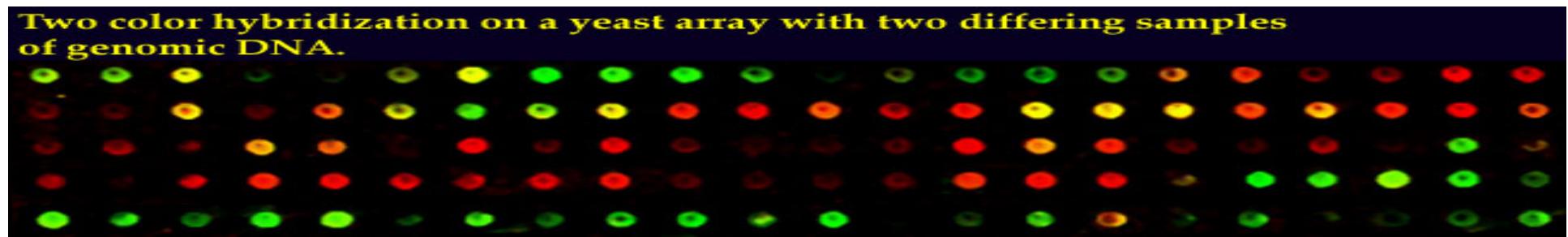
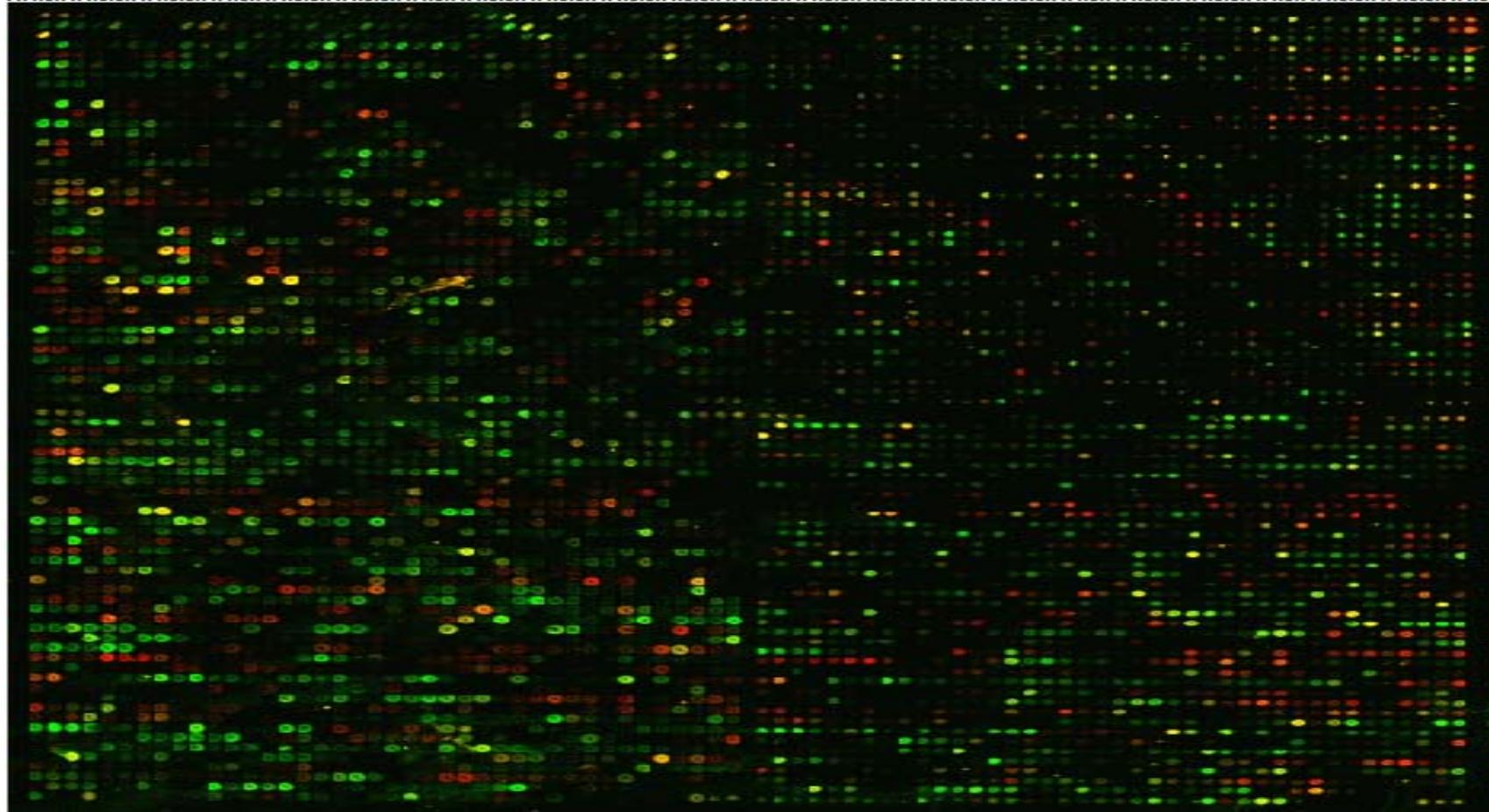


Genómica: Estudio de la expresión génica.

- El **patrón de genes** expresados en una célula brinda información del estado actual de la misma.
- Existe una correlación entre el estado de la célula y los **cambios en los niveles de mRNA** de muchos genes.
- Los patrones de expresiones de genes que nunca han sido caracterizados puede brindar nuevas pistas sobre su posible **función**.

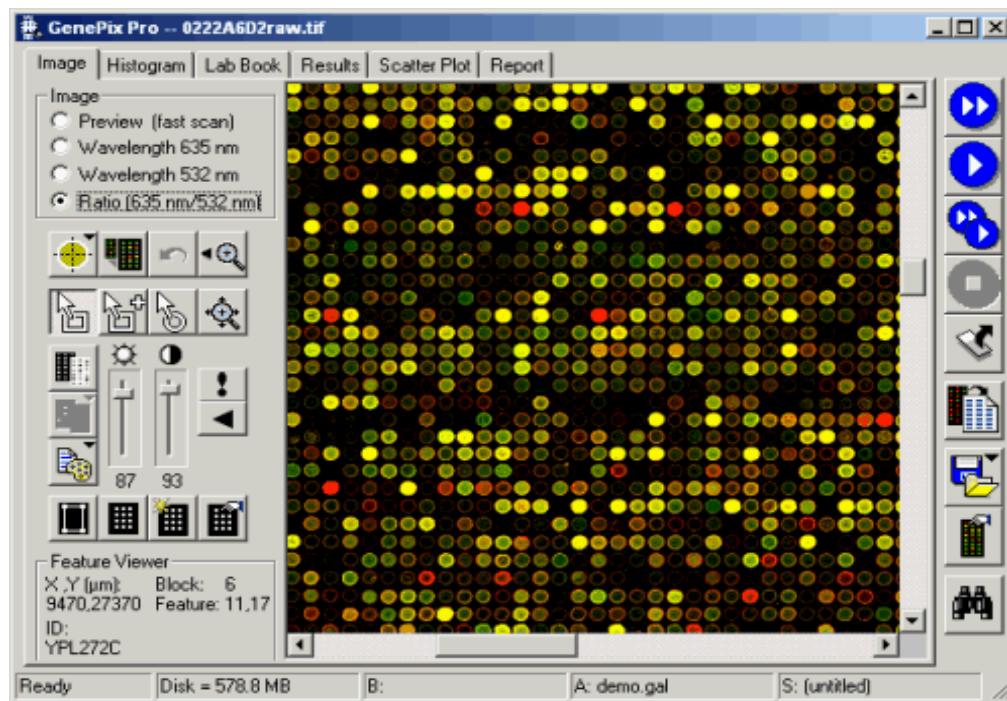
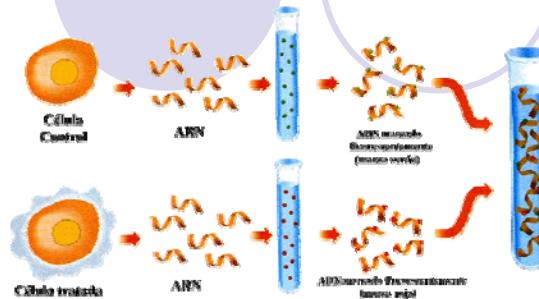


Genómica: Tecnología DNA Microarrays.



