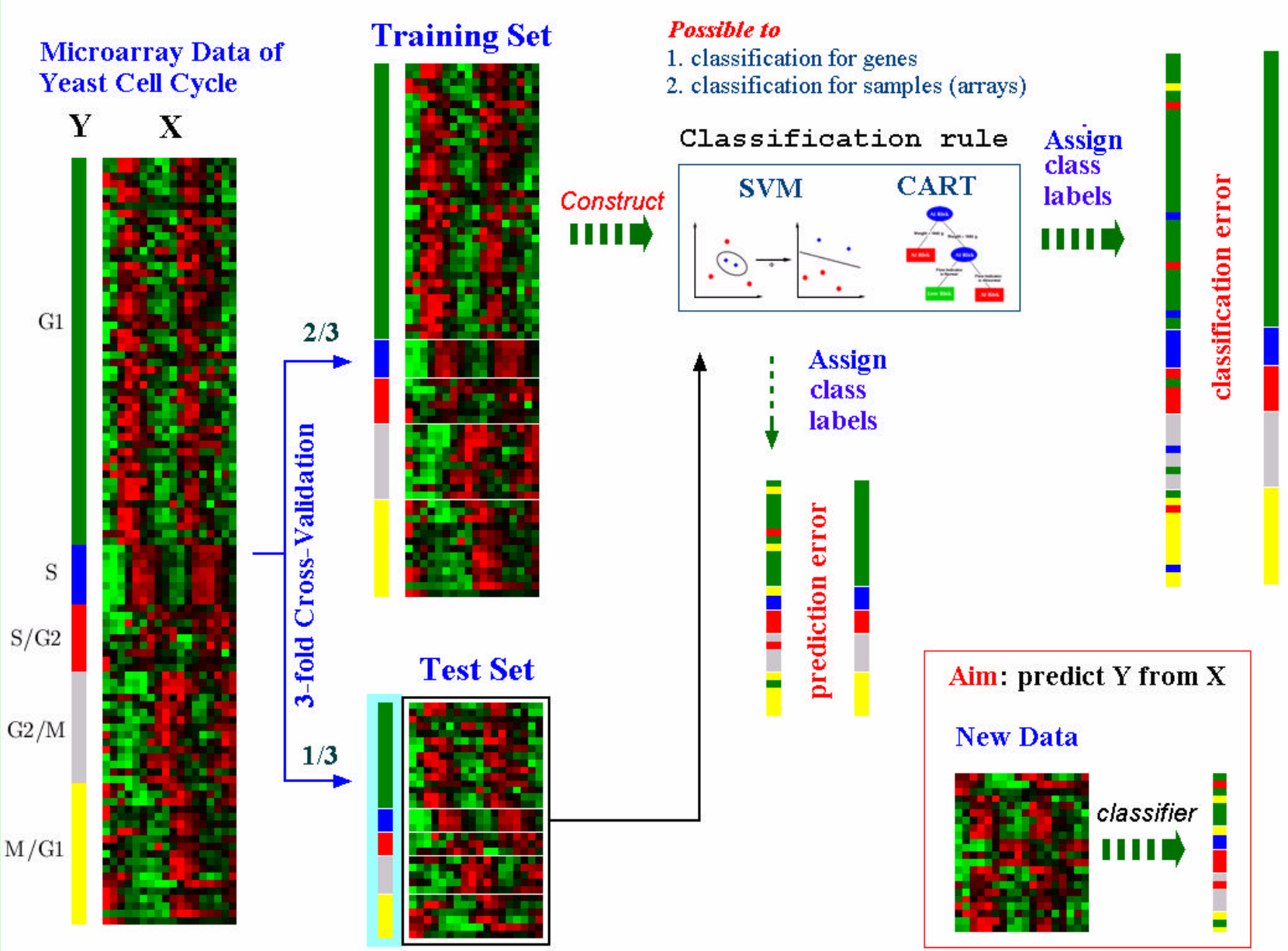
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

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.

Classification



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.

