陽明大學生物資訊研究所
2005 微陣列數據分析暑期課程

Microarray Data Analysis

Clustering and Visualization (I)

吳漢銘

2005年7月27日

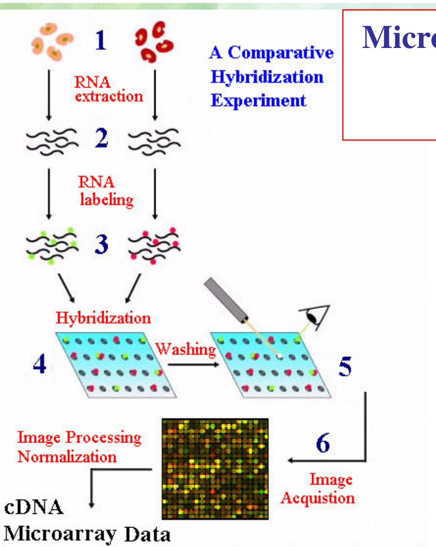
中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

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

Outlines

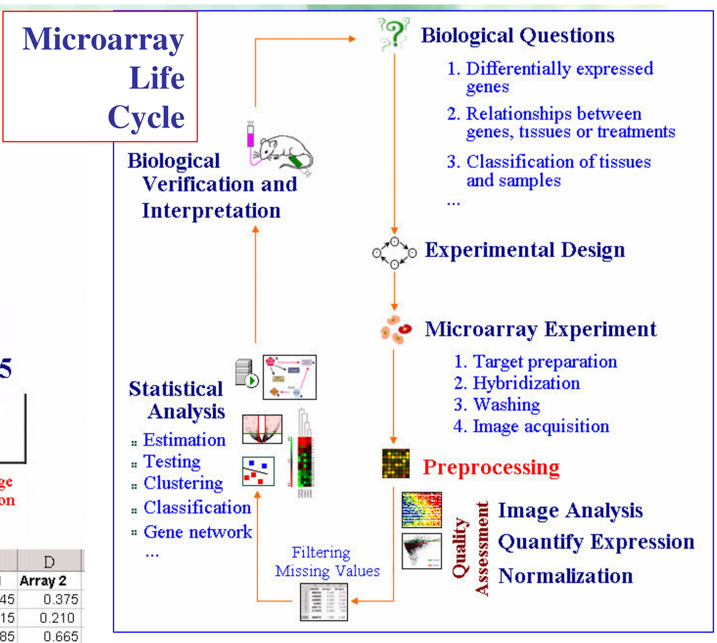
- **Overview of cDNA Microarray Experiment**
- **One/Two-dimensional Data**
  - Image Plot, Histogram, Boxplot, Scatterplot and MA Plot, Volcano Plot
- **High-dimensional Data : Dimension Reduction Techniques**
  - Distance and Similarity Measure
  - Principal Component Analysis (PCA) and Biplot
  - Multidimensional Scaling (MDS)
- **Clustering Analysis and Visualization**
  - Stages in Clustering
  - K-means
  - Self-Organizing Maps (SOM)
- **R, BioConductor and Lab Exercise**
- **Isomap (if we have time left)**

Overview of cDNA Microarray 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

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

Experimental Design

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

Preprocessing

Quality Assessment

Image Analysis
Quantify Expression
Normalization

Filtering
Missing Values

Statistical Analysis
- Estimation
- Testing
- Clustering
- Classification
- Gene network
...

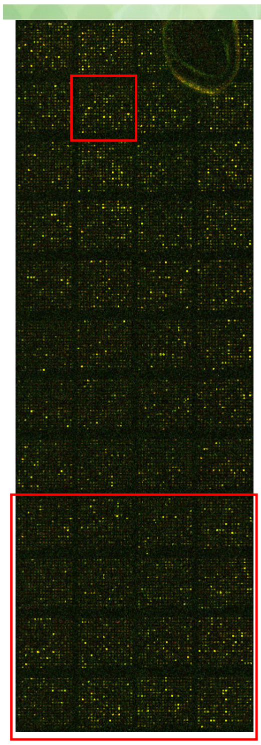
Biological Verification and Interpretation

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

R=Rf-Rb
G=Gf-Gb

M=log2R/G
A=1/2 log2RG

Array Image



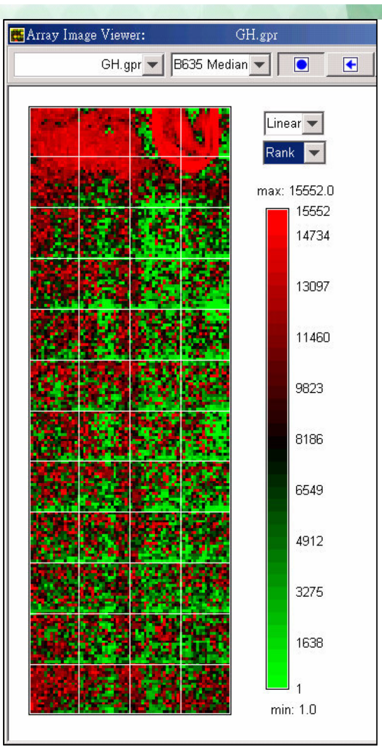
Blocks:
12 by 4

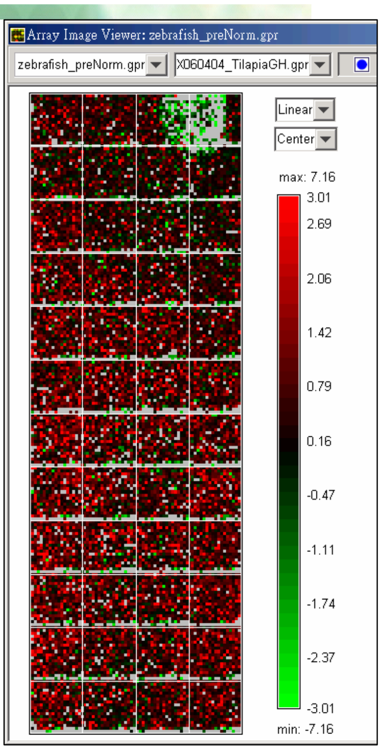
Features:
18 by 18

Signal
16-bit
0~65535

*.gpr

GAL

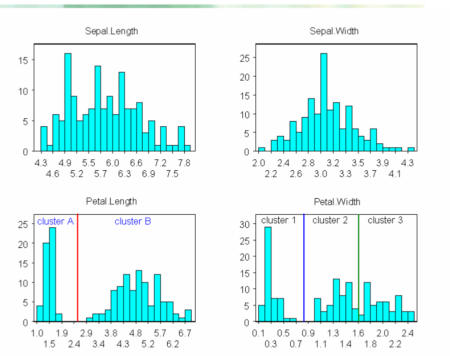




Histograms

The histogram shows:

1. center of the data (location)
2. spread of the data (scale)
3. skewness of the data
4. presence of outliers
5. presence of multiple modes in the data.