S

Genómica: Tecnología DNA Microarrays.

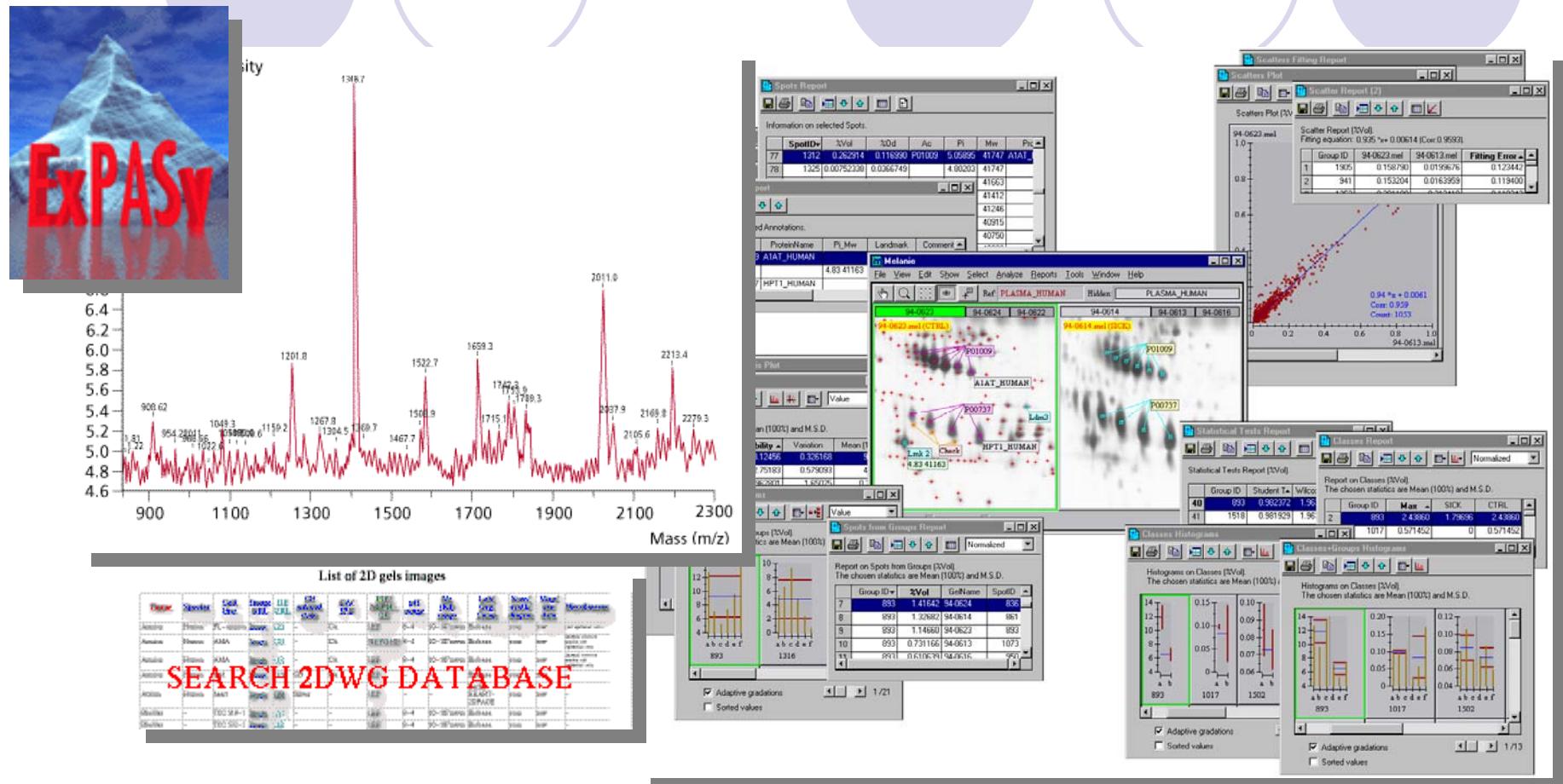


- Almacenamiento
- Análisis de Datos
- Visualización
- Interpretación/Anotación
- Publicación en repositorio público

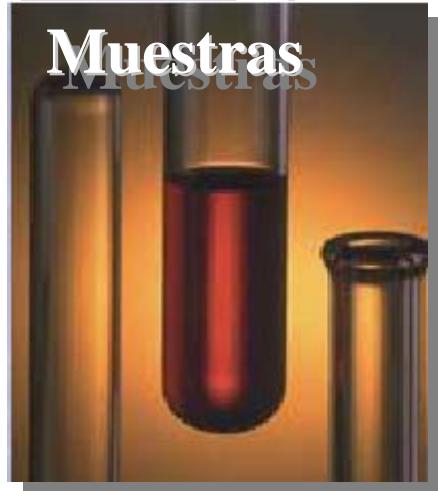
Proteómica: del Gen a la proteína.

- El **proteoma** es una imagen dinámica de todas las **proteínas** expresadas en un organismo, célula o compartimento subcelular concreto en un momento dado y bajo determinadas condiciones
- La **proteómica** se encarga del estudio y caracterización del proteoma

Proteómica.



LIMS: Laboratory Information Management Systems.

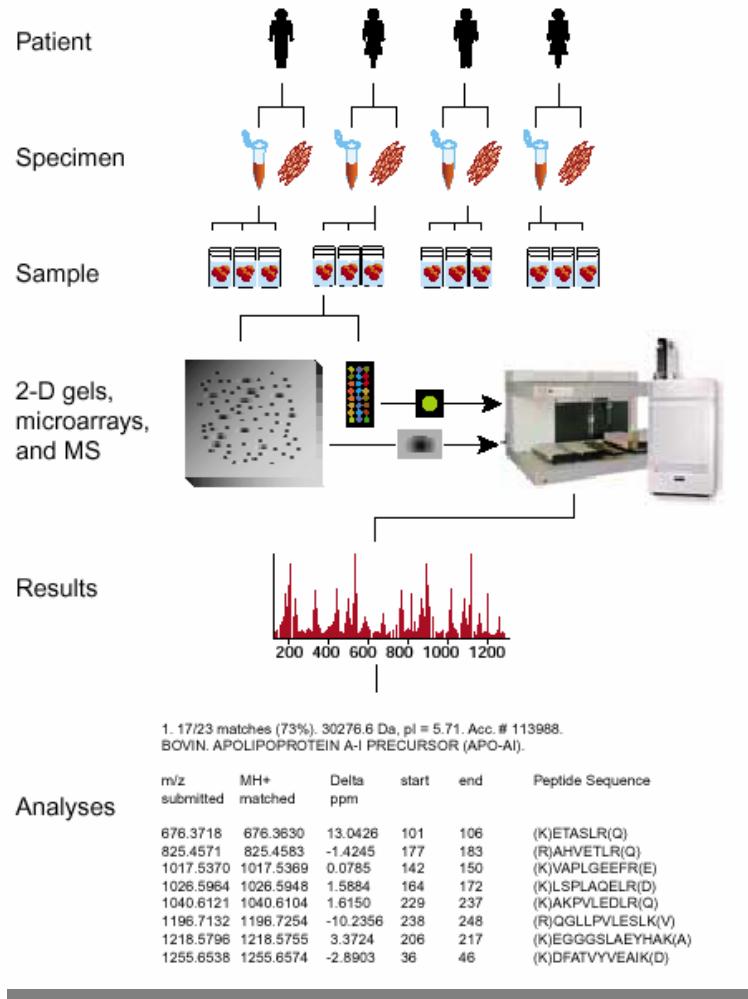


HEADER HORMONE
TITLE NMR STRUCTURE OF HUMAN INSULIN; 2HNU
TITLE 2 ZINC-FREE, 10 STRUCTURES
COMPND MOLECULE: INSULIN;
COMPND T: HETERO DIMER
SOURCE HOMO SAPIENS;
SOURCE MAN
KEYWDS GLUCOSE METABOLISM
MODEL
ATOM 1 N GLY A 1 6.735 1.016 1.00 0.00 N
ATOM 2 CA GLY A 1 -4.686 6.753 1.376 1.00 0.00 C
ATOM 3 C GLY A 1 -3.864 6.149 0.235 1.00 0.00 C
ATOM 4 O GLY A 1 -3.324 6.855 -0.593 1.00 0.00 O
ATOM 5 1H GLY A 1 -6.407 5.776 0.726 1.00 0.00 H

Datos:
- ficheros
- bases de datos



LIMS: Automatización completa en Proteómica.



