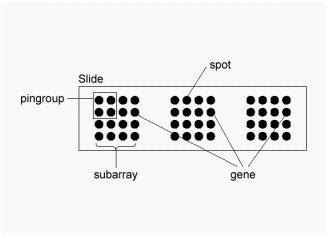
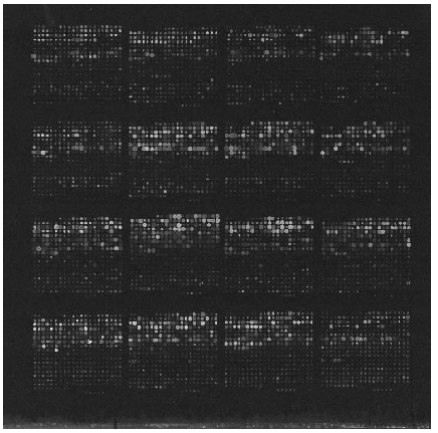
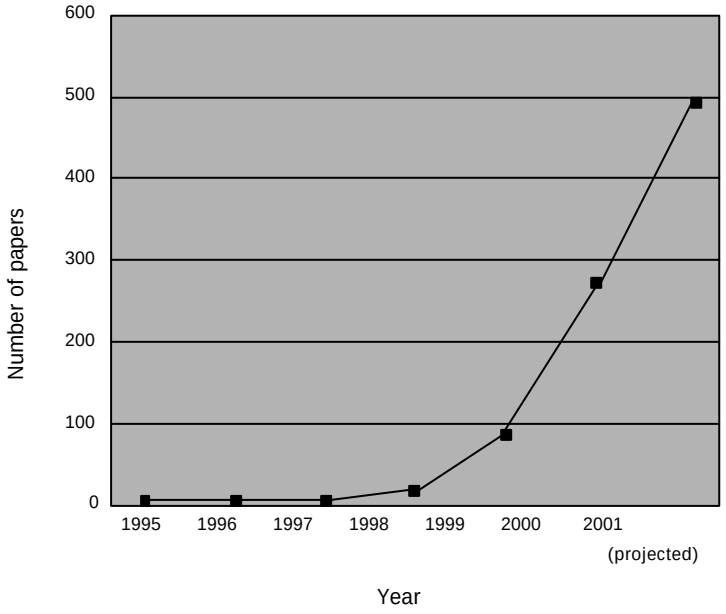
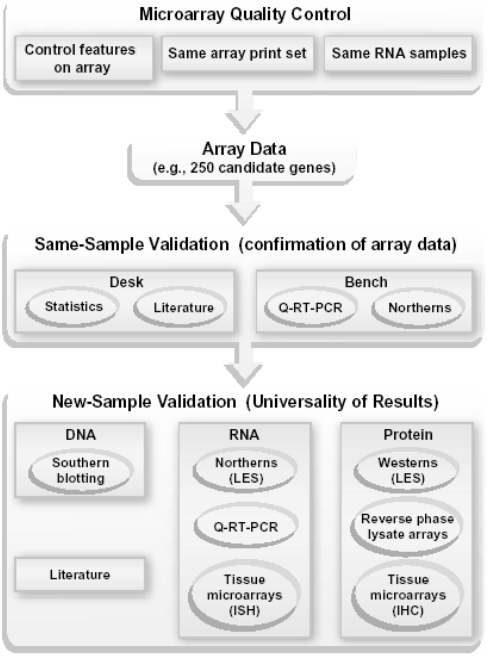
ANÁLISIS ESTADÍSTICO DE EXPRESIÓN GÉNICA

MICROARRAY



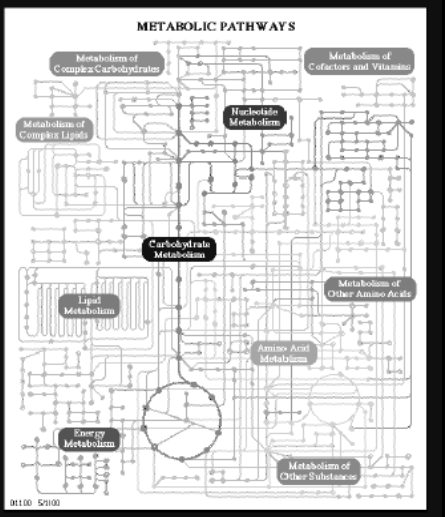



600
500
400
300
200
100
0
Number of papers
1995
1996
1997
1998
1999
2000
2001
(projected)
Year

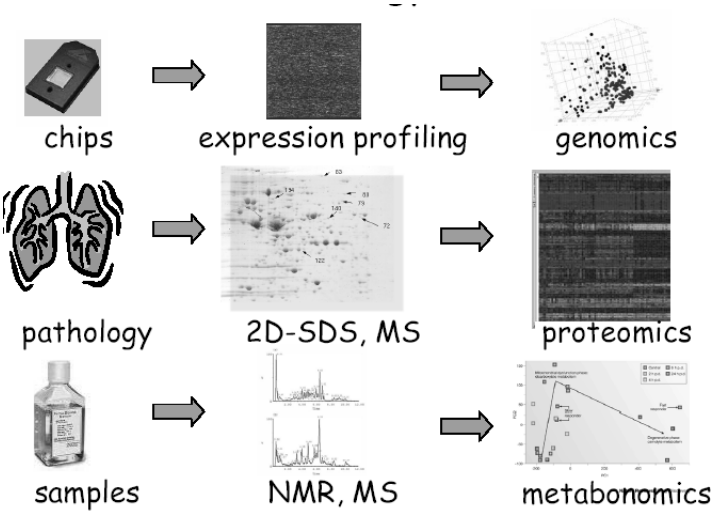

Microarray Quality Control
Control features on array
Same array print set
Same RNA samples
Array Data
(e.g., 250 candidate genes)
Same-Sample Validation (confirmation of array data)
Desk
Statistics
Literature
Bench
Q-RT-PCR
Northerns
New-Sample Validation (Universality of Results)
DNA
Southern blotting
Literature
RNA
Northerns (LES)
Q-RT-PCR
Tissue microarrays (ISH)
Protein
Westerns (LES)
Reverse phase lysate arrays
Tissue microarrays (IHC)

Current complexity

- Human genes (~35,000 - 120,000)
- ~15,000 genes expressed in one cell type
- ~500,000 - 1,000,000 proteins working
  - Alternative splicing
  - Post-translation modification

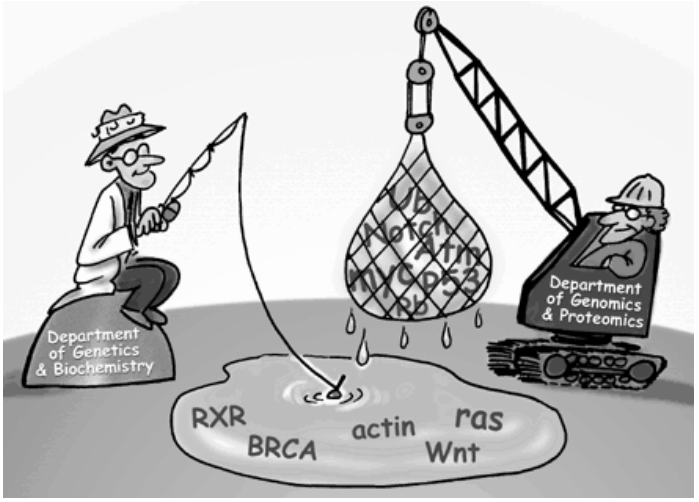


SYSTEM BIOLOGY

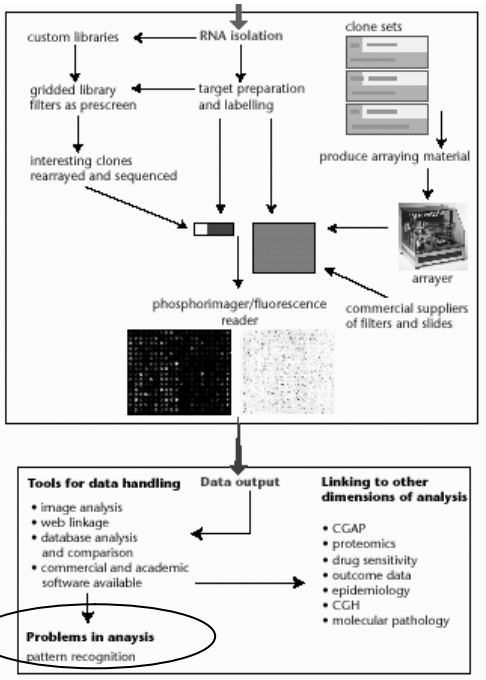


TRADICION

MICROARRAY



PUNTO CRITICO!!!



A sight for sore eyes? Statistical analysis of the data provided by DNA chips can be a major headache.