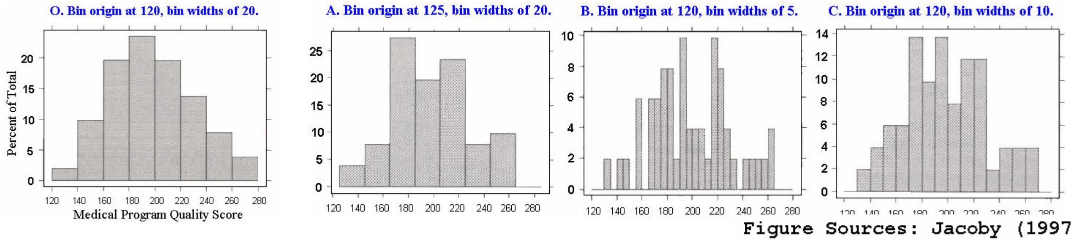
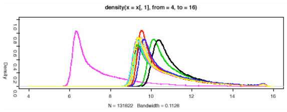
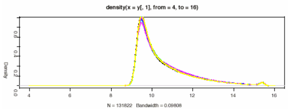


Figure Sources: Jacoby (1997).

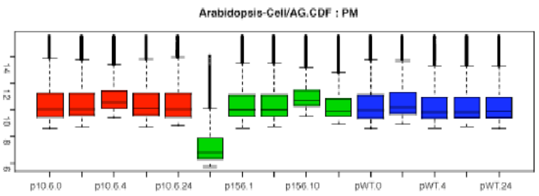
Density Plots

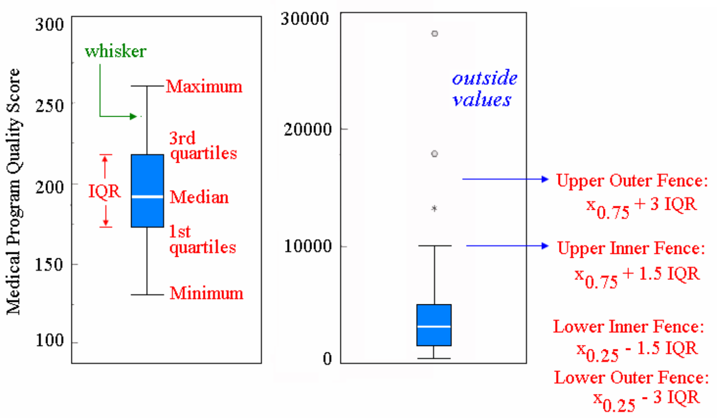




Box Plots

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





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

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

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

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

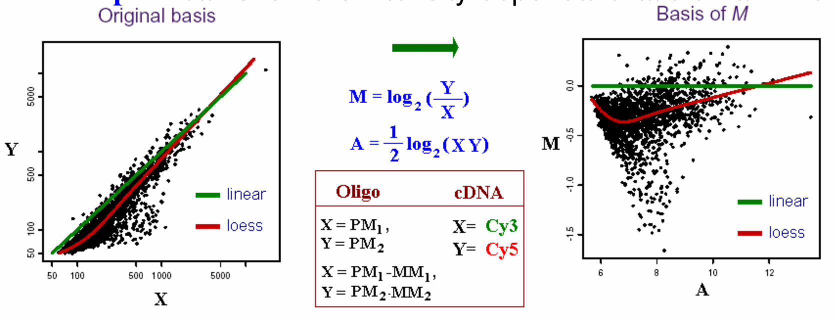
The box plot can provide answers to the following questions:

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

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

Scatterplot and MA plot

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



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

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

Oligo	cDNA

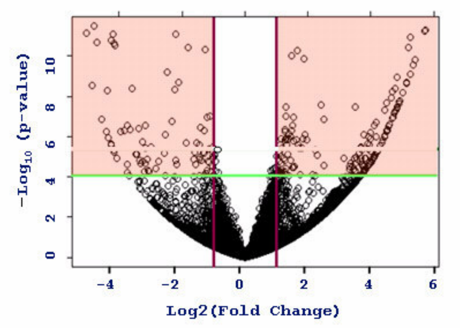
$X = PM_1$,
$Y = PM_2$
	X= Cy3
	Y= Cy5

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

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

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

Volcano Plot



A volcano plot is a heuristic device that arranges genes along dimensions of biological and statistical significance.

A volcano plot is helpful in identifying significance and magnitude of change in expression of a set of genes between two conditions.

- A volcano plot displays the negative log of p-values from a t-test on one axis and the log2 of change between two conditions on the other axis on the scatterplot view.

- The researcher can then make judgments about the most promising candidates for follow-up studies, by trading off both these criteria by eye.

Visualizing and Clustering High-dimensional Data: Dimension Reduction Techniques

- ◆ Principal Component Analysis (PCA)
- ◆ Biplot
- ◆ Multidimensional Scaling (MDS)

Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.