Esquema general

Preparación
de la muestra

Separación

Identificación

Caracterización

Requerimientos computacionales y matemáticos.

- Gestión de Datos: Sistemas de Gestión de Información para laboratorios (LIMS)
- Bases de datos
- Integración de fuentes de información.
- Reconocimiento de patrones
- Aprendizaje Automático
- Redes Neuronales
- Estadística
- Sistemas Expertos
- Minería de Datos biológicos y Texto
- Procesamiento de Imagen y Señal
- Computación de alto rendimiento: paralelismo y mallas

Tecnologías de Información y Ciencias de la Vida.

- La producción de datos en Ciencias de la Vida es cada vez más abundante en gran parte debido a la automatización de procesos experimentales.
 - ➡ La creación de Bases de Datos es un componente crítico en el flujo de trabajo de Ciencias de la vida.
- Los datos producidos a veces son de tipos complejos (imágenes, video).
 - ➡ Los motores de BDs deben ser suficientemente flexibles (p.ej.: OO-Relacional).
- La disponibilidad de los datos debe ser inmediata y global a la compañía y/o a la comunidad científica.
 - ➡ Publicación y distribución Web.

Tecnologías de Información y Ciencias de la Vida.

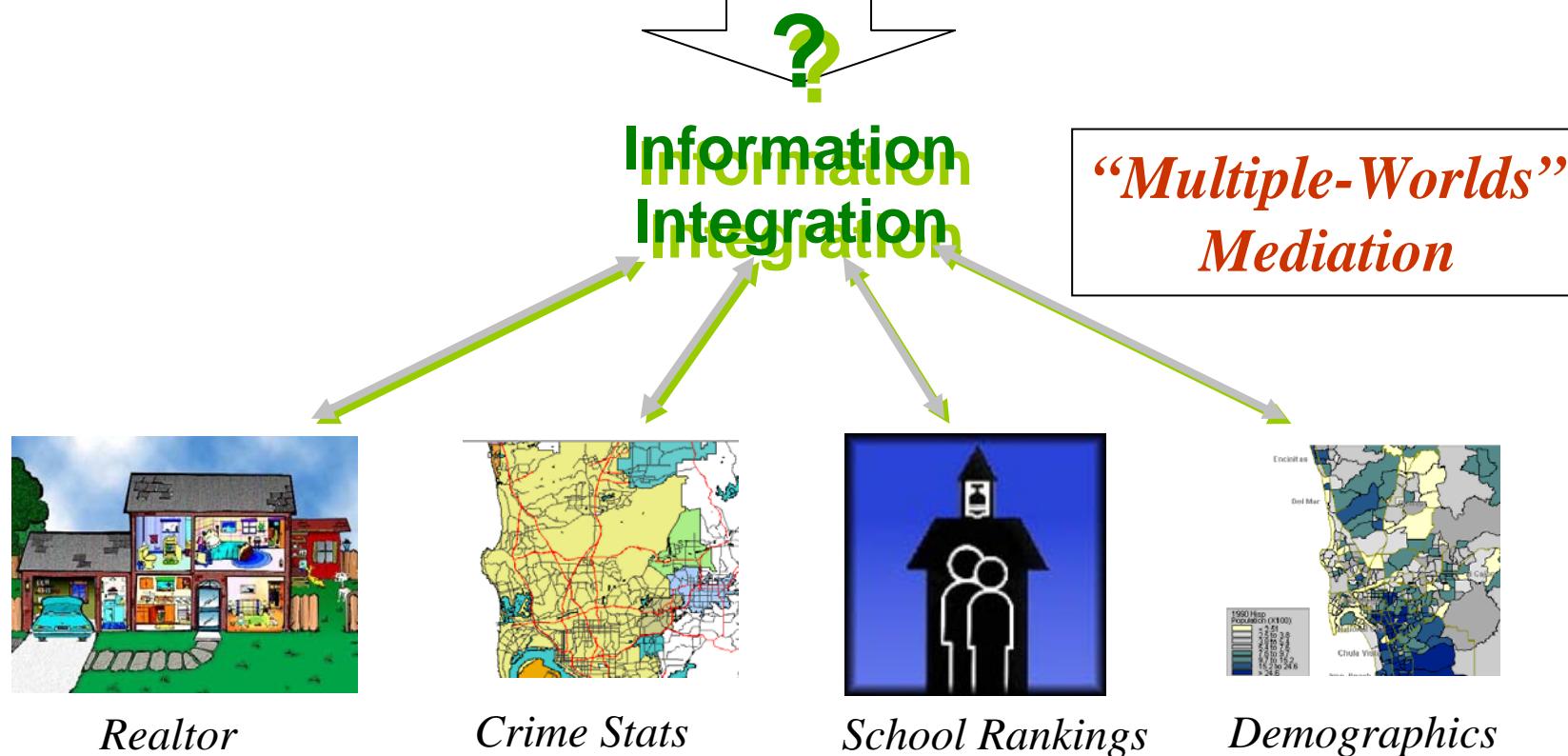
Information about biological processes and entities is totally inter-connected... in theory. In practice, one can find a disappointing landscape of autonomous, heterogeneous and poorly connected database infrastructures [...] Thus, information integration becomes a critical issue[...].

Fuente: BioIT-World Magazine

- Bases de Datos autónomas, cada vez más pobladas.
- Heterogeneidades a varios niveles.
- Database Federation vs. Data Warehousing.
- Integración con otras aplicaciones en los flujos de trabajo (workflows).

Integración de datos: Ejemplo.

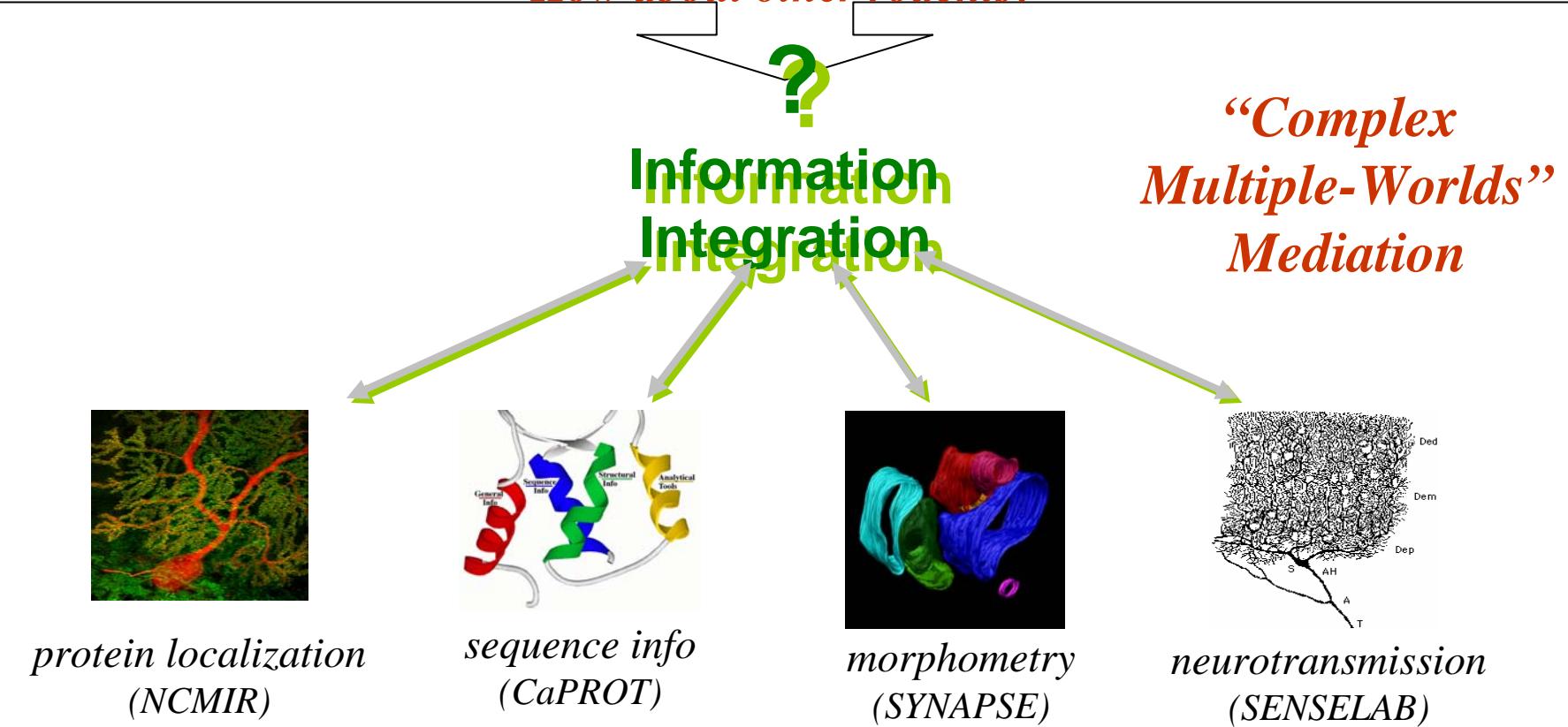
What houses for sale under \$500k have at least 2 bathrooms, 2 bedrooms, a nearby school ranking in the upper third, in a neighborhood with below-average crime rate and diverse population?