Linear Discriminant Analysis (LDA)

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

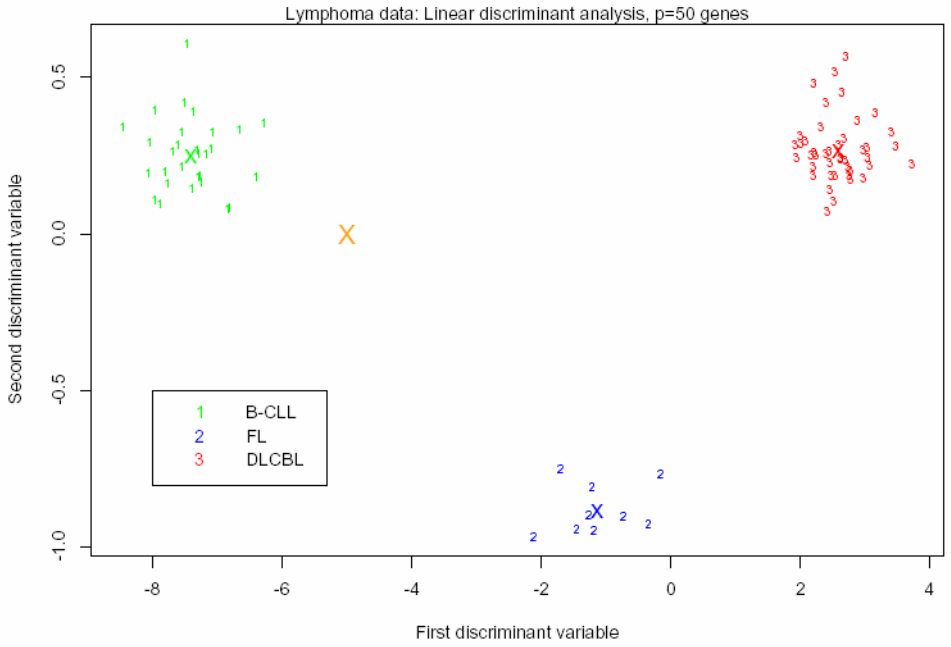
$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

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

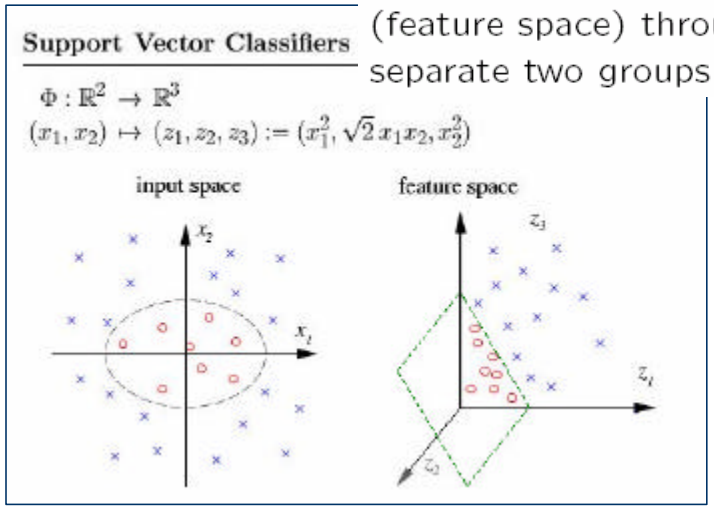
the p genes with the largest BSS/WSS ratios.

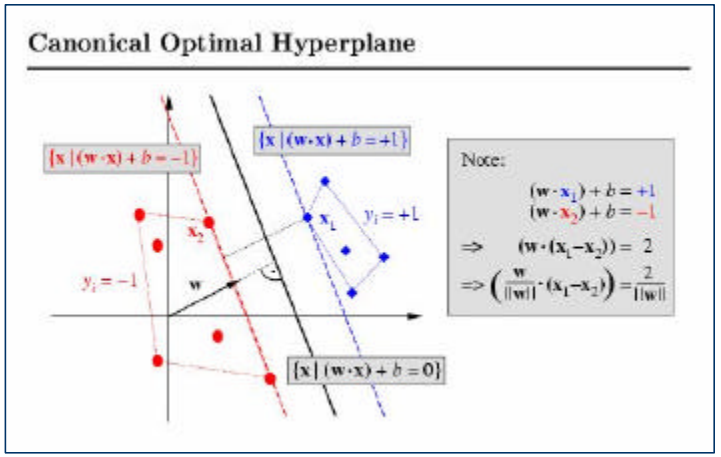


Dudoit S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 97 (457), 77-87.

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).





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

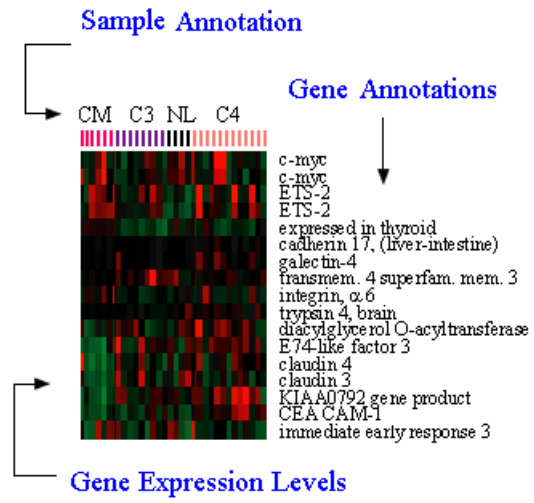
SVMTorch
(Collobert and Bengio, 2001)

LIBSVM
(Chang and Lin, 2002)

Brown et al. (2000). Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines, PNAS 97(1), 262-267.
Yeast Gene Expression [2467x 80] out of [6,221x 80] has accurate functional annotations.