Distance and Similarity Measure

Cov	x1	x2	x3	x4		xp
x1	0.69	0.48	0.10	-0.10		-0.28
x2	0.48	0.71	0.41	0.22		-0.23
x3	0.10	0.41	0.50	0.36		-0.05
x4	-0.10	0.22	0.36	0.44		0.10
xp	-0.28	-0.23	-0.05	0.10		0.41

Proximity Matrix

Data Matrix x y →

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

Pearson Correlation Coefficient

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

Euclidean Distance

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

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

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

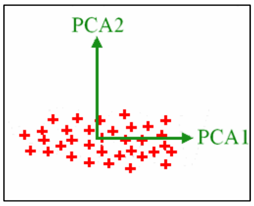
- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

Raw Data Matrix $\mathbf{X}$
Dispersion Matrix $\mathbf{S}_X^2 = \mathbf{X}^T\mathbf{X}$
Centered Data $\mathbf{C} = \mathbf{X} - \mu$
Covariance Matrix $\mathbf{\Sigma}_X = \mathbf{C}^T\mathbf{C}$
Scaled Data $\mathbf{Z} = \frac{\mathbf{X}-\mu}{\sigma}$
Correlation Matrix $\mathbf{R}_X = \mathbf{Z}^T\mathbf{Z}$

Principal Component Analysis (PCA)

(Pearson 1901; Hotelling 1933; Jolliffe 2002)

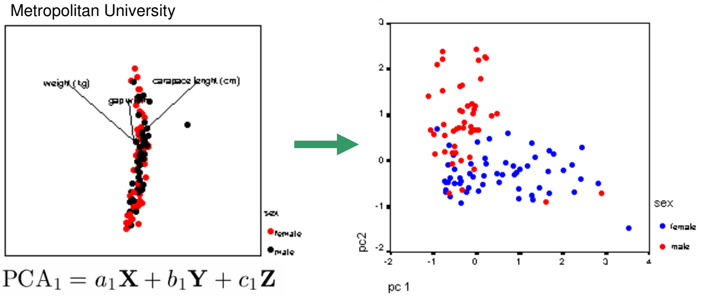
PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



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

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

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



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

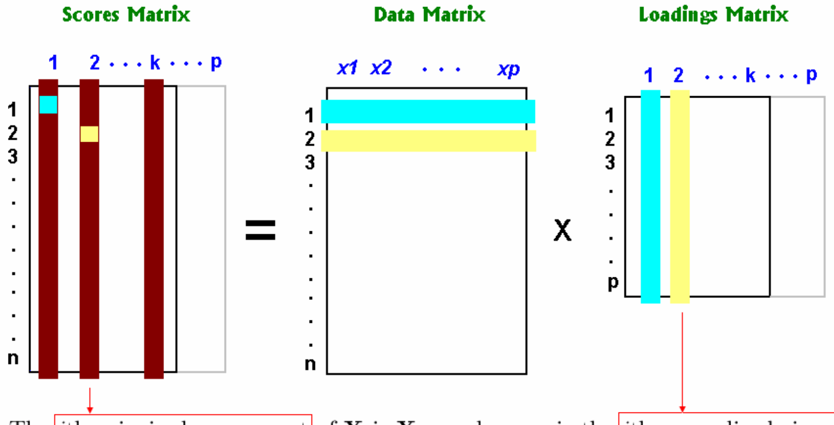
$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$

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

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

PCA: Loadings and Scores

$$\mathbf{U} = \mathbf{X}\ \mathbf{V}$$



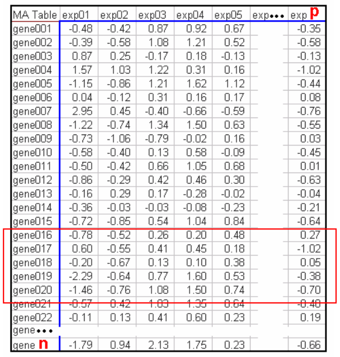
The $\boxed{i\text{th principal component}}$ of $\mathbf{X}$ is $\mathbf{X}\,\mathbf{v}_i$, where $\mathbf{v}_i$ is the $\boxed{i\text{th normalized eigenvector}}$ of $\Sigma_{\mathbf{x}}$ corresponding to the ith largest eigenvaules.

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

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

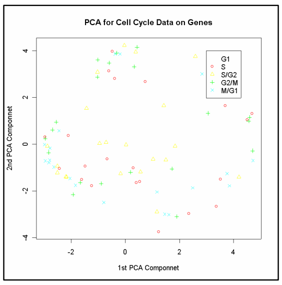
Interpretation of the PCA Results

Microarray Data Matrix



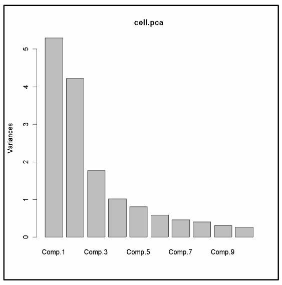
MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007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

PCA on Genes

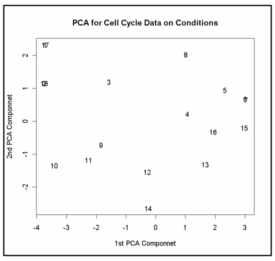


Scree plot

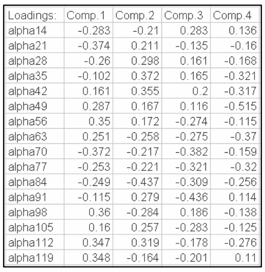
Variances ↔ Eigenvalues



PCA on Conditions

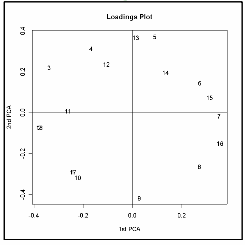


Loadings Matrix



Loadings:	Comp.1	Comp.2	Comp.3	Comp.4
alpha14	-0.283	-0.21	0.283	0.136
alpha21	-0.374	0.211	-0.135	-0.16
alpha28	-0.26	0.298	0.161	-0.168
alpha35	-0.102	0.372	0.165	-0.321
alpha42	0.161	0.355	0.2	-0.317
alpha49	0.287	0.167	0.116	-0.515
alpha56	0.35	0.172	-0.274	-0.115
alpha63	0.251	-0.258	-0.275	-0.37
alpha70	-0.372	-0.217	-0.382	-0.159
alpha77	-0.253	-0.221	-0.321	-0.32
alpha84	-0.249	-0.437	-0.309	-0.256
alpha91	-0.115	0.279	-0.436	0.114
alpha98	0.36	-0.284	0.186	-0.138
alpha105	0.16	0.257	-0.283	-0.125
alpha112	0.347	0.319	-0.178	-0.276
alpha119	0.348	-0.164	-0.201	0.11

Loadings Plot



```
> summary(cell.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5          Comp.15
Standard deviation     2.3012110 2.0542795 1.3300507 1.00895544 0.90053289      0.308577283
Proportion of Variance 0.3309732 0.2637540 0.1105647 0.06362444 0.05068497 ••• 0.005951246
Cumulative Proportion  0.3309732 0.5947272 0.7052919 0.76891637 0.81960134      1.000000000
```

The Biplot: Scores + Loadings

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

The purpose of the biplot is to show variables and observations together, in a way that represents graphically their joint interrelationships.