Integración de datos: Ejemplo.

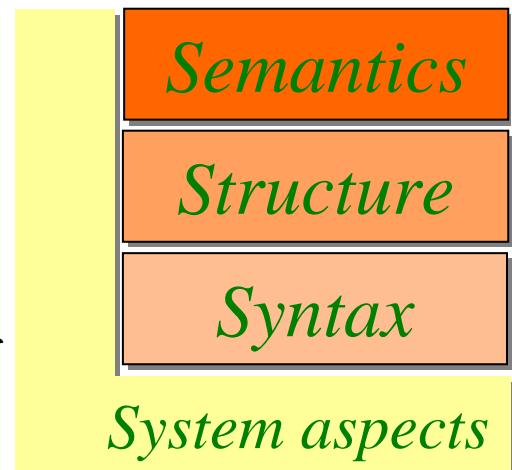
What is the cerebellar distribution of rat proteins with more than 70% homology with human NCS-1? Any structure specificity?

How about other rodents?

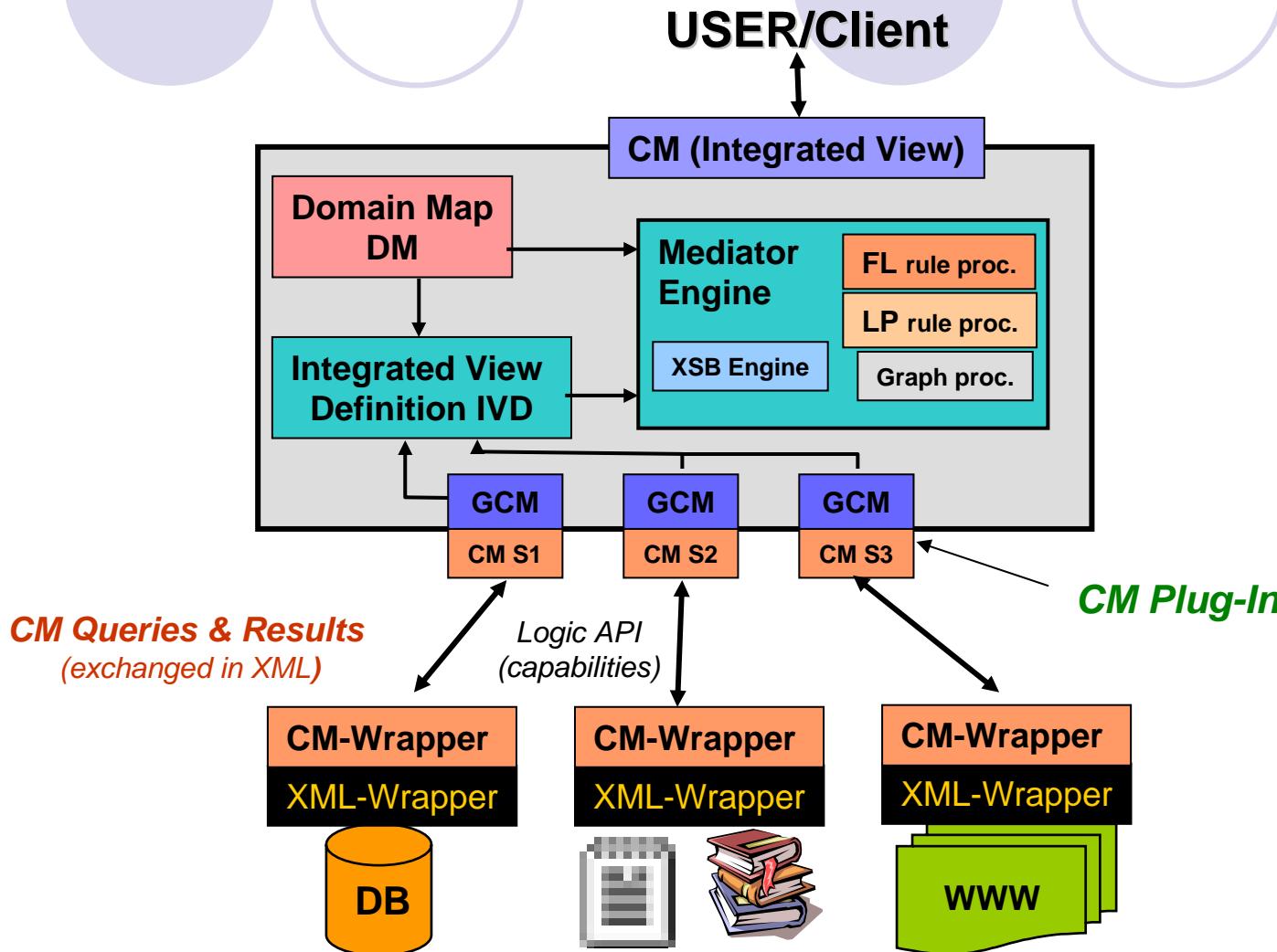


Integración de datos: Obstáculos.

- ... Plataformas (hardware, SOs, aplicaciones), dispositivos, almacenamiento/acceso, localización física...
 - ⇒ ***"Aspectos de Sistema"***
 - ⇒ Tecnologías Grid(s): data-G, computational-G, super-G, ...
- ... Formato de los datos (doc-2-rtf-2-xml, ...)
 - ⇒ ***"Aspectos Sintácticos"***
 - ⇒ *Herramientas de conversión de formatos*
- ... Estructuras de datos (rel-2-oo, rel-2-xml, oo-2-xml, ...)
 - ⇒ ***"Aspectos Estructurales"***
 - ⇒ *lenguajes de consulta y de transformación*
- ... semántica (ontologías, modelos conceptuales, ...)
 - ⇒ ***"Aspectos Semánticos"***
 - ⇒ *representaciones del conocimiento, lenguajes lógicos*