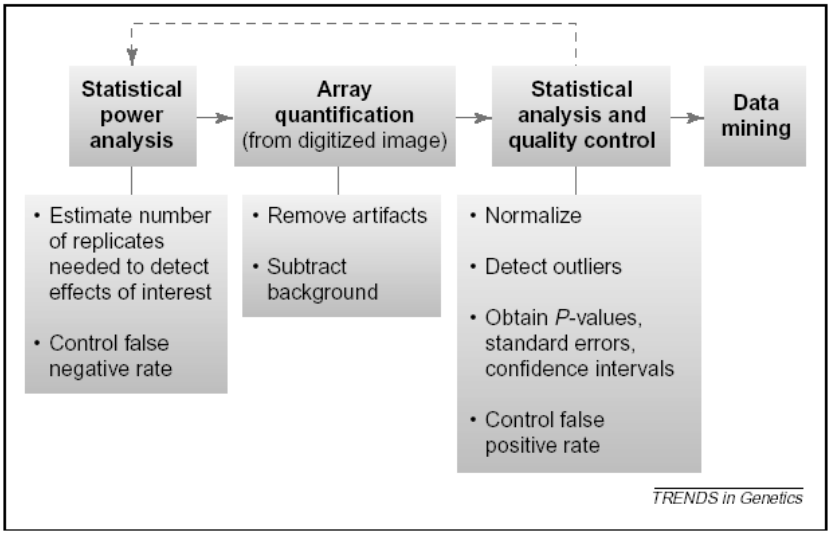


Fig. 1. Data analysis workflow.

COMPUTATIONAL ANALYSIS OF MICROARRAY DATA

John Quackenbush

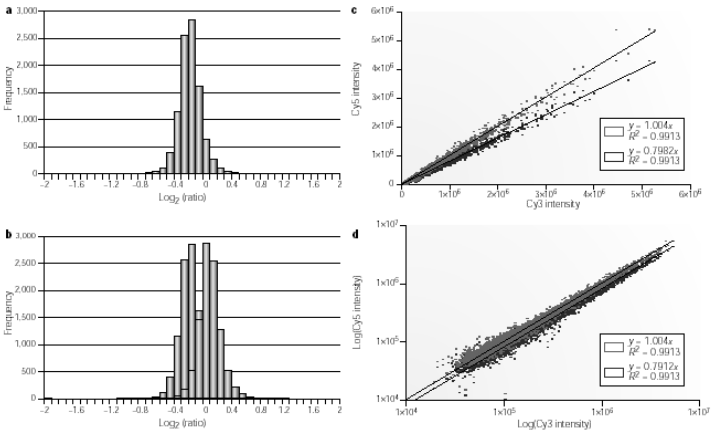
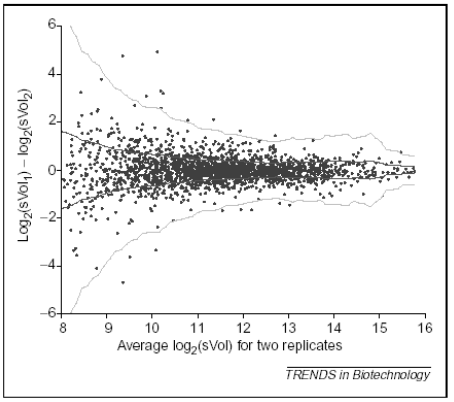


Figure 1 | **Data normalization. a** | A histogram representing the distribution of $\log_2$(ratio) values for a 'self–self' hybridization, in which the measured Cy5 intensity is generally less than the measured Cy3 intensity. Consequently, the $\log_2$(ratio) histogram is centred to the left of zero (as are, indeed, the vast majority of the data). **b** | The same data set shown before (red) and after (blue) normalization to illustrate how the data are transformed. The normalized distribution, shown in blue, is shifted and centred about zero. The perceived change in the shape of the distribution is an artefact of the process of placing data into 'bins' when making the histogram. Scatter plots before (red) and after (blue) normalization of the **c** | measured intensities and **d** | log(intensities) also illustrate the transformation of the data.

OTRO PUNTO NO CONSIDERADO EN ESTA CLASE!!



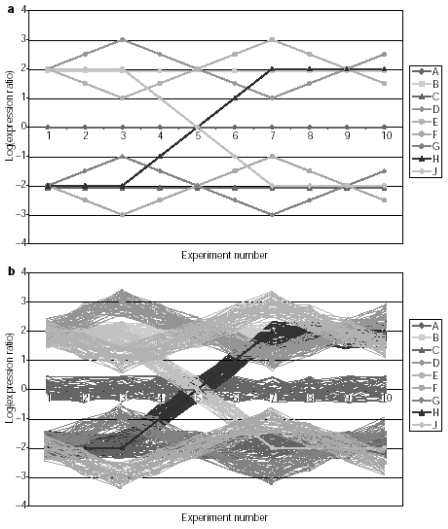
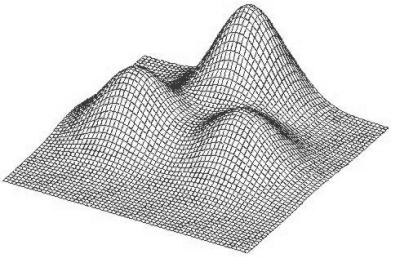


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments. **b** | For each expression pattern, 50 additional genes were generated, representing variations on the basic patterns.

RESPUESTA MULTIVARIADA



BIOINFORMATICS

Editorial

BIOINFORMATICS NEEDS TO ADOPT STATISTICAL THINKING

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

Cluster analysis and display of genome-wide expression patterns

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

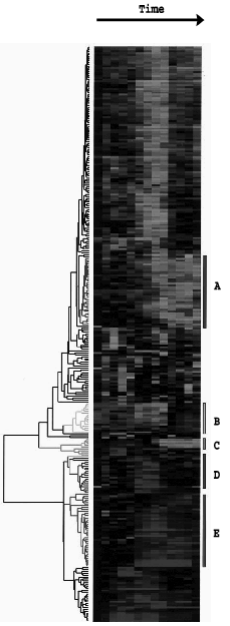
*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

An Information-Intensive Approach to the Molecular Pharmacology of Cancer

Coupled two-way clustering analysis of gene microarray data

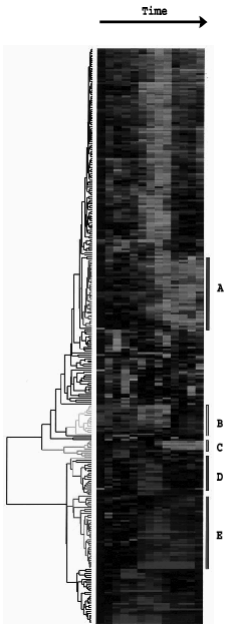
Gad Getz, Erel Levine, and Eytan Domany*

Validating Clustering for Gene Expression Data

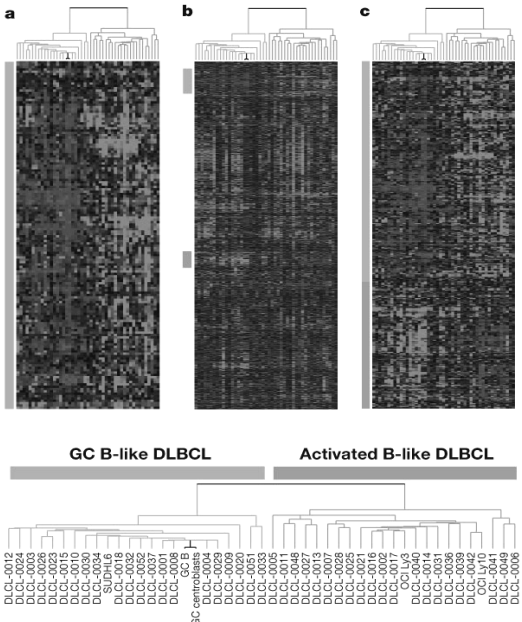


Clusters

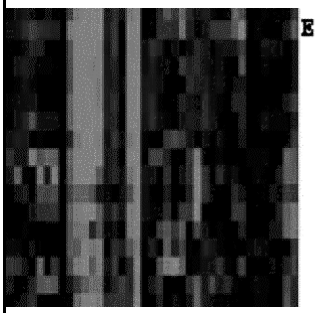
Tomada de
Nature February, 2000
Paper by A Alizadeh *et al*
Distinct types of diffuse large
B-cell lymphoma identified by
Gene expression profiling,



Descubriendo sub-grupos



Secuencias no relacionadas con función similar agrupan juntas



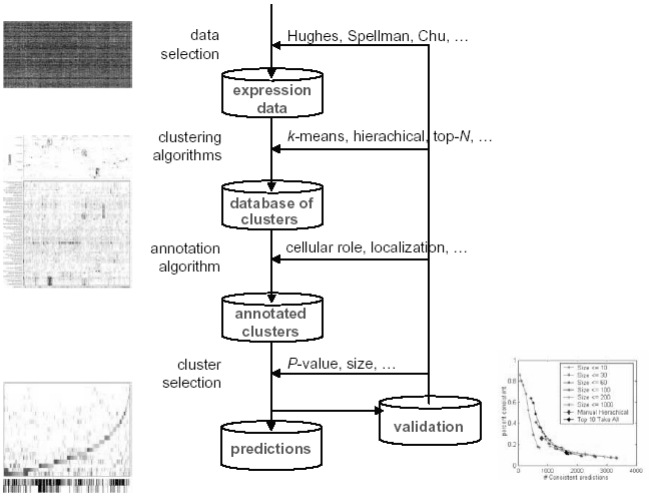
E	TPI1	GLYCOLYSIS	TRIOSEPHOSPHATE ISOMERASE
	GPM1	GLYCOLYSIS	PHOSPHOGLYCERATE MUTASE
	PGK1	GLYCOLYSIS	PHOSPHOGLYCERATE KINASE
	TDH3	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 3
	TDH2	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 2
	ENO2	GLYCOLYSIS	ENOLASE II
	TDH1	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 1
	FBA1	GLYCOLYSIS	ALDOLASE
	TKL1	PENTOSE PHOSPHATE CYCLE	TRANSKETOLASE
	PDC5	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
	PDC6	GLYCOLYSIS	PYRUVATE DECARBOXYLASE 3
	PDC1	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
	CDC19	GLYCOLYSIS	PYRUVATE KINASE
	HXK2	GLYCOLYSIS	HEXOKINASE II
	TYE7	GLYCOLYSIS	BASIC H-L-H TRANSCRIPTION FACTOR
	PFK1	GLYCOLYSIS	PHOSPHOFRUCTOKINASE

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression pattern. Proc. Natl. Acad. Sci. USA 95, 14863-14868.

Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters

Lani F. Wu[1,3*], Timothy R. Hughes[1,2*], Amitaabh P. Davierwala[2], Mark D. Robinson[2], Roland Stoughton[1] & Steven J. Altschuler[1,3]
*These authors contributed equally to this manuscript.

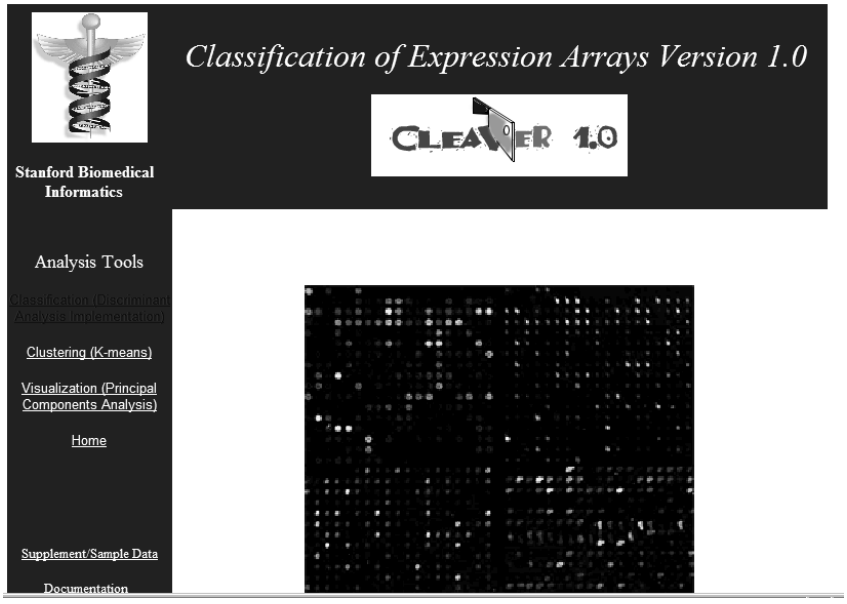
Published online: 24 June 2002, doi:10.1038/ng906



Microarray data analyses (web)

AFM
AMADA
Churchill
CLUSFAVOR
CLUSTER,
D-CHIP
GENE-CLUSTER
J-EXPRESS
PAGE
PLAID
SAM

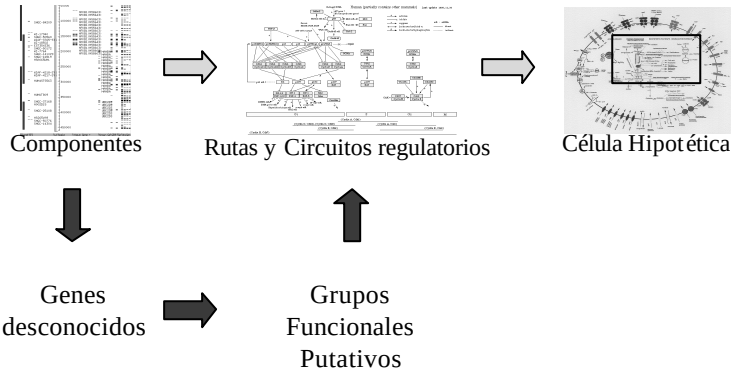
SMA
SVDMAN
TREE-ARRANGE & TREEPS
VERA &SAM
XCLUSTER
ArrayTools
- **ARRAY-VIEWER**
- **F-SCAN**
- **P-SCAN**
- **SCAN-ALYZE**
- **GENEX**
- **MAPS**



Análisis de Cluster

- Buenos cuando señales promedios son tomadas entre los grupos de genes (Eisen)
- Útil cuando se buscan nuevas clases de células, tumores, etc.
- Lleva a una gráfica de interpretación rápida.

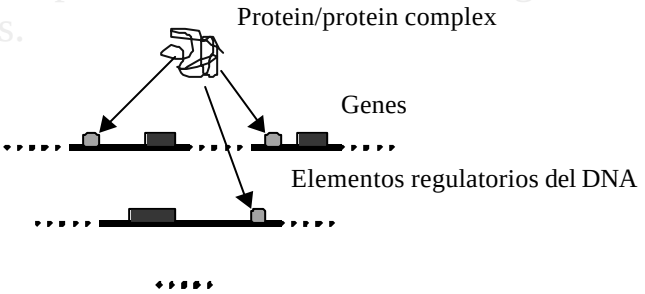
POR QUE AGRUPAR?



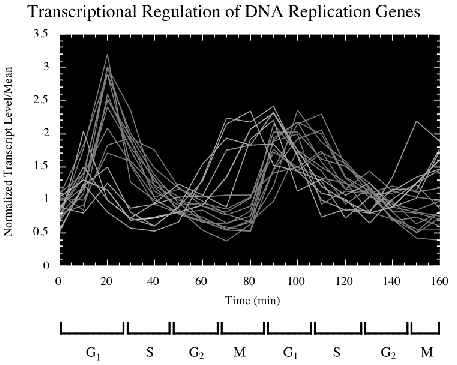
Análisis de Cluster

Propósito ESTADÍSTICO: Dividir las muestras en grupos homogéneos basados en ciertos atributos.

Análisis de Expresión Génica: Encontrar genes co-regulados.



Grupos de genes inducibles son co-regulados

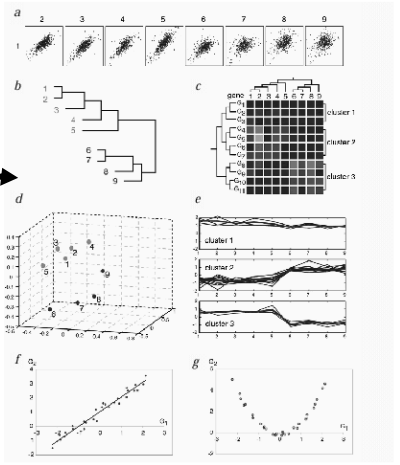


orange=pre-replication complex genes: mcm2, mcm3, cdc46, cdc47, cdc54, cdc6
blue=replication genes involved in DNA synthesis

Figura tomada de
http://genomics.stanford.edu/yeast/additional_figures_link.html

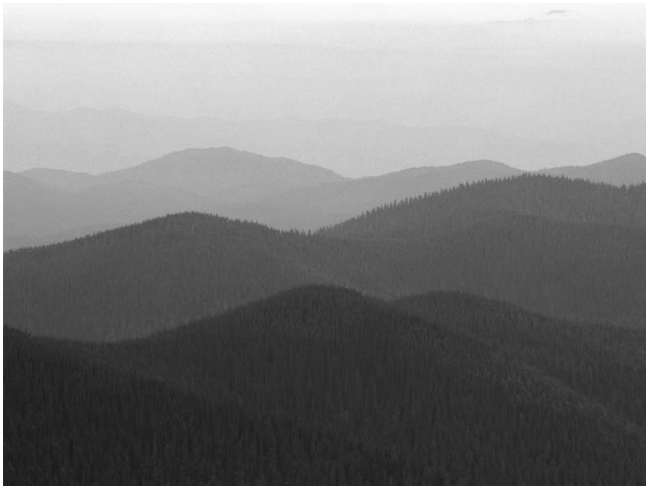
Questions

- **Are there groups in the data?**
  --> Cluster Analysis (groups not known *a priori*).
- **Given the groups, are there differences in the central tendency of the groups?**
  --> ANOVA or MANOVA (experimental manipulation).
- **To which groups does this new individual belong?**
  --> Discriminate Analysis (prediction and explanation).



PCA , SDV, etc.

Panorama de Decisiones para agrupar datos de expresión Génica y construir un árbol



Panorama de Decisiones para agrupar Expresión Génica y construir un árbol

Data Normalization | Distance Metric | Linkage | Clustering Method

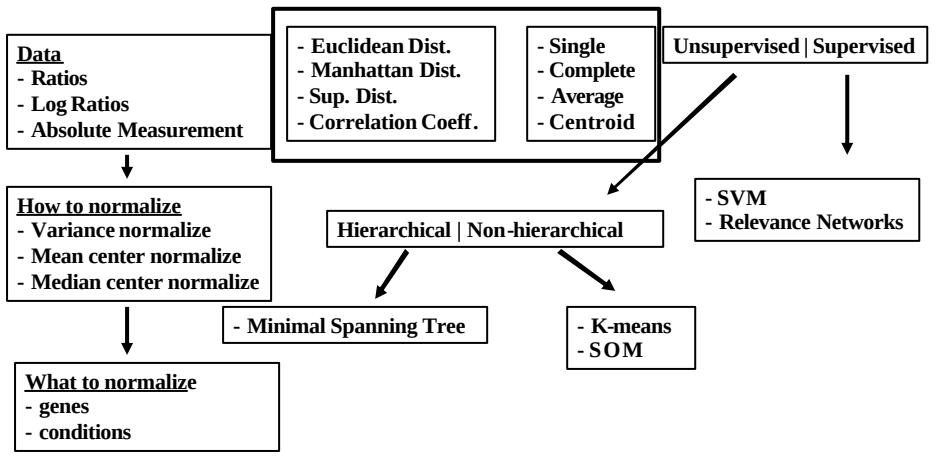


Table 1. Methods used for array data analysis.