The data matrix can be factored:

$$\mathbf{X} = \mathbf{A}\mathbf{B}'$$

$\mathbf{X}_{n\times p}$: 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 **A** and **B**, using Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

$\mathbf{A}_{[2]}$: 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.

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

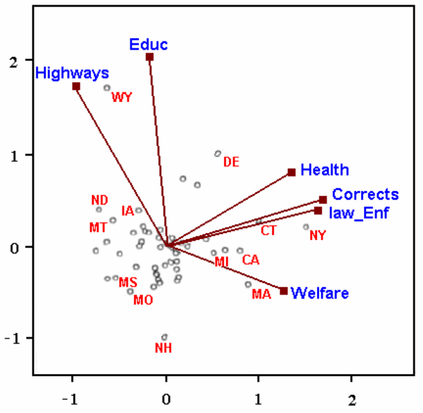
$$\mathbf{B}_{[2]} = \mathbf{V}_{[2]}\mathbf{D}^{1-c}_{[2]}$$

$\mathbf{U}_{[2]}$: the first two columns of **U**.
$\mathbf{V}_{[2]}$: the first two columns of **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

K. R. Gabriel (1971). The biplot graphical display of matrices with application to principal component analysis. Biometrika 58, 453-467.
J.C. Gower and D. J. Hand (1996). Biplots. Chapman & Hall.

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)

Seattle
San Francisco
Los Angeles
Denver
Chicago
Houston
Atlanta
Miami
New York
Washington D.C.
MEXICO
North Pacific Ocean
North Atlantic Ocean
Gulf of Mexico
THE BAHAMAS

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

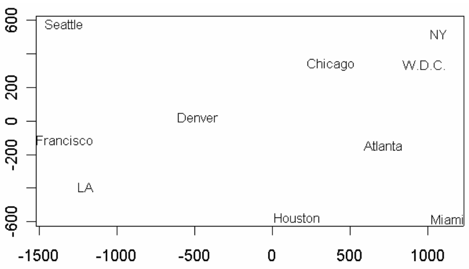
?

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

Flying Mileages Between Ten U.S. Cities

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

MDS



MDS: Metric and Non-Metric Scaling

Question

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

Goal of metric scaling

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

Goal of non-metric scaling

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

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

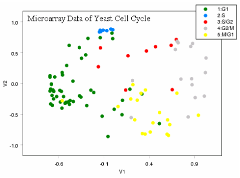
Mathematically: for given k, compute points x1,…,xn in k-dimensional space such that the object function is minimized.

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

2D MDS

Configuration Plot

for 103 known genes



Microarray Data of Yeast Cell Cycle

Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)

103 known genes: every 7 minutes and totally 18 time points.

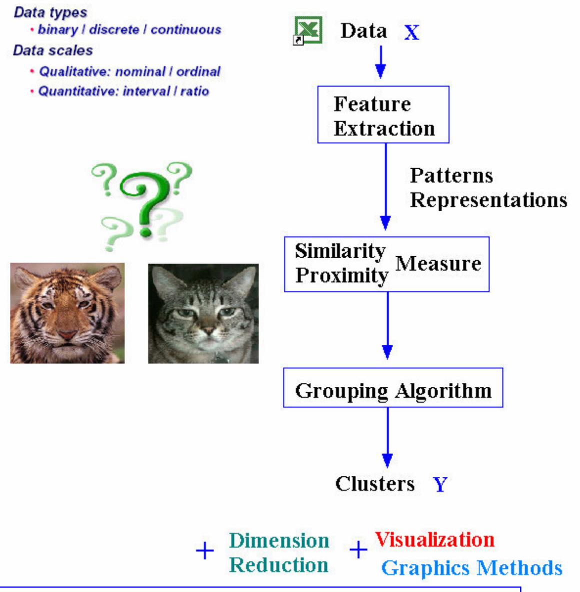
Clustering Analysis

What is Clustering?

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

Clustering Methods

- Hierarchical Clustering Algorithm
- Partitional Algorithm: k-means
- SOM
- Nearest Neighbor Clustering
- Fuzzy Clustering
- Artificial Neural Networks for Clustering
- Clustering Large data sets
- …



Data types
- *binary / discrete / continuous*

Data scales
- *Qualitative: nominal / ordinal*
- *Quantitative: interval / ratio*

Data X

Feature Extraction

Patterns Representations

Similarity Proximity Measure

Grouping Algorithm

Clusters Y

+ Dimension Reduction + Visualization Graphics Methods

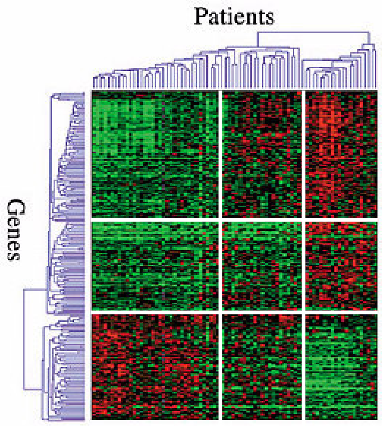
Two important properties of a clustering definition:
1. Most of data has been organized into non-overlapping clusters.
2. Each cluster has a within variance and one between variance for each of the other clusters. A good cluster should have a small within variance and large between variance.