Annotations in Microarray Data

- To learn the biological significance of the observed gene expression patterns.
- Rely on manual literature searches and expert knowledge to interpret microarray results.
- Database Referencing of Array Genes Online (DRAGON) allows microarray data to be annotated with data from publicly available databases such as UniGene.
- DRAGON offers a suite of visualization tools to identify genbe expression changes that occur in gene or protein families.
- The goal of annotation tools such as DRAGON is to provide insight into the biological significance of gene expression findings.



Examples of Some Available Databases

- Literature databases
  - PubMed, Medline, OMIM

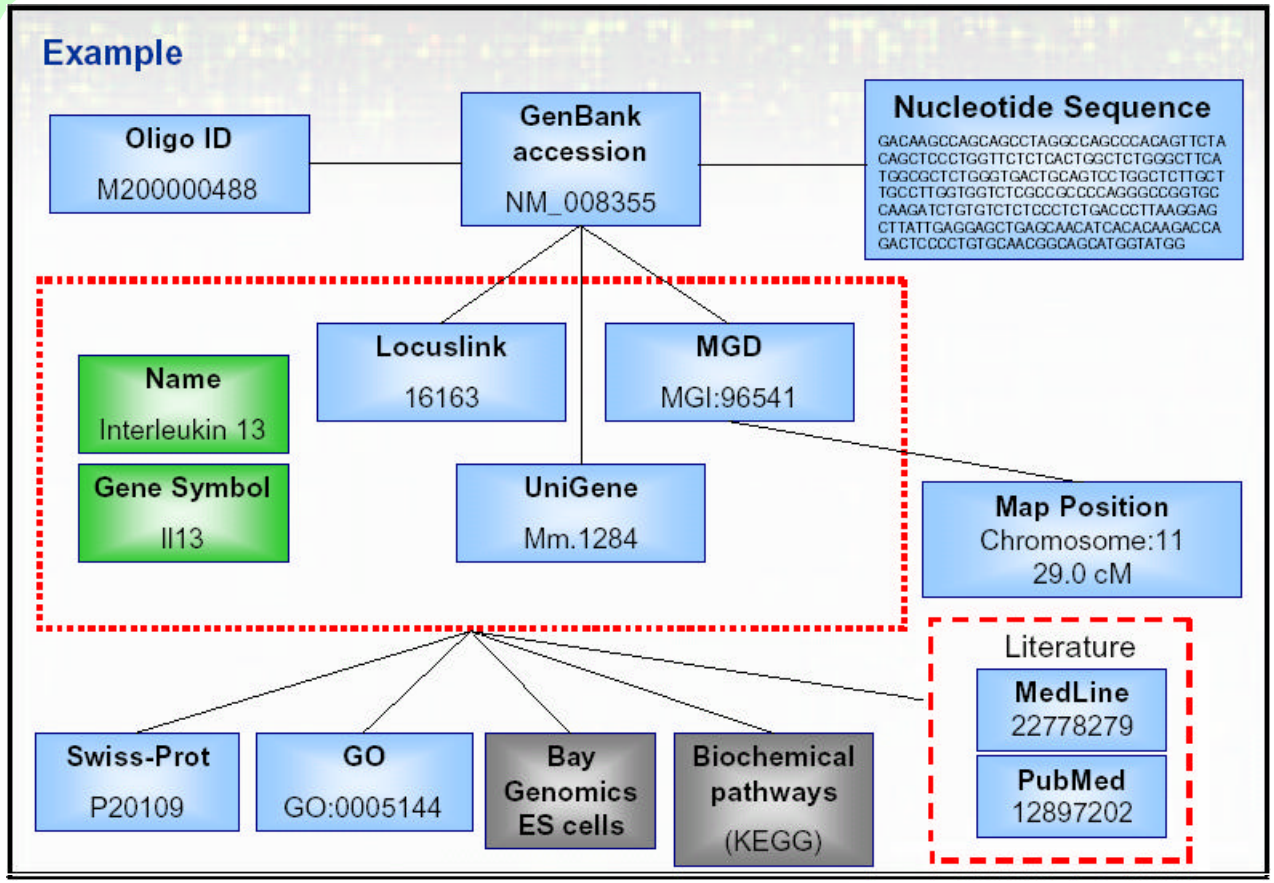
***Source**: UCSF Shared Functional Genomics Core Facility*

- Factual databases
  - Nucleic acid sequence: GenBank, EMBL, DDBJ, RefSeq.
  - Amino acid sequence: SwissProt.
  - 3D molecular structures: PDB.
- Knowledge and other databases
  - Gene classification: UniGene, Locuslink.
  - Gene Ontology: GO.
  - Motif libraries: Prosite.
  - Pathways: KEGG, WIT.
  - Transcription Factor: Transfac.