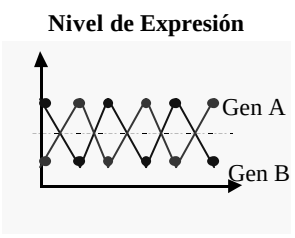
Category	Method
Unsupervised clustering	Hierarchical clustering, *K*-means clustering, MDS, self-organizing maps
Supervised discriminatory gene classifiers	F-test, t-test, Mann-Whitney U-test, Wilcoxon rank score, total number of mis-classifications score, signal-to-noise statistic, MDS weighted gene analysis, ANOVA
Supervised machine learning classifiers	Support vector machines, multi-layer perceptron artificial neural networks

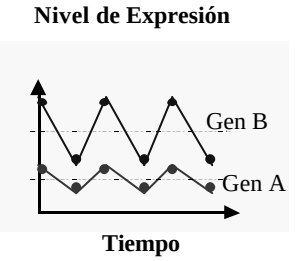
ANOVA: Analysis of variance; MDS: Multi-dimensional scaling.

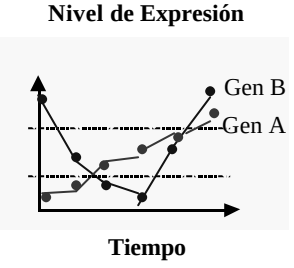
Términos Claves en Análisis de Cluster

- Medidas de similaridad (ej. Pearson, Euclidiana).
- Análisis Jerárquico o no Jerárquico.
- Single/complete/average linkage.
- Dendrograma.

Medidas de Similaridad







Medidas de Distancia : Métrica de Minkowski

Suponga dos objectos x e y ambos con p atributos :

$$x = (x_1, x_2 \dots x_p)$$

$$y = (y_1, y_2 \dots y_p)$$

La métrica Minkowski esta definida por

$$d(x, y) = \sqrt[r]{\sum_{i=1}^{p} | x_i - y_i |^r}$$

Derivados de la Métrica de Minkowski

$1, r = 2$ (Distancia Euclideana)

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p} | x_i - y_i |^2}$$

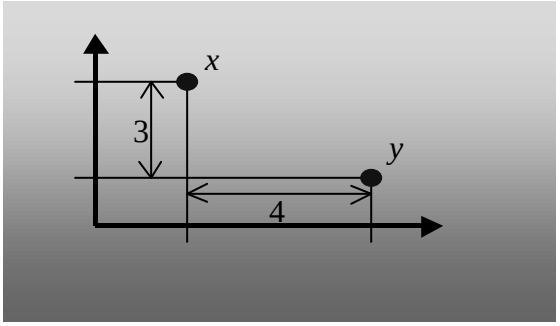
$2, r = 1$ (Distancia de Manhattan)

$$d(x, y) = \sum_{i=1}^{p} | x_i - y_i |$$

$3, r = +\infty$ (Distancia"sup")

$$d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$$

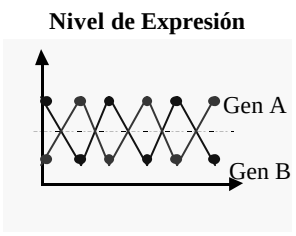
Ejemplo

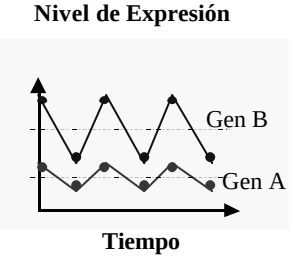


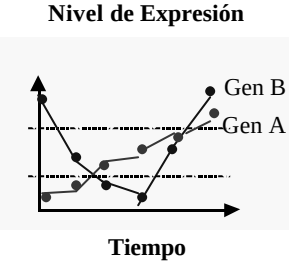
1, Dist. Euclidiana : $\sqrt[2]{4^2+3^2}=5.$

2, Dist. Manhattan : $4+3=7.$

Medidas de Similaridad: Coeficiente de Correlación







Medidas de Similaridad: Coeficiente de Correlación

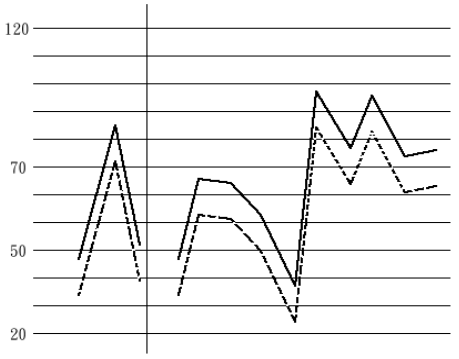
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

$$\text{donde } \bar{x} = \tfrac{1}{p}\sum_{i=1}^{p} x_i \quad \text{y} \quad \bar{y} = \tfrac{1}{p}\sum_{i=1}^{p} y_i.$$

$$|s(x,y)| \leq 1$$

Coeficiente de Correlación de Pearson

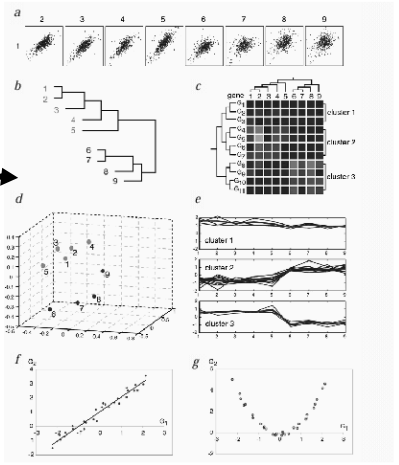
– Captura la similaridad en la “forma” de dos perfiles de expresión e ignora las diferencias entre sus magnitudes.



Questions

- **Are there groups in the data?**
  --> Cluster Analysis (groups not known *a priori*).
- **Given the groups, are there differences in the central tendency of the groups?**
  --> ANOVA or MANOVA (experimental manipulation).
- **To which groups does this new individual belong?**
  --> Discriminate Analysis (prediction and explanation).

PCA , SDV, etc.

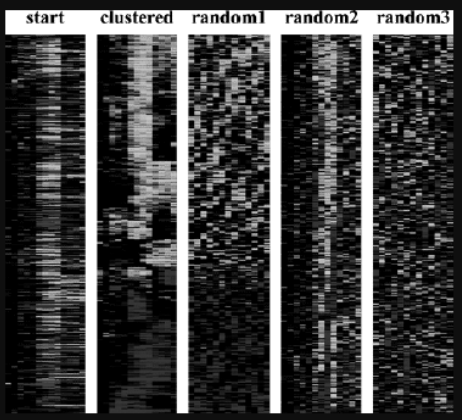


Clustering

- Distance measure: correlation coefficient(Pearson correlation)

$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sigma_x \sigma_y}$$

- Clustering of genes by expression patterns
  - (Eisen et al 1998, Spellman et al 1998, Michaels 1998, Petri et al 1999, Alon et al 1999)



Agrupación de datos

Métodos no supervisados:

-Cluster Jerárquico
-K-means
-SOM

CLUSTER JERARQUICO

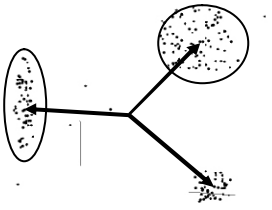
-Elección de medida de similaridad/distancia

-Elección del criterio de unión

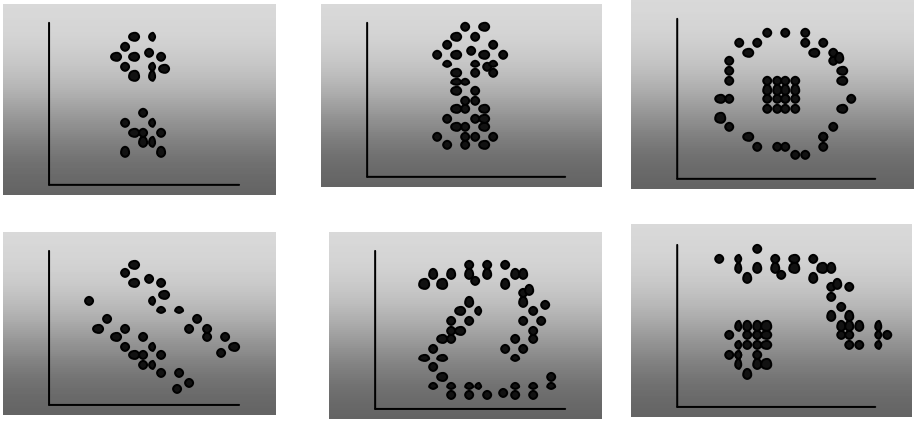
Agrupamiento Jerárquico y no Jerárquico.

•Jerárquico: una serie de sucesivas UNIONES de datos… hasta que un número de final cluster es obtenido.