Clustering Analysis in Microarray Experiments

Goals

- Find natural classes in the data
- Identify new classes/gene correlations
- Refine existing taxonomies
- Support biological analysis/discovery

- cluster genes based on samples profiles
- cluster samples based on genes profiles



Hypothesis:

- genes with similar function have similar expression profiles

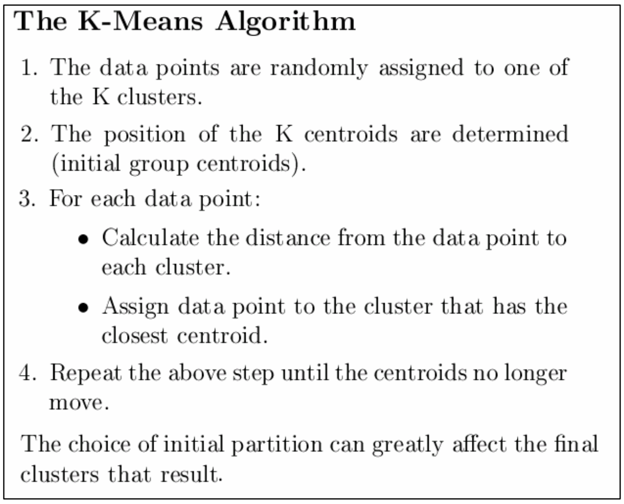
K-Means Clustering

- 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

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

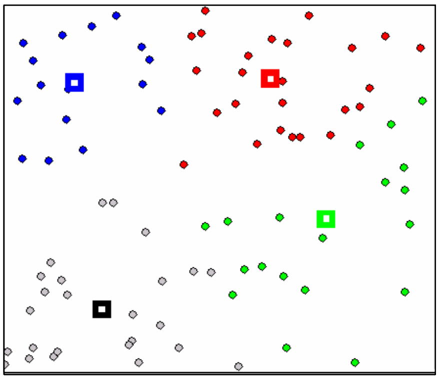


The K-Means Algorithm

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



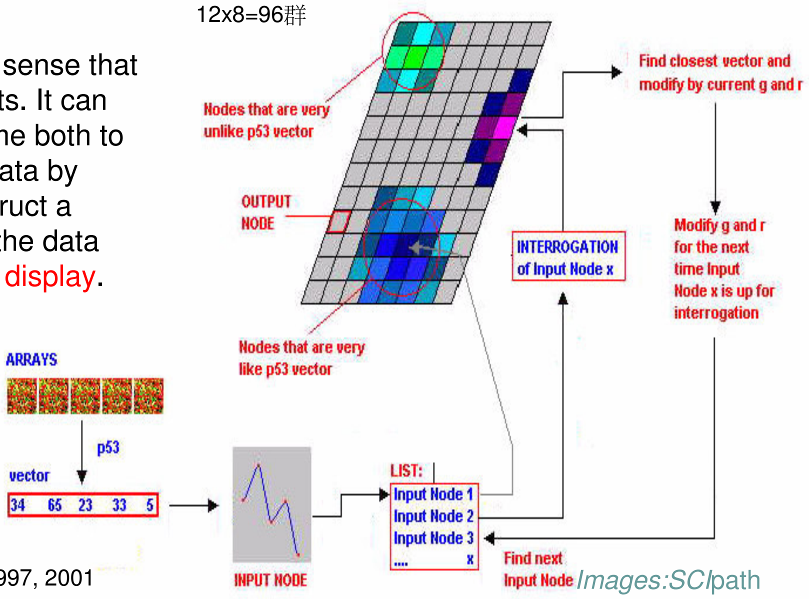
Self-Organizing Maps (SOM)

■ SOMs were developed by Kohonen in the early 1980's, original area was in the area of speech recognition.

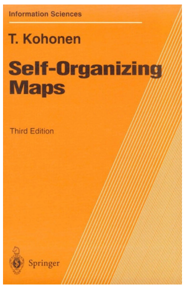
■ ***Idea:*** Organise data on the basis of similarity by putting entities geometrically close to each other.

■ 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.

12x8=96群


Nodes that are very unlike p53 vector
Find closest vector and modify by current g and r
OUTPUT NODE
INTERROGATION of Input Node x
Modify g and r for the next time Input Node x is up for interrogation
Nodes that are very like p53 vector
ARRAYS
p53
vector
34 65 23 33 5
INPUT NODE
LIST:
Input Node 1
Input Node 2
Input Node 3
.... x
Find next Input Node