NCBI http://www.ncbi.nlm.nih.gov/About/tools/index.html

EBI http://www.ebi.ac.uk/Databases/

Gene Annotations



***Source**: UCSF Shared Functional Genomics Core Facility*

Software

- Gene Spring
- ArrayMiner
- BioConductor
- …

BioConductor: Package: annaffy ***(For Affymetrix GeneChip Data Only)***

Probe	Symbol	Description	Function	Chromo -some	Chromosome Location	GenBank	LocusLink	Cytoband	UniGene	PubMed	Gene Ontology	Pathway
207173_x_at	CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)		16	-64755822	D21254	1009	16q22.1	Hs.443435	9	protein binding calcium ion binding ossification	

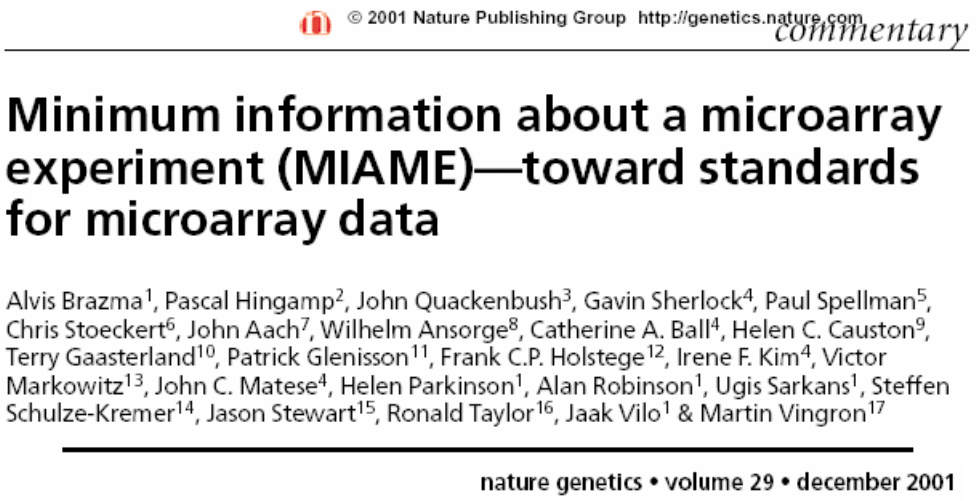
Sample Annotations

Microarray Gene Expression Data Society - MGED Society
http://www.mged.org/
MIAME
http://www.mged.org/Workgroups/MIAME/miame.html

- MIAME is a set of guidelines [that] will then assist with the development of microarray repositories and data analysis tools.

- MIAME aims to outline the minimum information required to unambiguously interpret microarray data and to subsequently allow independent verification of the data at a later stage if required.

- Store microarray parameters, such as sample name, mouse strain, amplification protocols used.



Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma[1], Pascal Hingamp[2], John Quackenbush[3], Gavin Sherlock[4], Paul Spellman[5], Chris Stoeckert[6], John Aach[7], Wilhelm Ansorge[8], Catherine A. Ball[4], Helen C. Causton[9], Terry Gaasterland[10], Patrick Glenisson[11], Frank C.P. Holstege[12], Irene F. Kim[4], Victor Markowitz[13], John C. Matese[4], Helen Parkinson[1], Alan Robinson[1], Ugis Sarkans[1], Steffen Schulze-Kremer[14], Jason Stewart[15], Ronald Taylor[16], Jaak Vilo[1] & Martin Vingron[17]