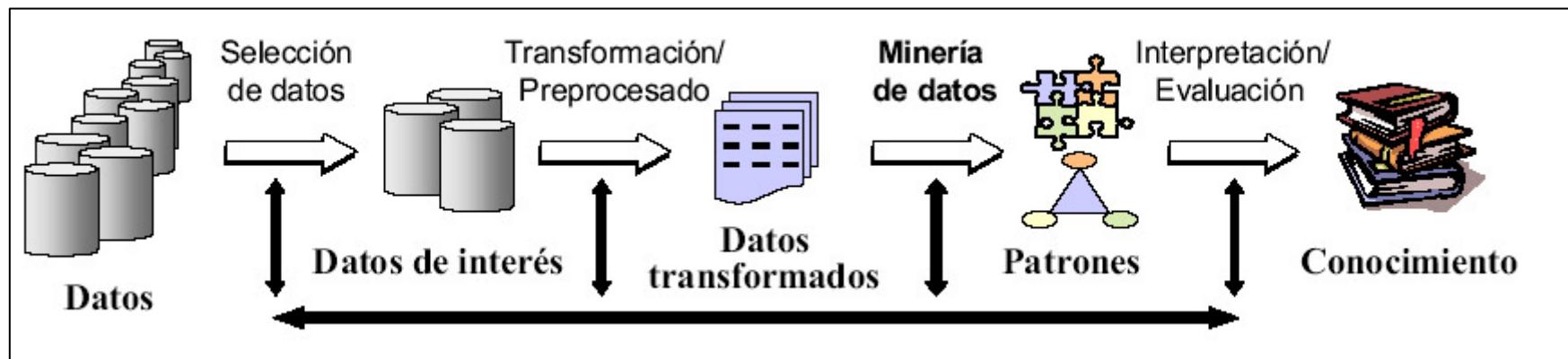


Integración de datos: *Model-based Mediation*.



Minería de datos (*Data Mining*).

Minería de datos: Proceso no trivial de identificar en bases de datos patrones anteriormente desconocidos, válidos, potencialmente útiles y comprensibles por el usuario.



- Patrón o modelo a buscar.
- Criterio de selección.
- Algoritmo de búsqueda.
- Alta dimensionalidad.
- Significación estadística.
- Interacción con el usuario.
- Patrones comprensibles
- Integración.

Minería de datos: Reglas asociativas.

Hipótesis: Aquellos genes con patrones de expresión similares en respuesta a determinadas condiciones experimentales forman parte de grupos de expresión coordinados.



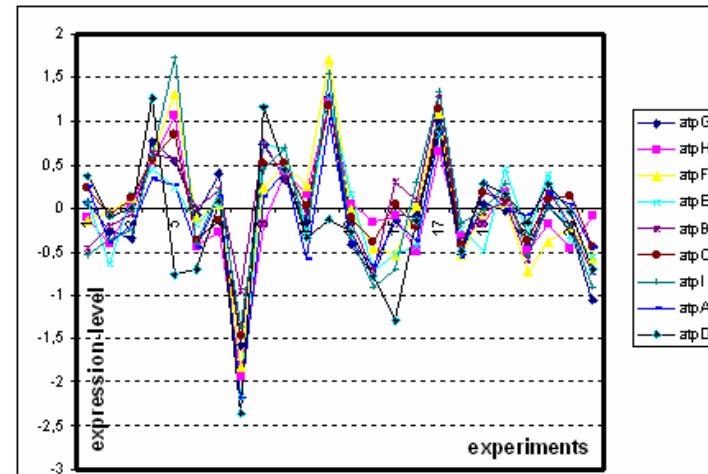
Conf. Soporte

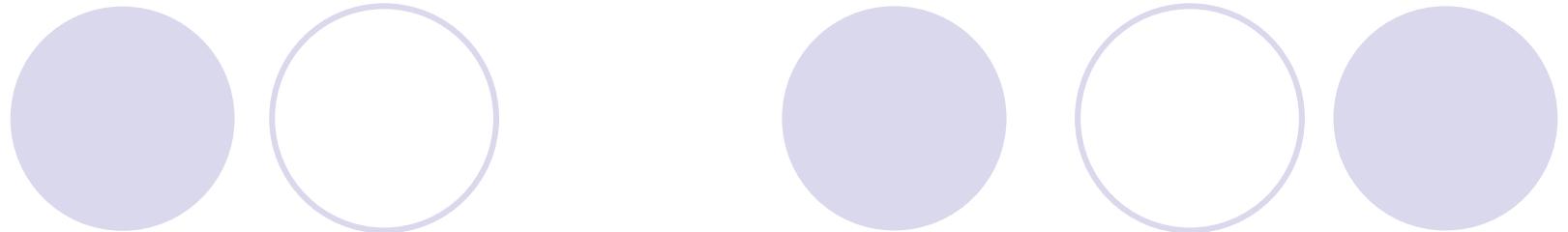
100% , 39% 4 [+]**ex267** [+]**ex279** [-]**ex270** ⇒ ATP SYNTASE*

100% , 49% 5 [+]**ex274** [+]**ex279** [-]**ex270** [-]**ex276** ⇒ ATP SYNTASE*

Regla

Clúste	genes	/product=
1	atpG	ATP synthase (subunit gamma)
1	atpH	ATP synthase (subunit delta)
1	atpF	ATP synthase (subunit b)
1	atpE	ATP synthase (subunit c)
1	atpB	ATP synthase (subunit a)
2	atpC	ATP synthase (subunit epsilon)
2	atpI	ATP synthase (subunit i)
3	atpA	ATP synthase (subunit alpha)
4	atpD	ATP synthase (subunit beta)



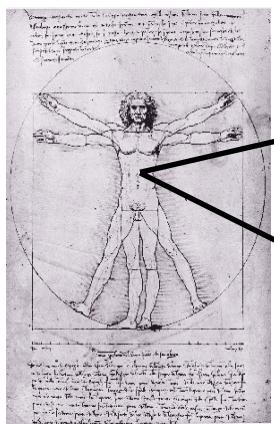


PARTE II. Análisis de expresión génica con microarrays de ADN

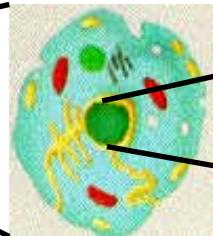
Biología y Genómica Funcional

- Poseer información completa de todos los genes es muy importante, pero también lo es el estudio de la interacción entre estos genes en el organismo.
- Últimamente los esfuerzos se han concentrado en estudiar las interrelaciones de todos los genes simultáneamente.
- Para esto se requiere una tecnología de medición masiva de la expresión génica.

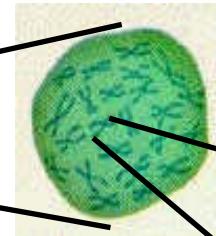
Biología y Genómica Funcional (II)



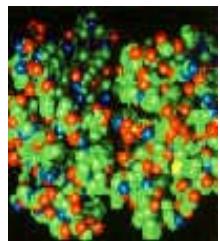
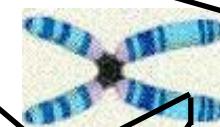
Cell



Nucleus

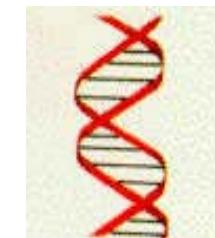
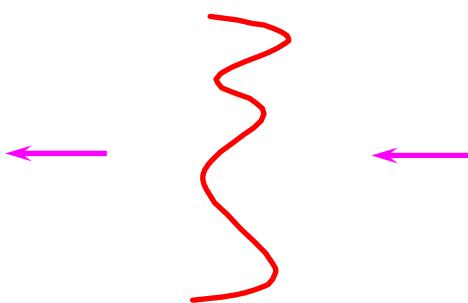