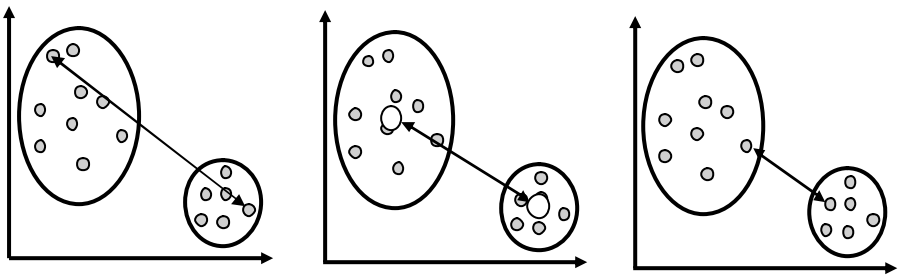
•No- Jerárquico: ej. K-mean: *K clusters* son elegidos de modo que tales puntos son mutuamente los más distantes.
Cada componente en la población asignada a cada cluster es asignado a un cluster por la mínima distancia. La posición del centroide es recalculada y esto se repite hasta que todos los componentes son agrupados. El criterio es la minimización de la varianza



¿Cuál método entonces se puede sugerir como el más adecuado para este tipo de datos?

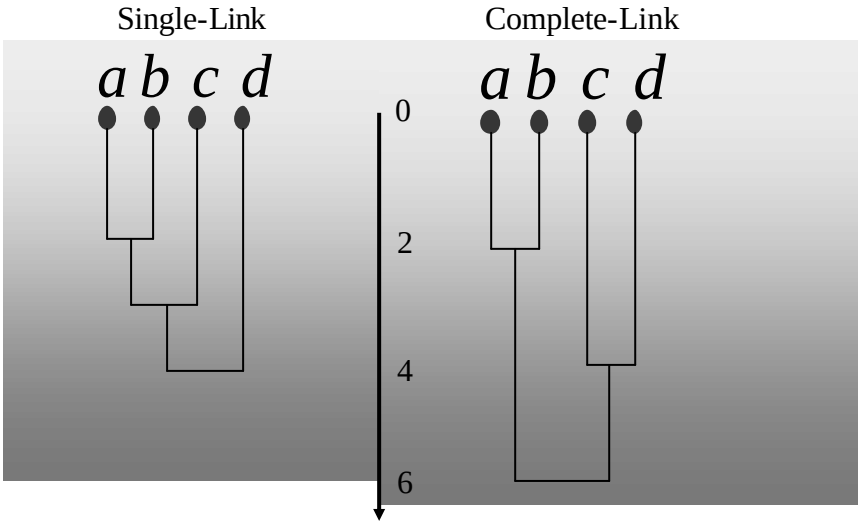


Elección criterio de unión

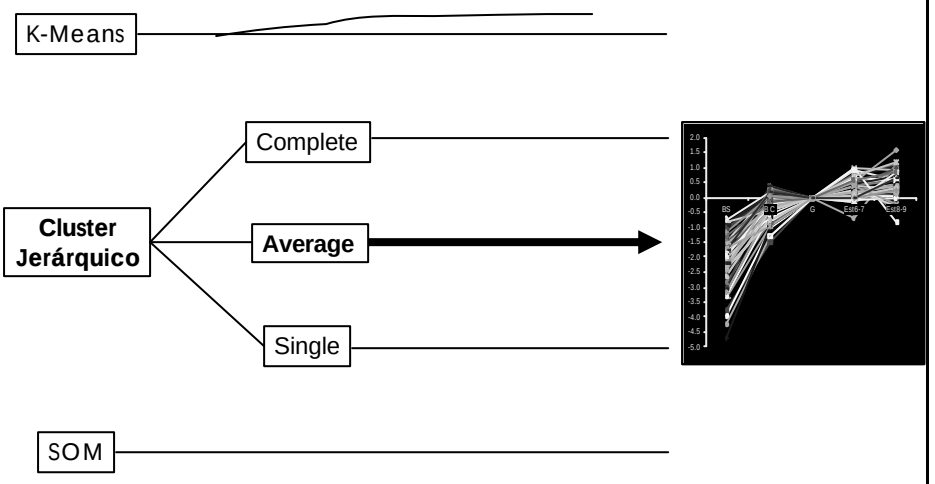


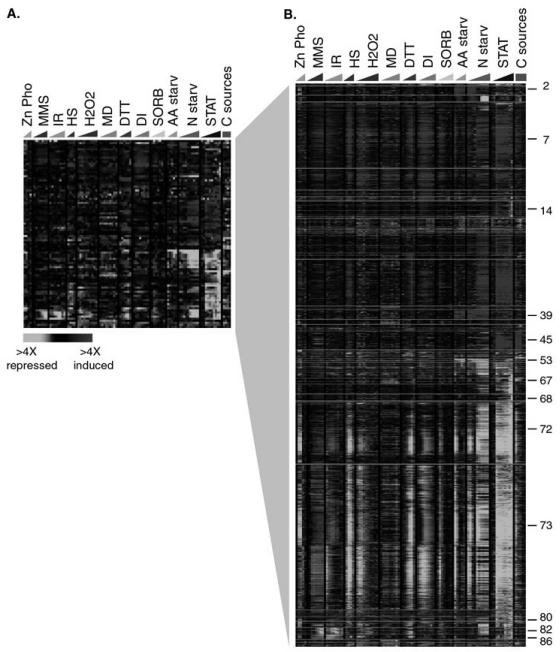
Complete Linkage **Average Linkage** **Single Linkage**

Dendrogramas

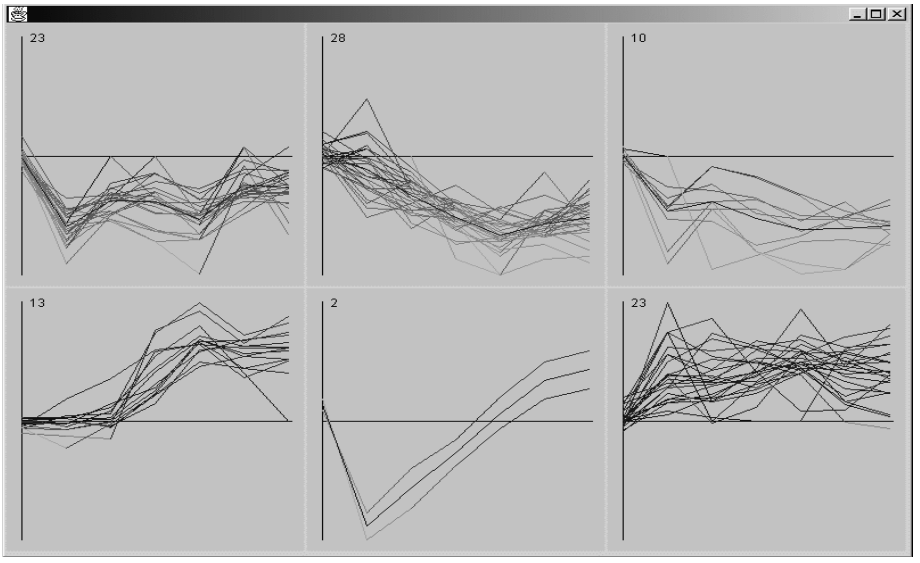


METODOS




A.
B.
Zn Pho
MMS
IR
HS
H2O2
MD
DTT
DI
SORB
AA starv
N starv
STAT
C sources
>4X repressed
>4X induced
2
7
14
39
45
53
67
68
72
73
80
82
86

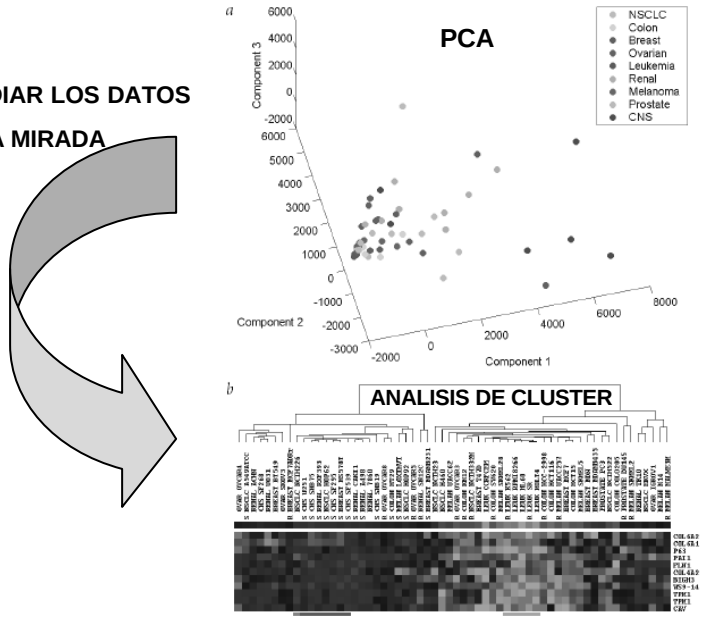
Resultados??