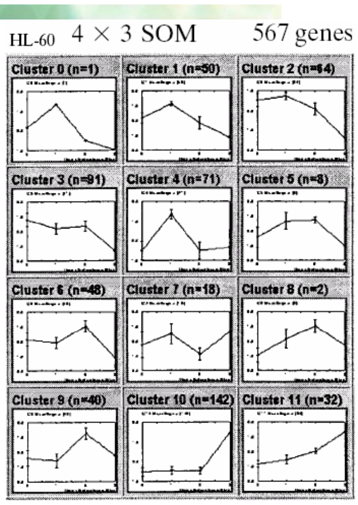
Images:SCIpath

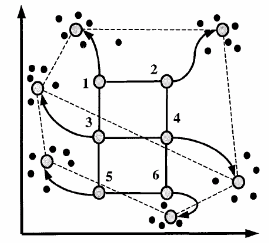


1995, 1997, 2001

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.
```



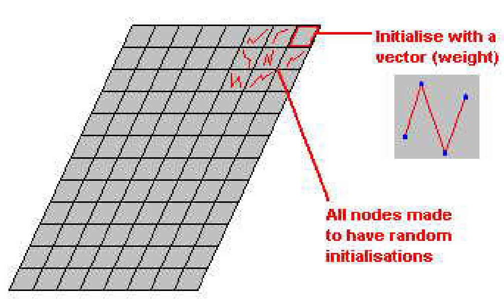


Macrophage Differentiation in HL-60 cells

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.

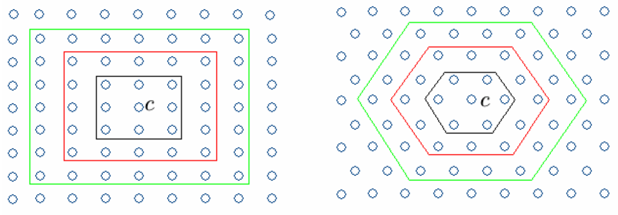
SOM - Initialization

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


Initialise with a vector (weight)
All nodes made to have random initialisations

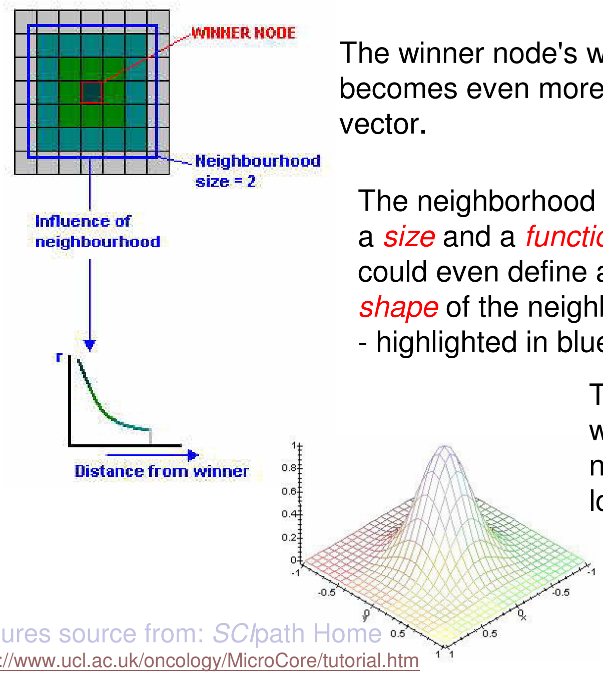
SOM initialization means to give each weight of the output node a random (or determined) vector value. *The dimensionality of the vector values put in* ***must match*** *the dimensionality of the raw data!* So if the raw data consists of 5 arrays, then the vectors must have 5 elements (dimensions).

Two examples of topological neighborhood.



■ $N_c(t_1) = 1$, 🟥 $N_c(t_2) = 2$, 🟩 $N_c(t_3) = 3$, $t_1 < t_2 < t_3$

Neighborhood Functions



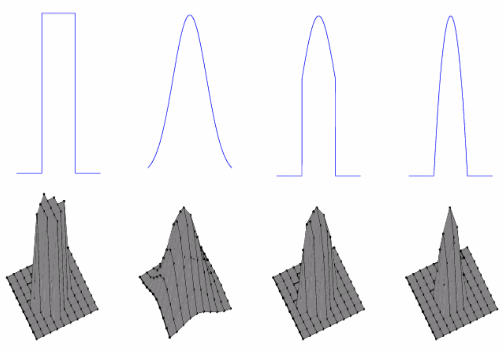
The winner node's weight is *modified* such that it becomes even more *similar* to the original input node's vector.

The neighborhood value has a two-fold character - a *size* and a *function of distance to influence*. One could even define a further third character - *the shape* of the neighborhood (in this case, a square - highlighted in blue).

The peak of the Gaussian function would be the location of the winner node. As one moves out from that location, the *r* value decreases.

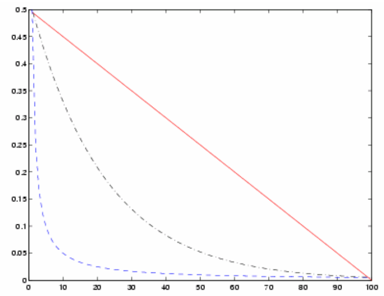
Figures source from: *SCI*path Home
http://www.ucl.ac.uk/oncology/MicroCore/tutorial.htm

Neighborhood Functions and Learning Rate



Different neighborhood functions. From the left
`'bubble'` $h_{ci}(t) = \mathbf{1}(\sigma_t - d_{ci})$,
`'gaussian'` $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2}$,
`'cutgauss'` $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2}\mathbf{1}(\sigma_t - d_{ci})$, and
`'ep'` $h_{ci}(t) = \max\{0, 1-(\sigma_t - d_{ci})^2\}$, where
σ_t is the neighborhood radius at time t,
$d_{ci} = ||\mathbf{r}_c - \mathbf{r}_i||$ is the distance between map units c and i on the map grid
$\mathbf{1}(x)$ is the step function: $\mathbf{1}(x) = 0$ if $x < 0$ and $\mathbf{1}(x) = 1$ if $x \geq 0$.
The neighborhood radius used is $\sigma_t = 2$.

Source from Technical report on SOM Toolbox 2.0 for Matlab.



Different learning rate functions:
`'linear'` (solid line) $\alpha(t) = \alpha_0\,(1 - t/T)$,
`'power'` (dot-dashed) $\alpha(t) = \alpha_0(0.005/\alpha_0)^{t/T}$ and
`'inv'` (dashed) $\alpha(t) = \alpha_0/(1 + 100\,t/T)$, where T
is the training length and α_0 is the initial learning rate.

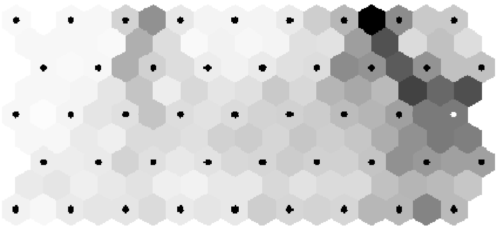
Possible Parameters used in SOM Analysis

1. Grid dimension: 2D, 3D
2. Grid shape: in 2D → Rectangle, Hexagon, ...
3. Number of node: in 2D Rectangle → 4×6, 5×5, 3×8,...
4. Neighborhood function: Bubble kernel, Gaussian kernel, ...
5. Neighborhood size: radius of $N_c(t)$
6. Learning rate function: $\alpha(t)$
7. Initial weights: random, use input vector
8. Order of input vectors: random, ...
9. Ways of learning: number of iteration,...

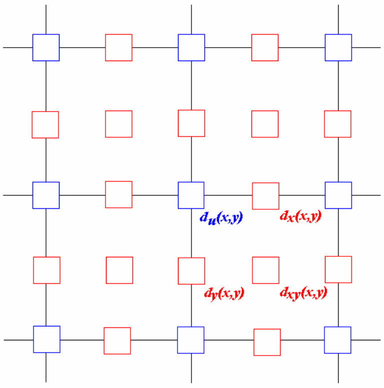
U-matrix: Unified Matrix Method

(Ultsch and Siemon 1989, Ultsch 1993)

U-matrix representation of SOM visualizes the distance between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes.



U-matrix representation of the SOM



$b(x,y)$: matrix of neurons, of size $n_x \times n_y$.

$w_i(x,y)$: matrix of weights.

$u(x,y)$: U-matrix of size $(2n_x - 1) \times (2n_y - 1)$.

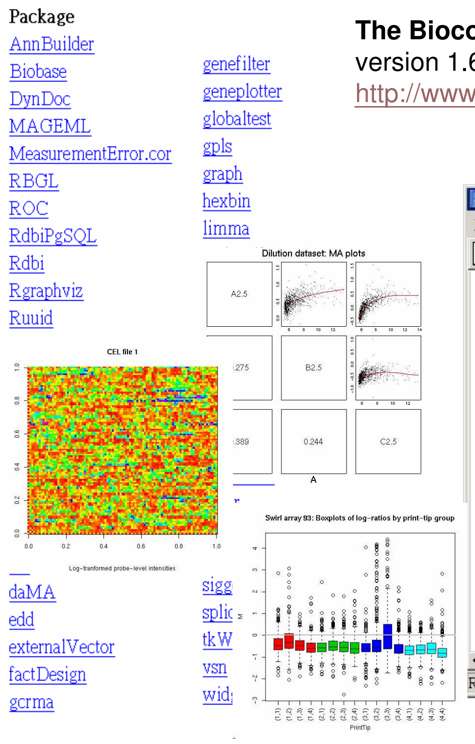
$d_x(x,y)$: $\|b(x,y) - b(x+1,y)\| = \sqrt{\sum_i [w_i(x,y) - w_i(x+1,y)]^2}$

$d_y(x,y)$: $\|b(x,y) - b(x,y+1)\| = \sqrt{\sum_i [w_i(x,y) - w_i(x,y+1)]^2}$

$d_{xy}(x,y)$: $\frac{1}{2}[\frac{\|b(x,y)-b(x+1,y+1)\|}{\sqrt{2}} + \frac{\|b(x,y+1)-b(x+1,y)\|}{\sqrt{2}}]$

$d_u(x,y)$: the median of the surrounding elements.

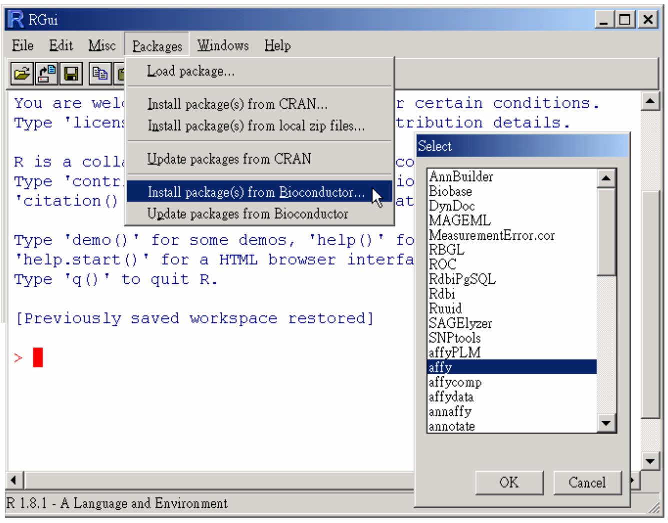
The Bioconductor



The Bioconductor
version 1.6
http://www.bioconductor.org

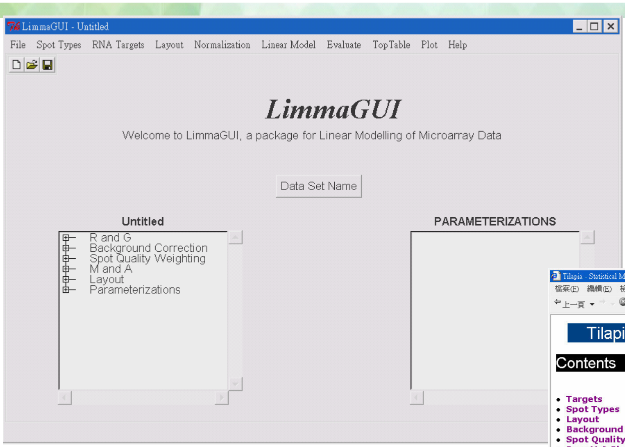
The R Project for Statistical Computing

R version 2.1.1 (**2005-06-20)**
http://www.r-project.org

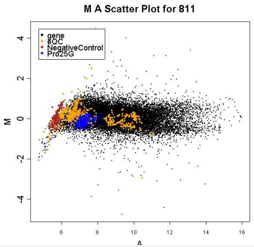

RGui
File
Edit
Misc
Packages
Windows
Help
Load package...
Install package(s) from CRAN...
Install package(s) from local zip files...
Update packages from CRAN
Install package(s) from Bioconductor...
Update packages from Bioconductor
Select
AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Ruuid
SAGElyzer
SNPtools
affyPLM
affy
affycomp
affydata
annaffy
annotate
OK
Cancel
R 1.8.1 - A Language and Environment

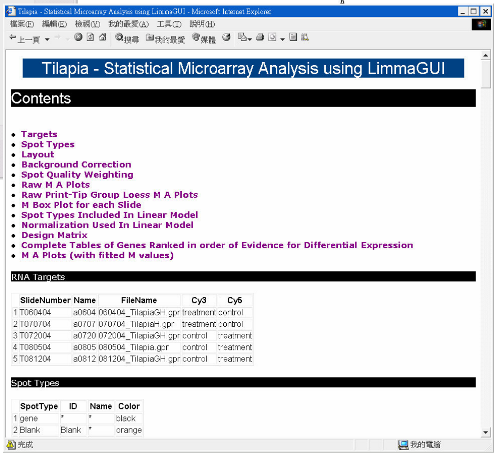
```
You are welc                                    r certain conditions.
Type 'licens                                    tribution details.