nature genetics • volume 29 • december 2001

Software

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

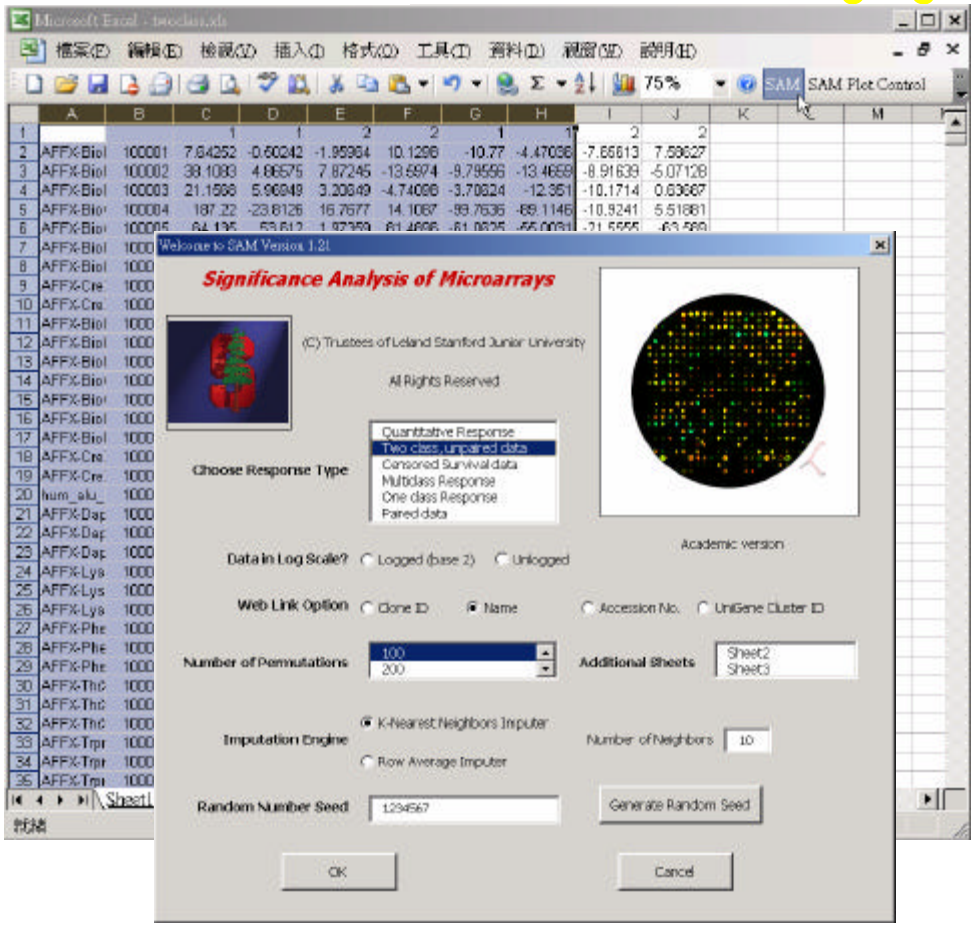
Significance Analysis of Microarray (SAM)

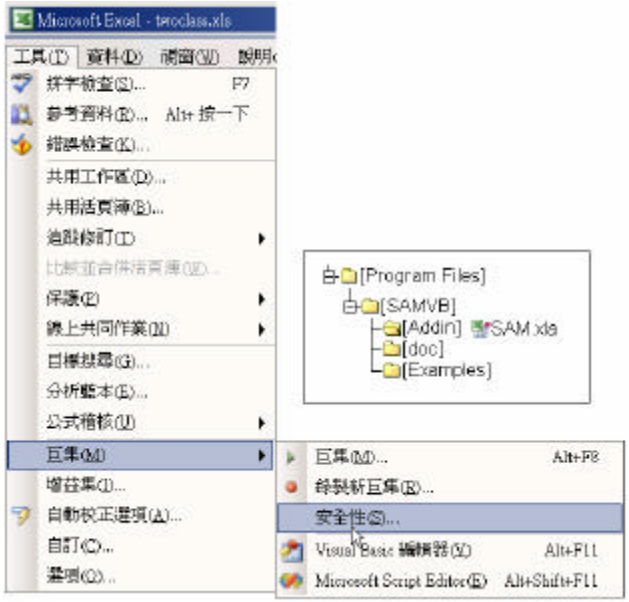
SAM assigns a score to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

- ***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.