Chromosome



Protein

Gene (mRNA),
single strand

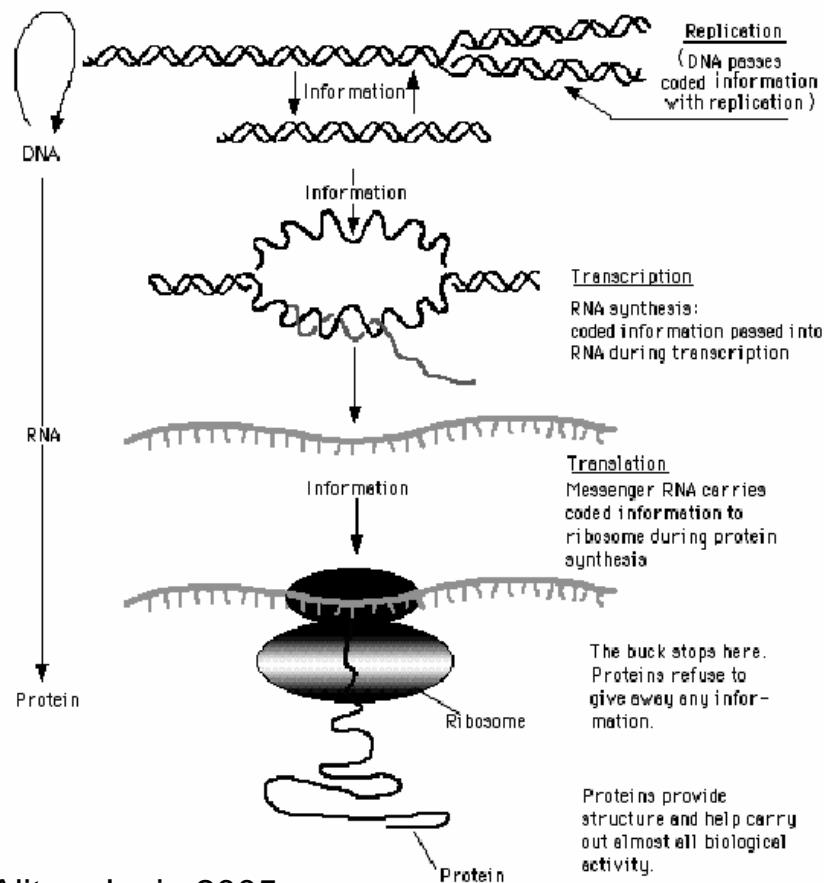


Gene (DNA)

Graphics courtesy of the National Human Genome Research Institute

El dogma central de la Biología Molecular

The Central Dogma of Molecular Biology

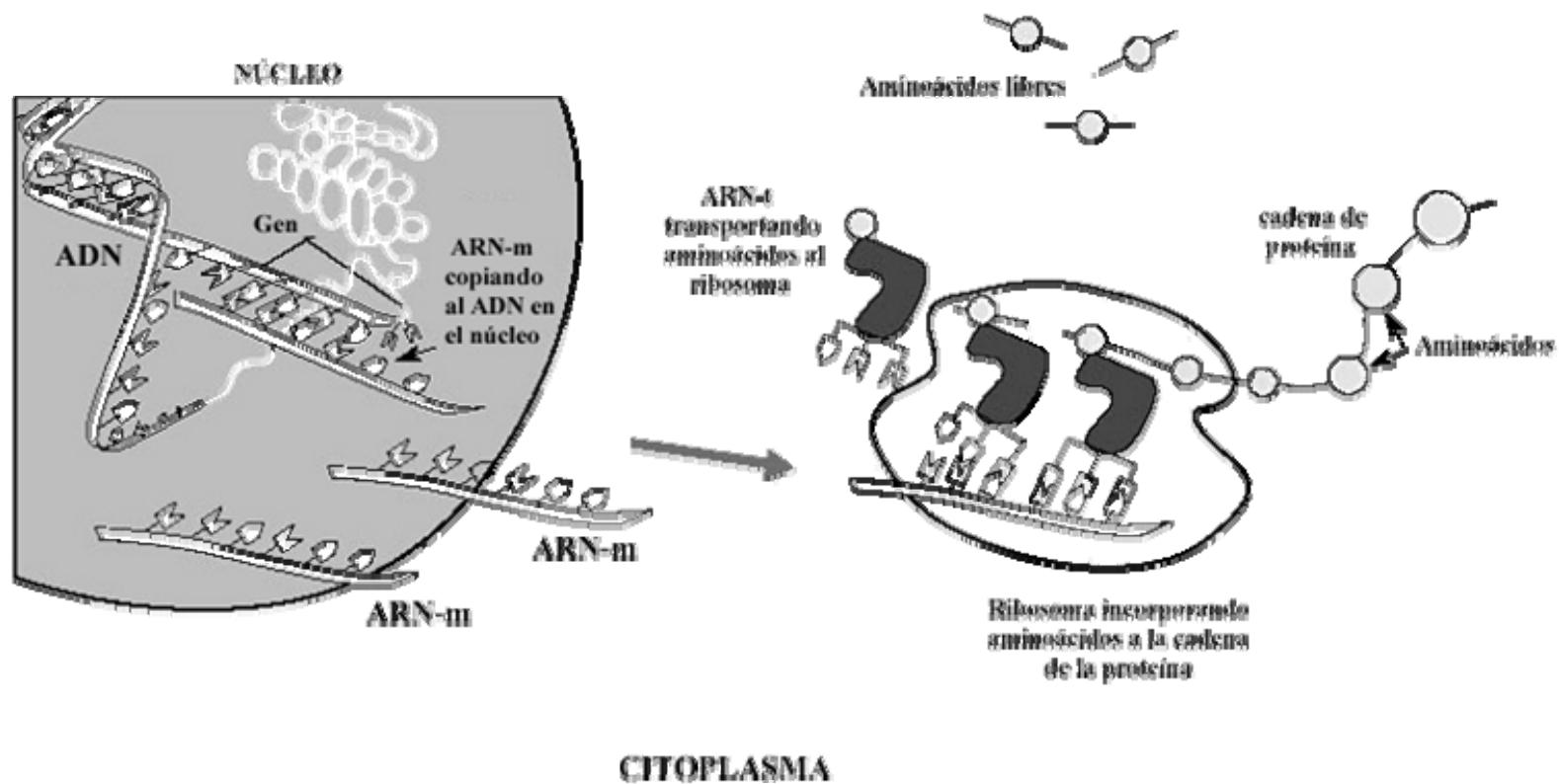


Aliter, Junio 2005.

Transcripción del DNA al RNA a la proteína:

1. El DNA replica su información en un proceso llamado **replicación**
2. El DNA codifica para la producción del RNA mensajero (mRNA) durante la **transcripción**.
3. El RNA mensajero lleva información codificada a los ribosomas. El ribosoma "lee" esta información y la utiliza para la síntesis de proteínas. A este proceso se le llama **traducción**.

Expresión génica



¿Por qué el estudio de la expresión génica?

- El **patrón de genes expresados** en una célula brinda información del estado actual de la misma.
- Existe una **correlación** entre el estado de la célula y los cambios en los niveles de mRNA de muchos genes.
- Los **patrones de expresión** de genes que nunca han sido caracterizados pueden proporcionar nuevas pistas sobre su posible función.