23
28
10
13
2
23

EVALUACION DE LOS CLUSTER

VOLVER A ESTUDIAR LOS DATOS

CON OTRA MIRADA



PRINCIPAL COMPONENT ANALYSIS (PCA)

FEBS Letters 507 (2001) 114–118 FEBS 25367

The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data

Marco Crescenzi, Alessandro Giuliani*

PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES

Soumya Raychaudhuri*, Joshua M. Stuart*, and Russ B. Altman*

Stanford Medical Informatics

Stanford University, 251 Campus Drive, MSOB X-215, Stanford CA 94305-5479

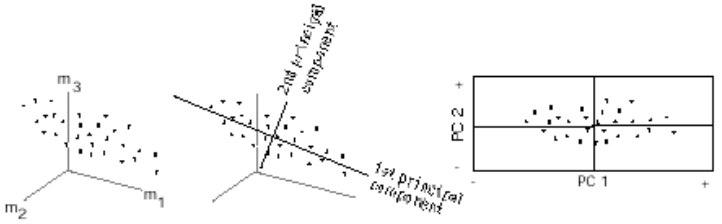
{sxr, stuart, altman} @smi.stanford.edu

PRINCIPAL COMPONENT ANALYSIS

- Es una técnica general de reducción de la dimensionalidad de los datos. Esta libre de supuestos (no docima hipótesis). Transforma un grupo de variables en otro nuevo set de variables.
- Se extraen eigenvectors y eigenvalues (“eigengenes”) a partir de una matriz de correlación o covarianza.
- Eigenvectors describen la posición de los componentes principales en el espacio m-dimensional, y los eigenvalues son proporcionales al largo de cada eje de componentes principales.

- The principal components are axes (=vectors) which define the maximum amount of variance in a data set.
- Can have up to m (the number of variables) principal components, but for most cases usually only the first few describe most of the variation. It is therefore a data reduction technique- reducing the number of dimensions (variables).
- Each principal component is orthogonal to all others and describes an increasingly smaller proportion of the variance.

e.g. with an original 3-dimensional data set (m=3)



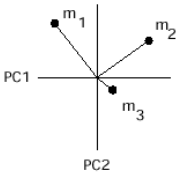
- Principal component scores are the object values onto the principal components.

Interpretation

Each eigenvector (principal component) is comprised of "loadings" or normalized "coefficients" for each variable.

The loadings can be either + or - and indicate the contribution of each variable for that particular principal component.

Common to look at plots of variable loadings (frequently as vectors) in principal component space.



m_1	has a high - contribution on PC1 and a high + contribution on PC2
m_2	has a high + contribution on PC1 and a high + contribution on PC2
m_3	has a low + contribution on PC1 and a low - contribution on PC2

The principal component loadings may show relationships between variables. Variables that plot together may be highly correlated.

PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES

Soumya Raychaudhuri*, Joshua M. Stuart*, and Russ B. Altman*
Stanford Medical Informatics
Stanford University, 251 Campus Drive, MSOB X-215, Stanford CA 94305-5479
{sxr, stuart, altman} @smi.stanford.edu

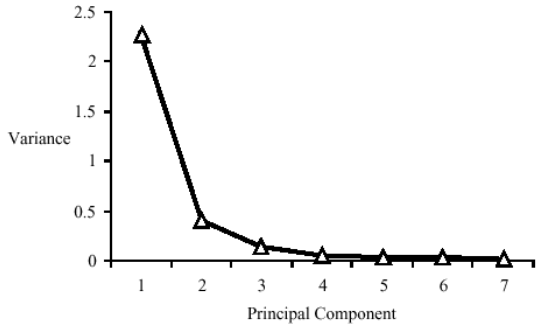
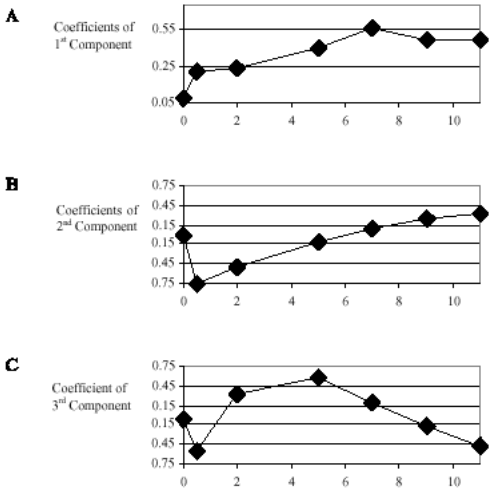


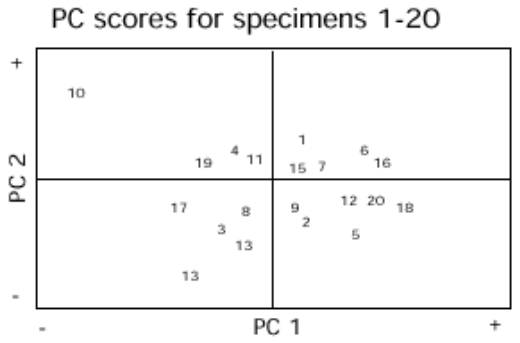
Figure 1. Plot of eigenvalues of the principal components. Most of the variance in the sporulation data set is contained in the first two principal components.

PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES

Soumya Raychaudhuri*, Joshua M. Stuart*, and Russ B. Altman*
Stanford Medical Informatics
Stanford University, 251 Campus Drive, MSOB X-215, Stanford CA 94305-5479
{sxr, stuart, altman} @smi.stanford.edu



e.g.



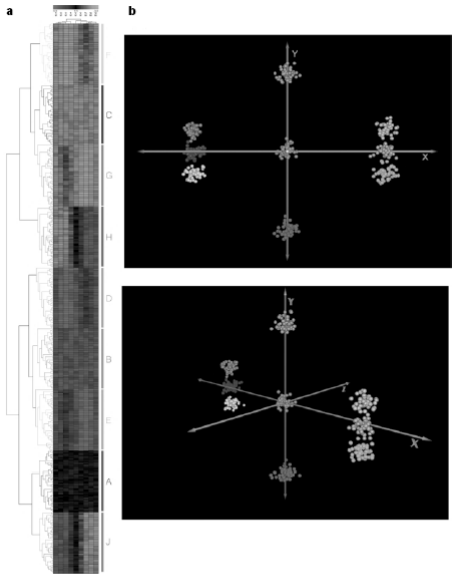
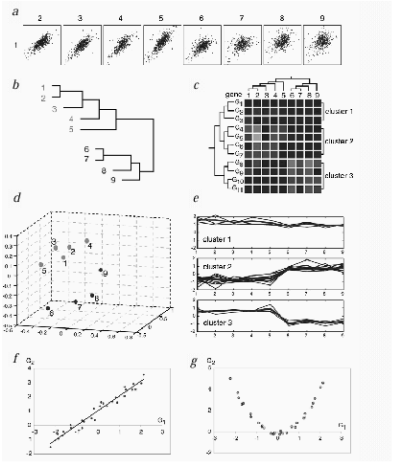


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using a | hierarchical (average-linkage) clustering and b | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

Questions

- **Are there groups in the data?**
  --> Cluster Analysis (groups not known *a priori*).
- **Given the groups, are there differences in the central tendency of the groups?**
  --> ANOVA or MANOVA (experimental manipulation).
- **To which groups does this new individual belong?**
  --> Discriminate Analysis (prediction and explanation).



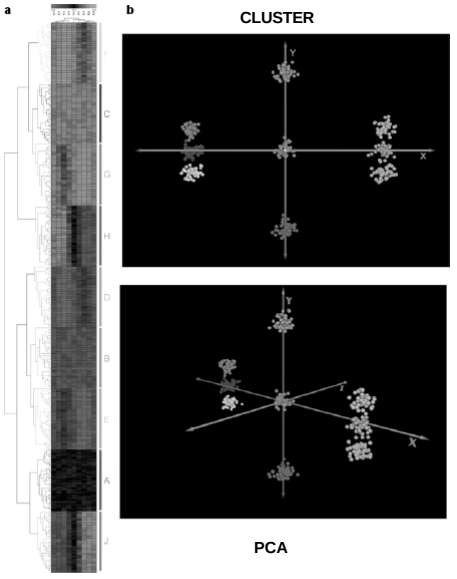


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using a | hierarchical (average-linkage) clustering and b | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

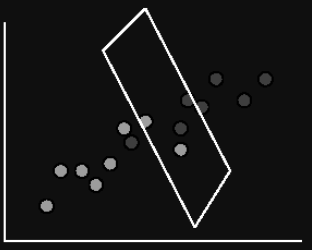
INCORPORARLOS EN EL MISMO ANÁLISIS?

ANALISIS DE COMPONENTES PRINCIPALES

- Es una técnica general de reducción de la dimensionalidad de los datos. Esta libre de supuestos (no docima hipótesis).
- Nuestra pregunta: Validación estadística de los grupos reconocidos en el Análisis de Cluster.
- Necesitamos otro tipo de análisis que docime hipótesis.

Discriminant analysis (linear or non-linear)

- Traditional statistics
  - Bayesian
  - Mahalanobis' distance
- Neural net
- Hidden Markov Model
- SVM(Support Vector Machine)



Comparison of Discrimination Methods for the Classication of Tumors Using Gene Expression Data

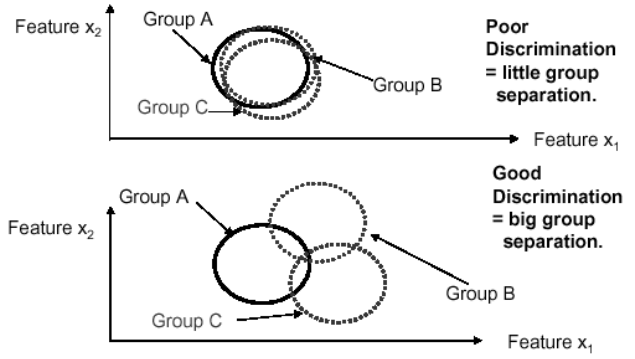
Sandrine Dudoit*
Mathematical Sciences Research Institute, Berkeley, CA.