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

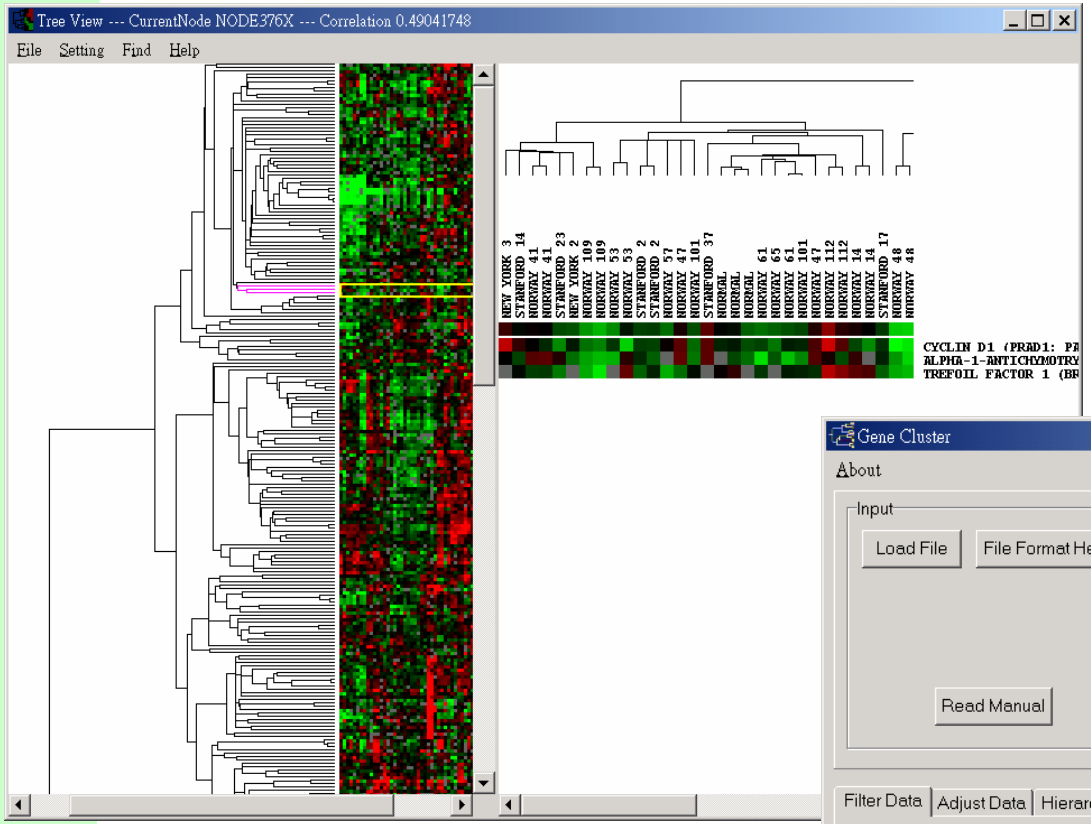
SAM does not do any normalization!





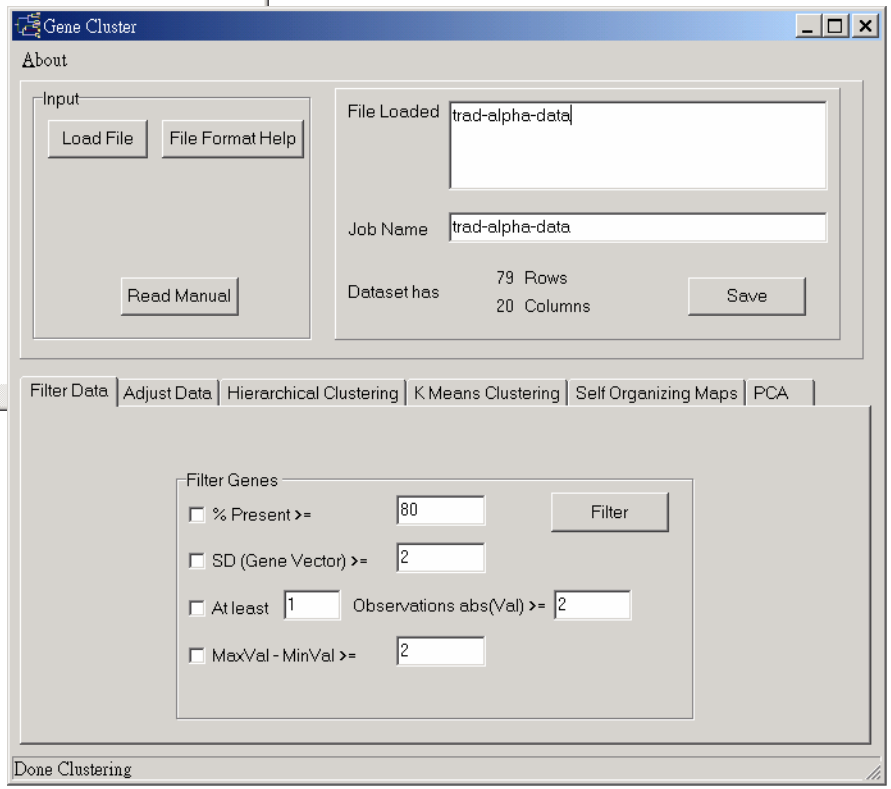
Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.

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.