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

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

[Previously saved workspace restored]

> 
```

Limma, LimmaGUI, LimmaAffy



Welcome to LimmaGUI, a package for Linear Modelling of Microarray Data




Tilapia - Statistical Microarray Analysis using LimmaGUI - Microsoft Internet Explorer
檔案(F)
編輯(E)
檢視(V)
我的最愛(A)
工具(T)
說明(H)
上一頁
搜尋
我的最愛
媒體
完成
我的電腦

Tilapia - Statistical Microarray Analysis using LimmaGUI

Contents

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

RNA Targets

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

Spot Types

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

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.

Reference

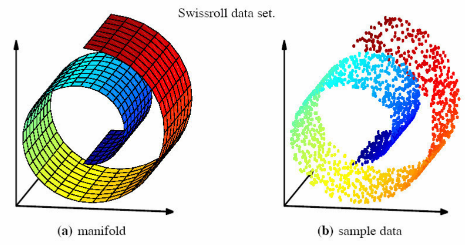
- Dr. Alexander Strehl: http://www.lans.ece.utexas.edu/~strehl/
- Michael Friendly's Home Page:http://www.math.yorku.ca/SCS/friendly.html

- Cox, T. F. and Cox, M.A.A. (2001), Multidimensional Scaling, London: Chapman & Hall.
- Hartigan, J. (1975), Clustering Algorithms, John Wiley and Sons, New York.
- Jacoby, W. G. (1998), Statistical Graphics for Visualizing Multivariate Data, Thousand Oaks, Calif. : Sage Publications.
- Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
- Kohonen, T. (2001), Self-Organizing Maps, Berlin: Springer.

Concept of Manifolds and Nonlinearity

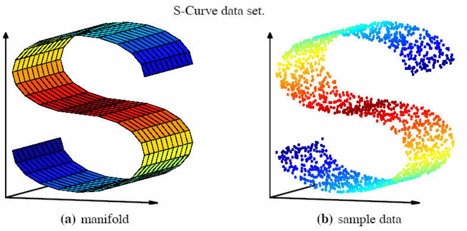
- A manifold is a topological space which is locally Euclidean. (i.e., around every point, there is a neighborhood that is topologically the same as the open unit ball in).
- In general, any object which is nearly "flat" on small scales is a manifold.
- Euclidean space is a simplest example of a manifold.
- More formally, any object that can be "charted" is a manifold.
- Intuitively, a manifold can be considered as a ``nice'' topological space that behaves at every point like our intuitive notion of a surface
- Manifolds arise naturally whenever there is a smooth variation of parameters [like pose of the face]

- The dimension of a manifold is the minimum integer number of co-ordinates necessary to identify each point in that manifold.



Swissroll data set.

(a) manifold (b) sample data



S-Curve data set.

(a) manifold (b) sample data

Isometric Mapping (isomap)

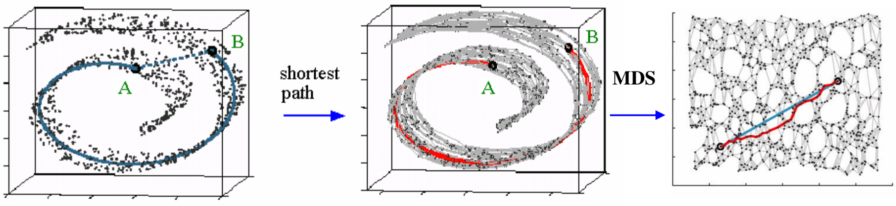
Isomap finds the projection that preserves the global, nonlinear geometry of the data by preserving the geodesic manifold interpoint distances.

- For neighboring points Euclidean distance is a good approximation to the geodesic distance.
- For farway points estimate the distance by a series of short hops between neighboring points.
- Find shortest paths in a graph with edges connecting neighboring data points.
- Once we have all pairwise geodesic distances use classical metric MDS

Algorithm of Isomap (Tenenbaum *et al.*, 2000)

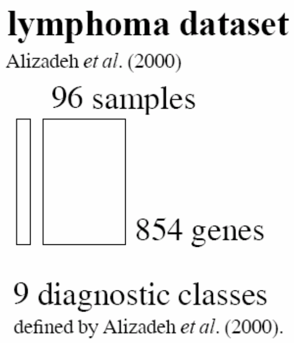
1. Calculate the distance $d_X(i,j)$ between all pairs i,j from n data points in the p-dimensional input space.
2. Construct the graph by determining the neighbors for each data point with ϵ-Isomap or k-Isomap.
3. Pursue the shortest paths in the graph G. Initialize $d_G(i,j) = d_X(i,j)$ if i,j are neighbors; otherwise, set $d_G(i,j) = \infty$. For each value of $l = 1, 2, \cdots, n$ and for all i,j, $d_G(i,j)$ are replaced by $\min\{d_G(i,j), d_G(i,l) + d_G(l,j)\}$.
4. Apply classical MDS to D_G.

What is important is the geodesic distance!



Tenenbaum , J. B., Silva, V. de, and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 290, 2319-2323.

Example



lymphoma dataset

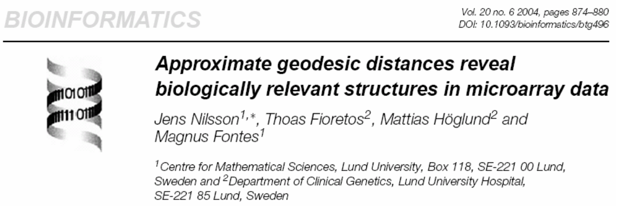
Alizadeh *et al.* (2000)

96 samples

854 genes

9 diagnostic classes

defined by Alizadeh *et al.* (2000).



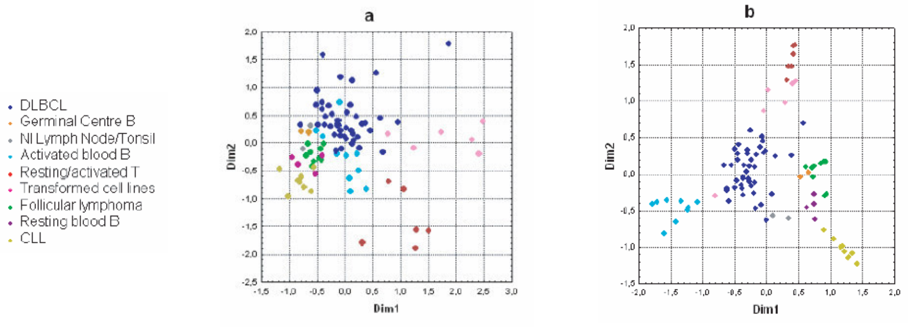
BIOINFORMATICS

Vol. 20 no. 6 2004, pages 874–880
DOI: 10.1093/bioinformatics/btg496

Approximate geodesic distances reveal biologically relevant structures in microarray data

Jens Nilsson[1,], Thoas Fioretos[2], Mattias Höglund[2] and Magnus Fontes[1]*

[1] Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden and [2] Department of Clinical Genetics, Lund University Hospital, SE-221 85 Lund, Sweden


DLBCL
Germinal Centre B
Nl Lymph Node/Tonsil
Activated blood B
Resting/activated T
Transformed cell lines
Follicular lymphoma
Resting blood B
CLL
a
Dim2
Dim1
b
Dim2
Dim1