Discriminate Analysis seeks to:

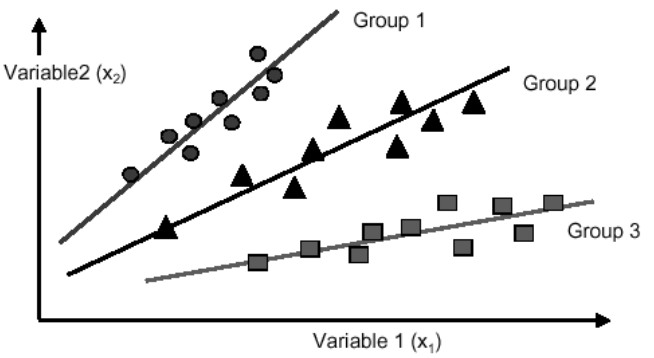
- Establish relationships useful for classifying objects/individuals into one of several populations based on multi-dimensional observations.
- The relationship is between a group membership label (a categorical response) $\mathbf{y}$ and a p-dimensional *feature* vector $\mathbf{x}$.

Group membership is defined before the analysis begins.

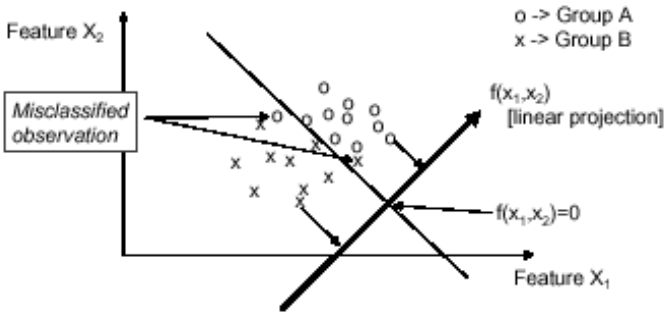
Three groups - Two features



Linear Relationship Example



Linear Discrimination Rule



Rule: Project data onto the line. Object is in A if $f(x_1,x_2) > 0$.

Try to get as few miss-classifications as possible.

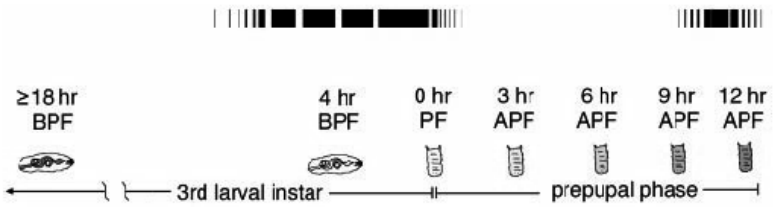
Discriminant analysis to evaluate clustering of gene expression data

Marco Méndez, Christian Hödar, Chris Vulpe[†] , Mauricio González, and Verónica Cambiazo[*]

Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile. Macul 5540, Macul, Santiago, Chile. [†]Nutrition and Toxicology, 119 Morgan Hall, University of California, Berkeley, USA

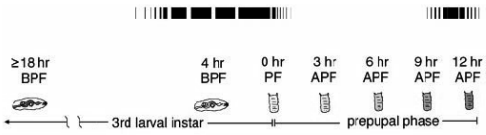
Microarray Analysis of *Drosophila* Development During Metamorphosis

Kevin P. White,* Scott A. Rifkin,† Patrick Hurban,‡
David S. Hogness



Microarray Analysis of *Drosophila* Development During Metamorphosis

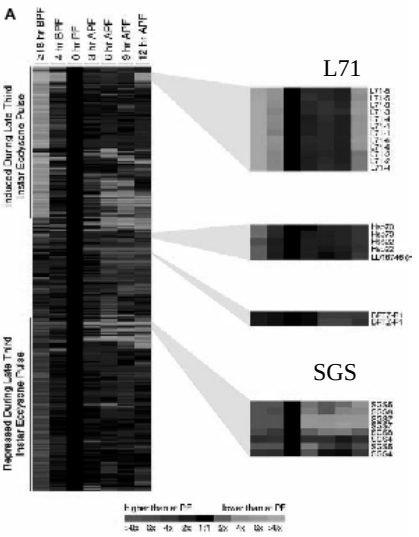
Kevin P. White,* Scott A. Rifkin,† Patrick Hurban,‡
David S. Hogness



Se seleccionó dos niveles de análisis:

-Global (todos los genes)

-Dos nodos en particular



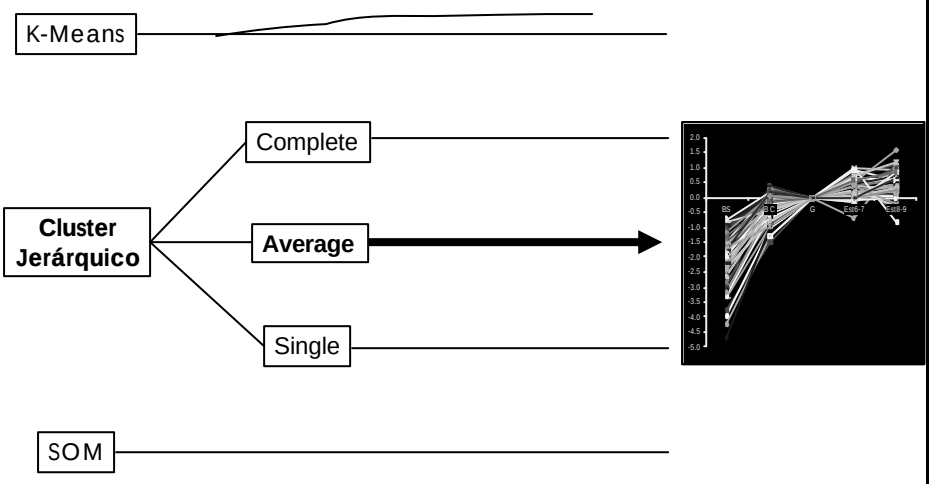
Utilización combinada de Análisis de Componentes Principales (PCA) y Análisis Discriminante (DA)

PCA:

-Disminución de la dimensionalidad de los datos

DA:

-Verificación estadística de la existencia de los grupos resueltos por análisis de cluster, utilizando distintos métodos de agrupamiento
-Asignación de probabilidad a los elementos que forman cada grupo



Agrupación de datos

Métodos no supervisados:

-Cluster Jerárquico
-K-means
-SOM

CLUSTER JERARQUICO

-Elección de medida de similaridad/distancia

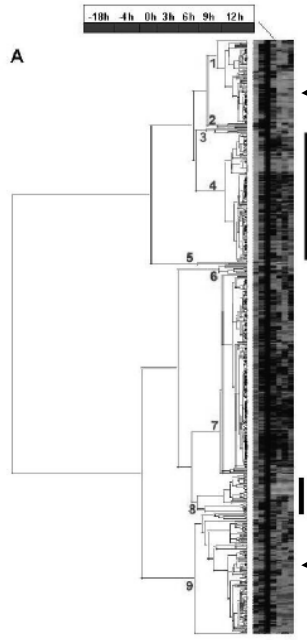
-Elección del criterio de fusión

Análisis de Cluster

Opciones

Average-linkage.

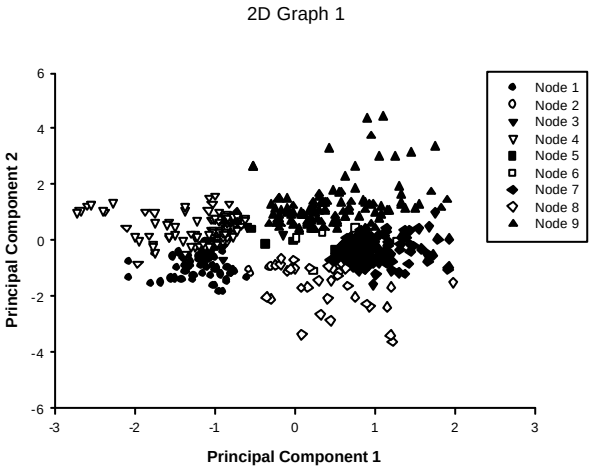
Distancia Euclidiana



Baja expresión

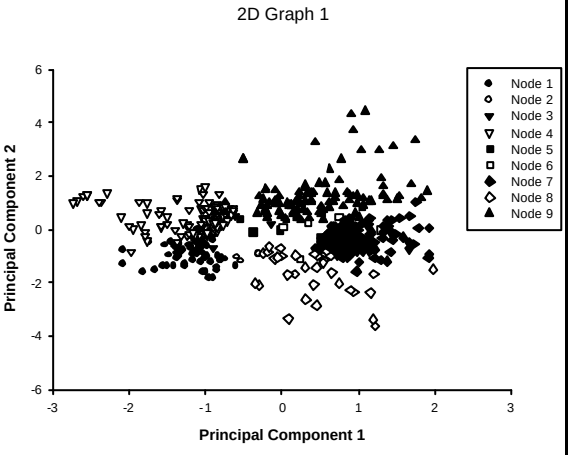
Alta expresión

Análisis Global



Resultados PCA

	% Variabilidad
1a Componente	55.5
2a Componente	27.2
3a Componente	9.2
Total	91.9



-Los datos pueden ser llevados a un nuevo espacio que resume mas del 90% de la variabilidad total del experimento.