The Bioconductor (v1.3.28)

The R Project for Statistical Computing

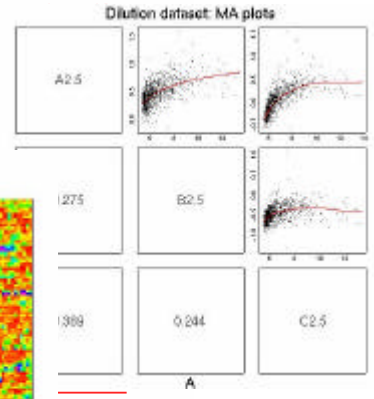
The Bioconductor Project Release 1.3
http://www.bioconductor.org/

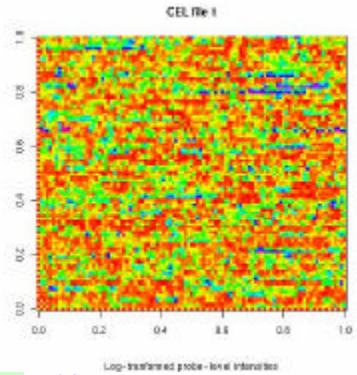
R version 1.9.0 has been released on 2004-04-12

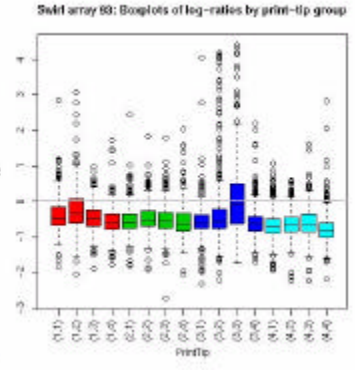
Package

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

genefilter
geneplotter
globaltest
gpls
graph
hexbin
limma

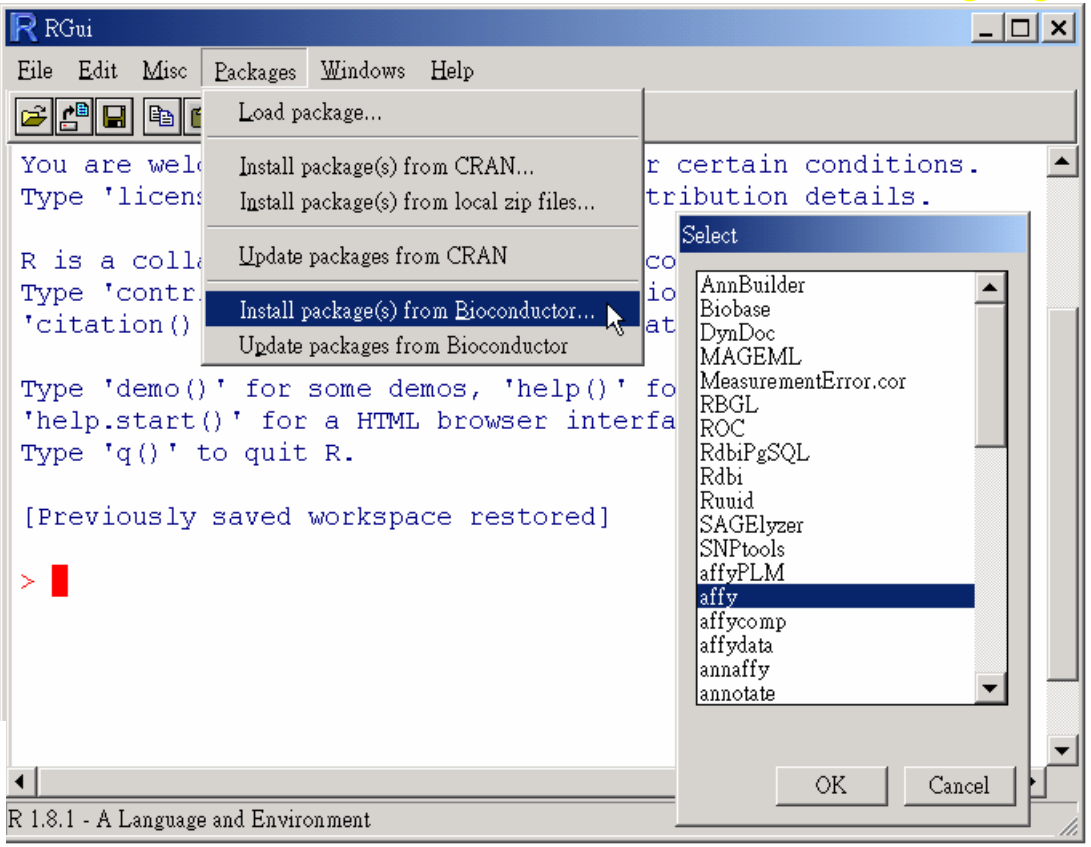






edd
externalVector
factDesign
gcrma

sigg
splic
tkW
vsn
wid



Installation

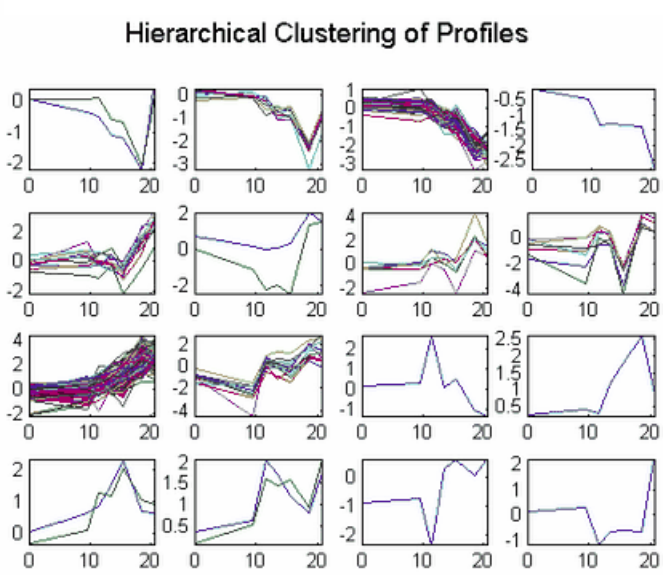
```
source("http://www.bioconductor.org/getBioC.R")
getBioC(relLevel="release")
```

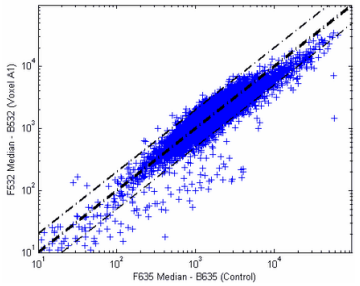
Matlab: Bioinformatics ToolBox

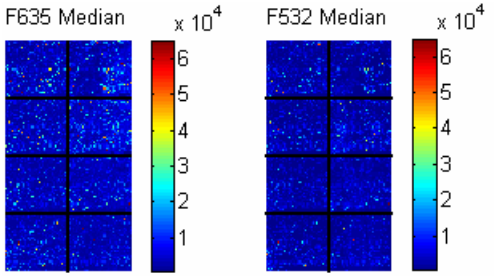
The MathWorks
Bioinformatics Toolbox

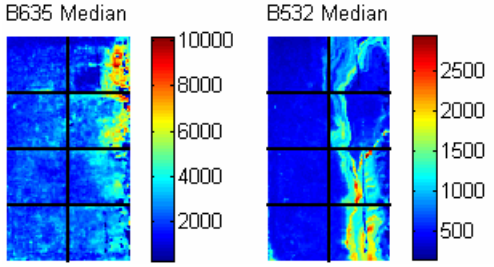
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html

- 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.









Useful Links

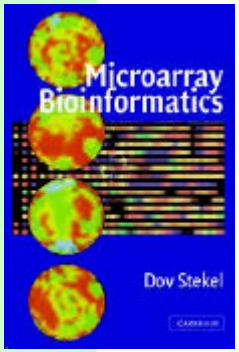
http://ihome.cuhk.edu.hk/~b400559/

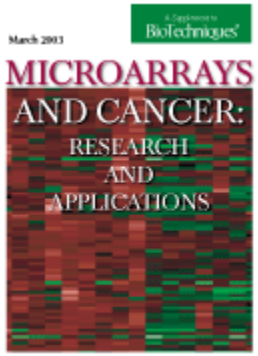
Bibliography on **Microarray Data Analysis**

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

- 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

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





BioConductor: open source software for bioinformatics

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

Some Related Issues

- Image Processing
- Analysis of Oligonucleotide Microarrays
- Analysis of Replicates Arrays
- Spatial Normalization
- Time Series Samples
- Experimental Design

Microsoft Excel - Mldemo.xls

	A	B	C
1	Probeset	Gene Name	Array 1 Signal
2	103941_at	alpha-spectin 1, erythroid	33.7625
3	104432_at	aplysia ras-related homolog N (Rhon)	127.736
4	104137_at	ATP-binding cassette, sub-family A (ABC1), member 2	109.522
5	98458_at	baculoviral IAP repeat-containing 5	128.96
6	93243_at	bone morphogenetic protein 7	174.85
7	95061_at	breast carcinoma amplified sequence 2	34.8
8	102632_at	calmodulin binding protein 1	69.888

Dimension Free Data Visualization
GAP

陳君厚
E-mail: cchen@stat.sinica.edu.tw
http://gap.stat.sinica.edu.tw/

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