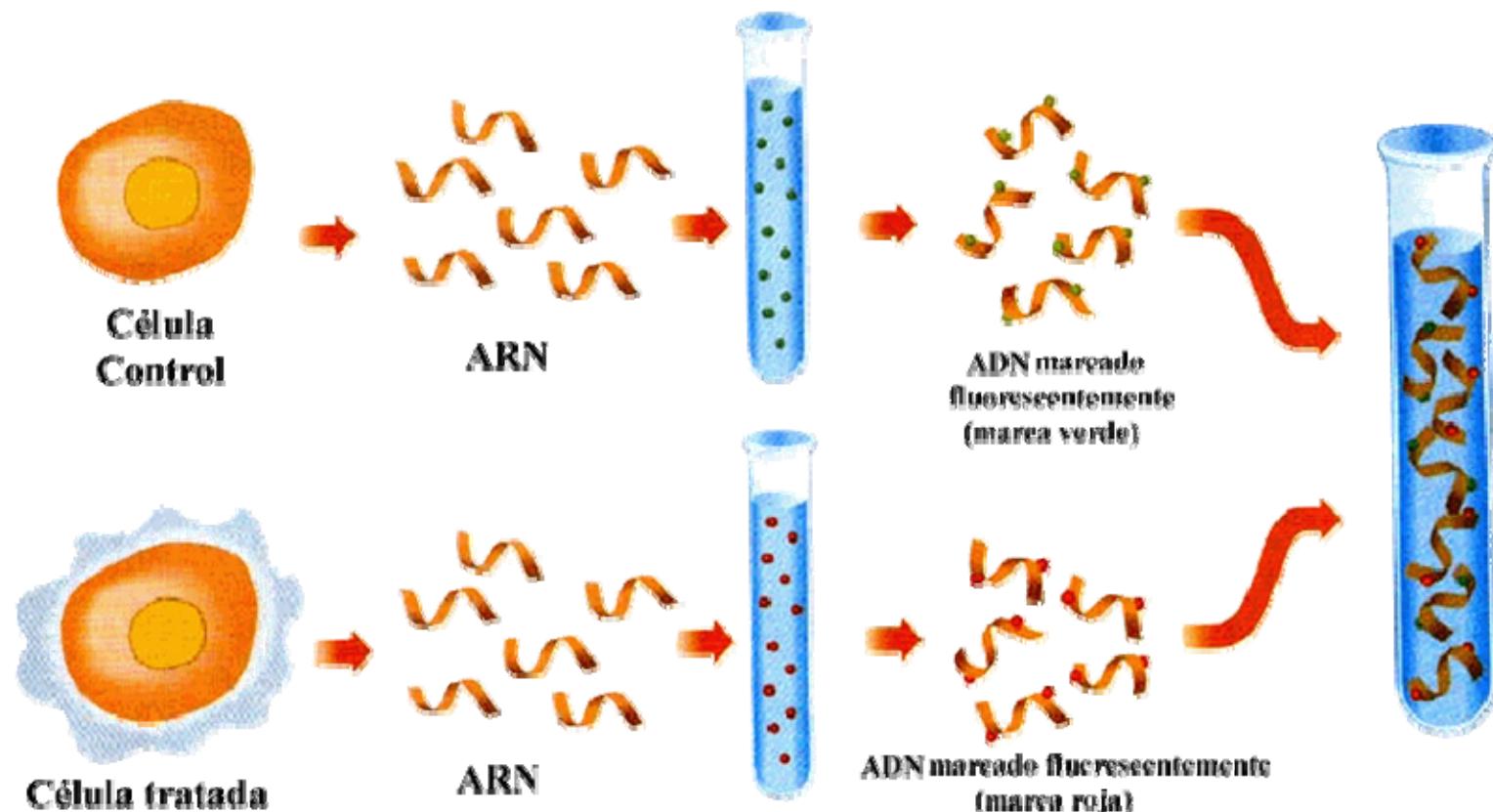
La tecnología de los DNA microarray

- Es una tecnología concebida para detectar la **expresión de miles de genes simultáneamente**.
- Presenta multitud de aplicaciones potenciales:
 - Identificación de enfermedades genéticas complejas.
 - Estudio toxicológicos
 - Drug discovery
 - Múltiples estudios de expresión de los genes a través del tiempo (distintos tejidos, distintos estadíos de enfermedades...)

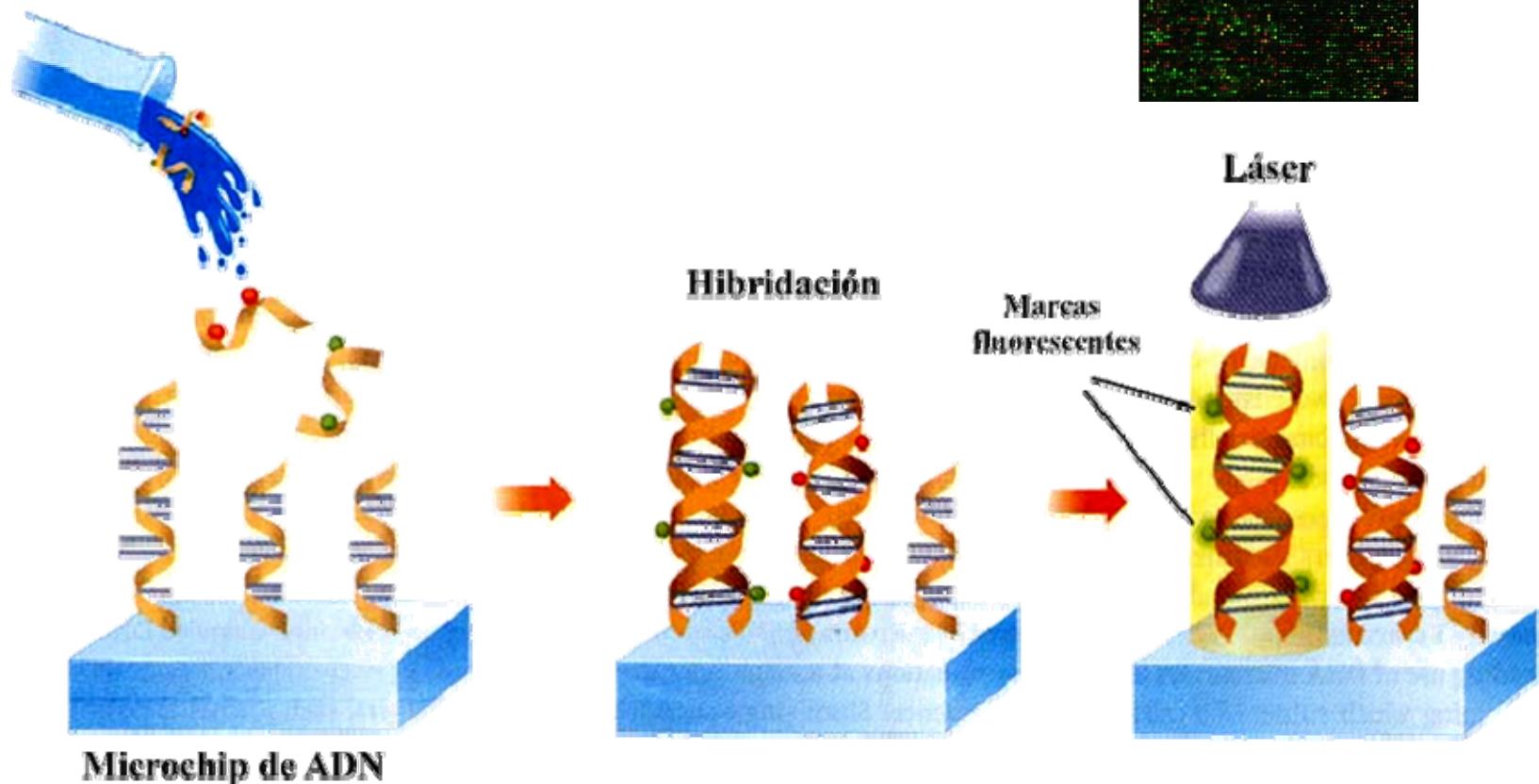
Principios de Microarrays de ADN

- El ADN complementario a los genes de interés es generado y depositado en posiciones específicas en placas de vidrio.
- El ADN de las muestras a analizar es marcado con sustancias fluorescentes y vertido sobre la superficie de estas placas. Como el el DNA complementario tiende a unirse (hibridación), aquellos genes que se han expresado en la célula se fijarán a su copia en la placa.
- La presencia del DNA expresado se detecta por fluorescencia al excitarse con láser.