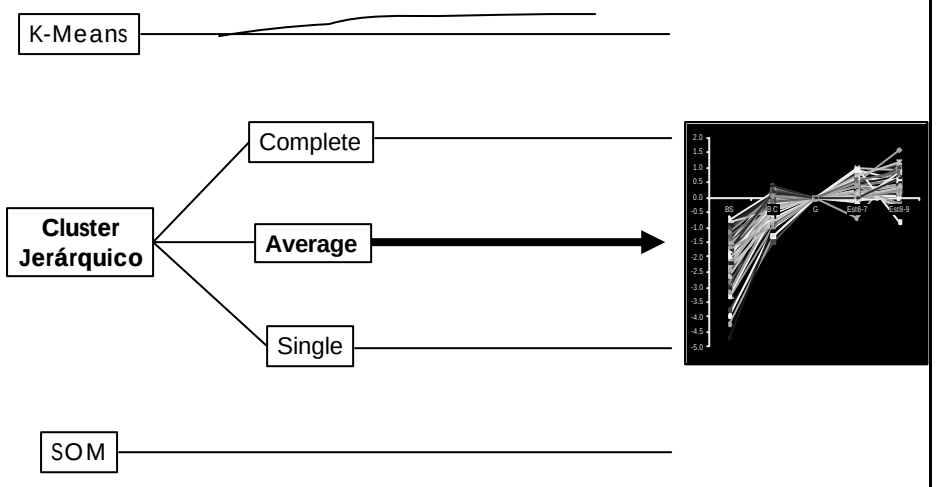
Resultados PCA

	% Variabilidad
1a Componente	55.5
2a Componente	27.2
3a Componente	9.2
Total	91.9

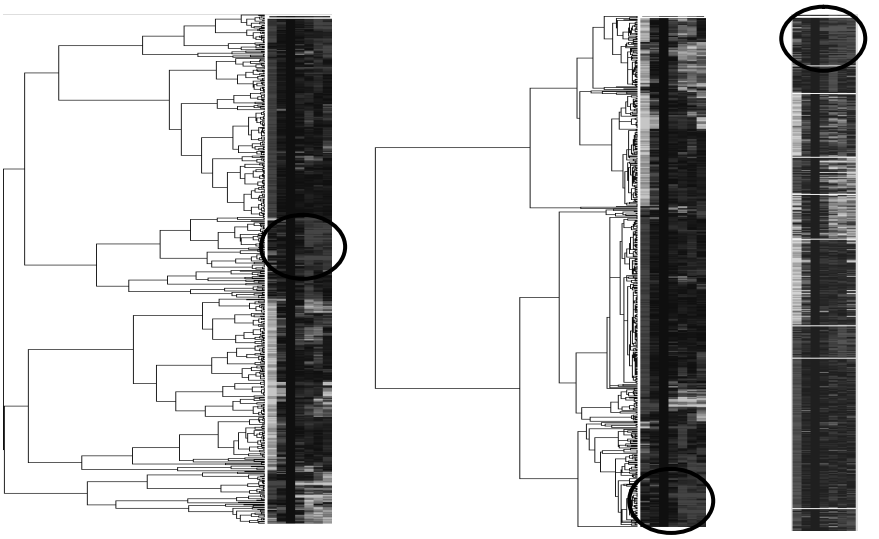
Resultados DA

	Average **(p<0.0001)**	
Grupos	Nº Elementos	% Elementos bien clasificados
1	74	91.9
2	4	100
3	5	80
4	115	59.1
5	3	100
6	9	44.4
7	180	68.3
8	34	35.3
9	110	24.5
Total	**534**	**58.6**

-**Los grupos seleccionados del análisis de cluster son estadisticamente confiables y distintos.**

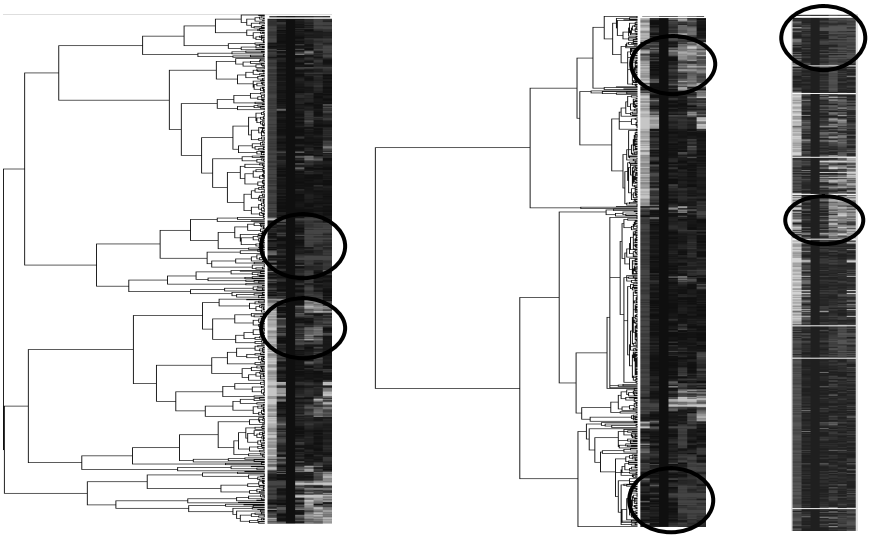


EFECTO METODO DE AGRUPAMIENTO

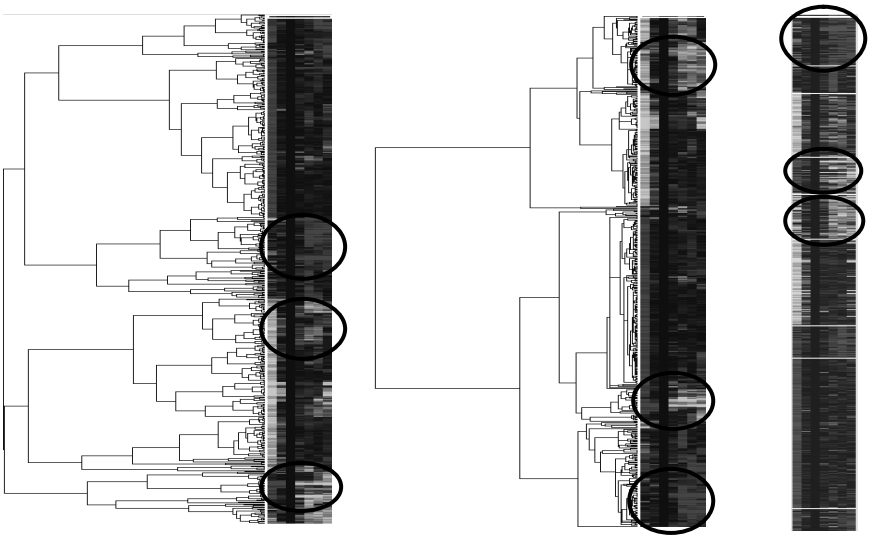


Complete Linkage **Average Linkage** **K-means**

EFECTO METODO DE AGRUPAMIENTO



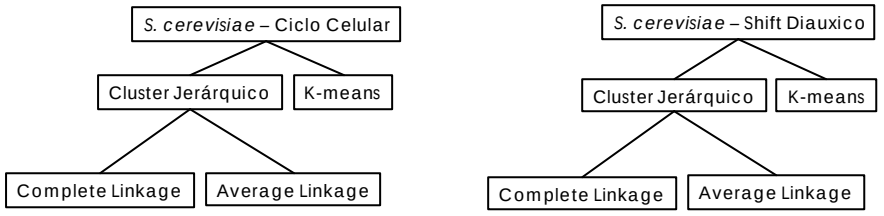
Complete Linkage **Average Linkage** **K-means**



Complete Linkage **Average Linkage** **K-means**

Los resultados obtenidos, ¿son propiedad del sistema biológico o consecuencia de la elección de las variables usadas en el análisis?

Dos bases de datos provenientes de experimentos en *S. cerevisiae:*



Resultados

1. Ciclo celular

	Nº Grupos
Complete	5
Average	12

	Complete P<0.001	Kmeans(5) P<0.001	Average P<0.001	Kmeans (12) P<0.001
Nº Inicial	831	831	831	831
Nº bien clasificados	623 **75.0%**	759 **91.3%**	596 **71.7%**	657 **79.1%**

	Nº Elementos comunes	% del total inicial
Average - Complete	435	**52.3%**
Average – Kmeans(5)	559	**67.3%**
Complete – Kmeans (9)	468	**56.3%**
Todos	320	**38.5%**

Resultados

2. Shift Diauxico

	Nº Grupos
Complete	5
Average	9

	Complete P<0.001	Kmeans(5) P<0.001	Average P<0.001	Kmeans (12) P<0.001
Nº Inicial	731	731	731	731
Nº bien clasificados	713 **97.5%**	668 **91.4%**	607 **83.0%**	632 **86.4%**

	Nº Elementos comunes	% del total inicial
Average - Complete	589	**80.1%**
Average – Kmeans(5)	649	**88.8%**
Complete – Kmeans (9)	538	**74.0%**
Todos	492	**67.3%**

Conclusiones

1. **El uso combinado de la técnica de Análisis de Componentes Principales y Análisis Discriminante permita entregar un soporte de confianza estadística a los grupos que se seleccionan mediante un análisis de cluster jerárquico.**

2. **El uso del Análisis Discriminante a través de la identificación de elementos correctamente asignados a un grupo y la reducción del volumen de información, posibilita el posterior análisis de estos elementos por técnicas biológicas tradicionales.**

3. **La búsqueda de consistencia mediante el uso paralelo de otras formas de asociación, permite determinar aquellos elementos que con mas fuerza podrían explicar el fenómeno estudiado, independiente del origen biológico de los datos.**