Proceso (I)

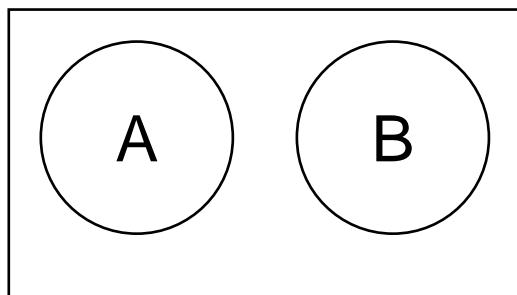


Proceso (II)



Proceso (III)

On the surface

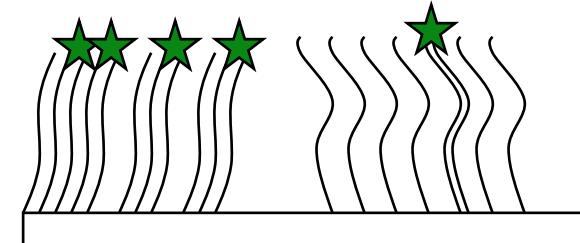
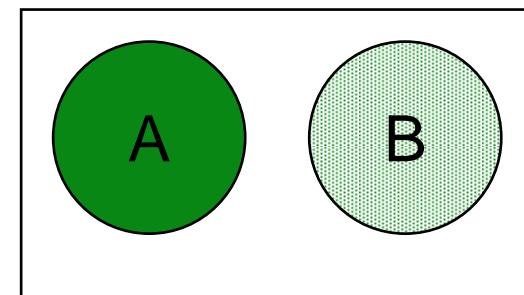


In solution

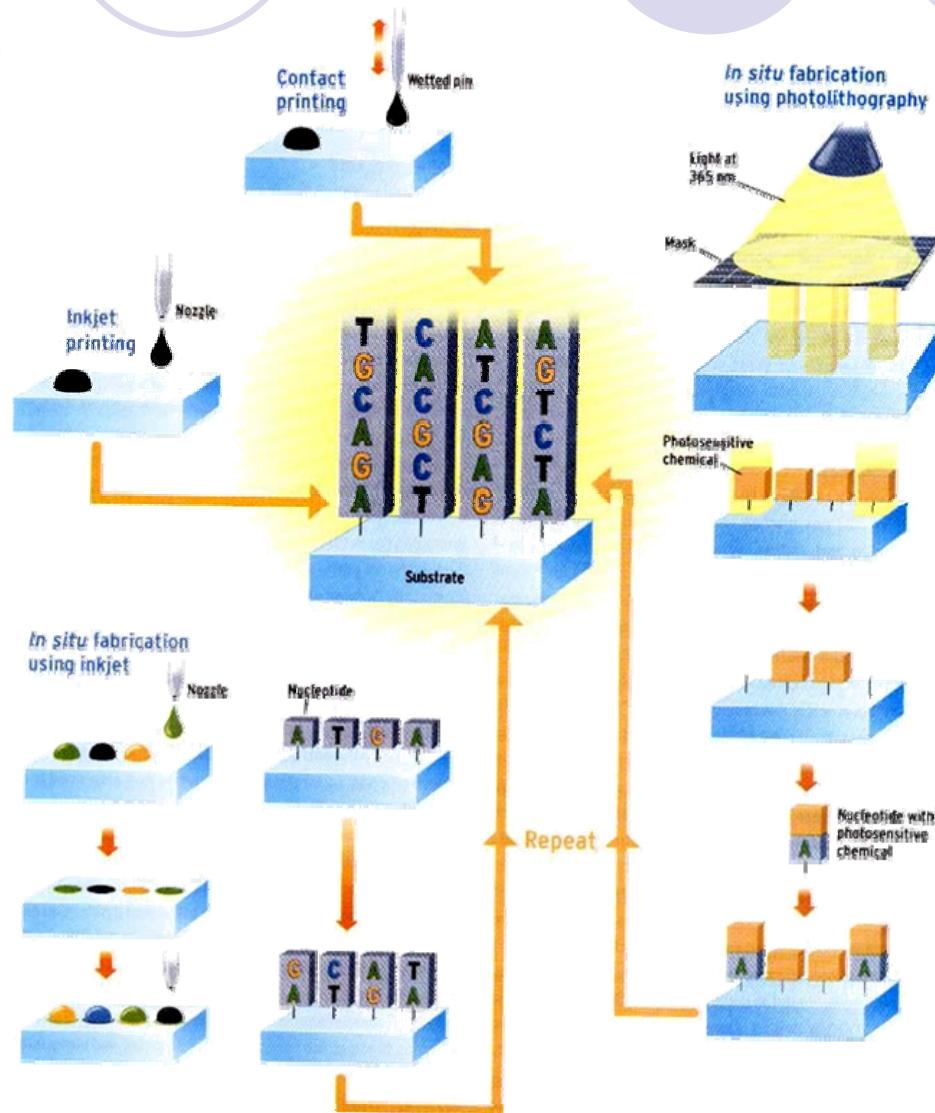
4 copies of gene A,
1 copy of gene B



After Hybridization



¿Cómo se hace el microarray?



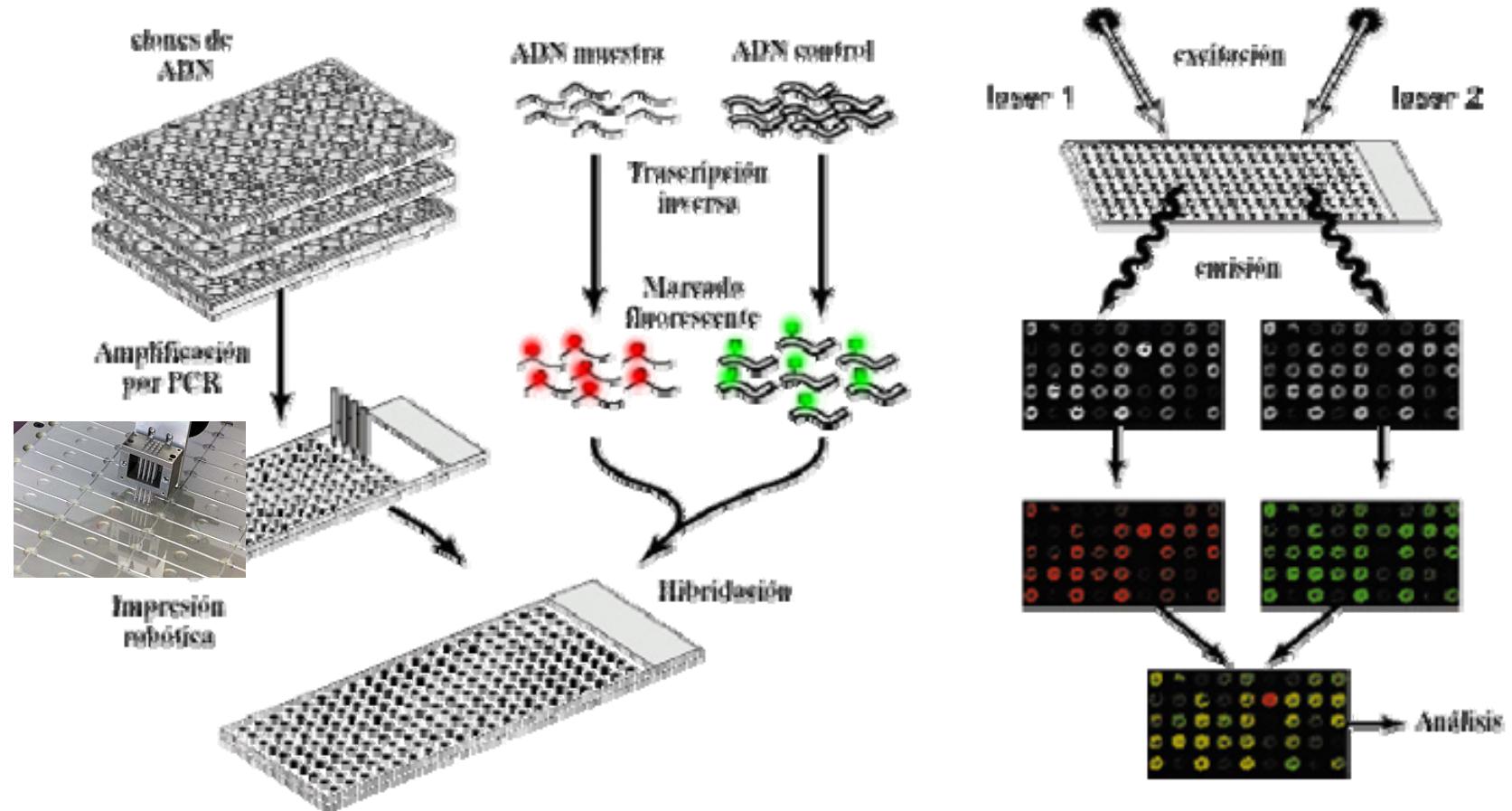
Tecnologías de Microarrays:

- Short oligonucleotide arrays (**Affymetrix**)
- cDNA or spotted arrays (**Brown/Botstein**).
- Long oligonucleotide arrays (**Agilent Inkjet**)
- Fiber-optic arrays

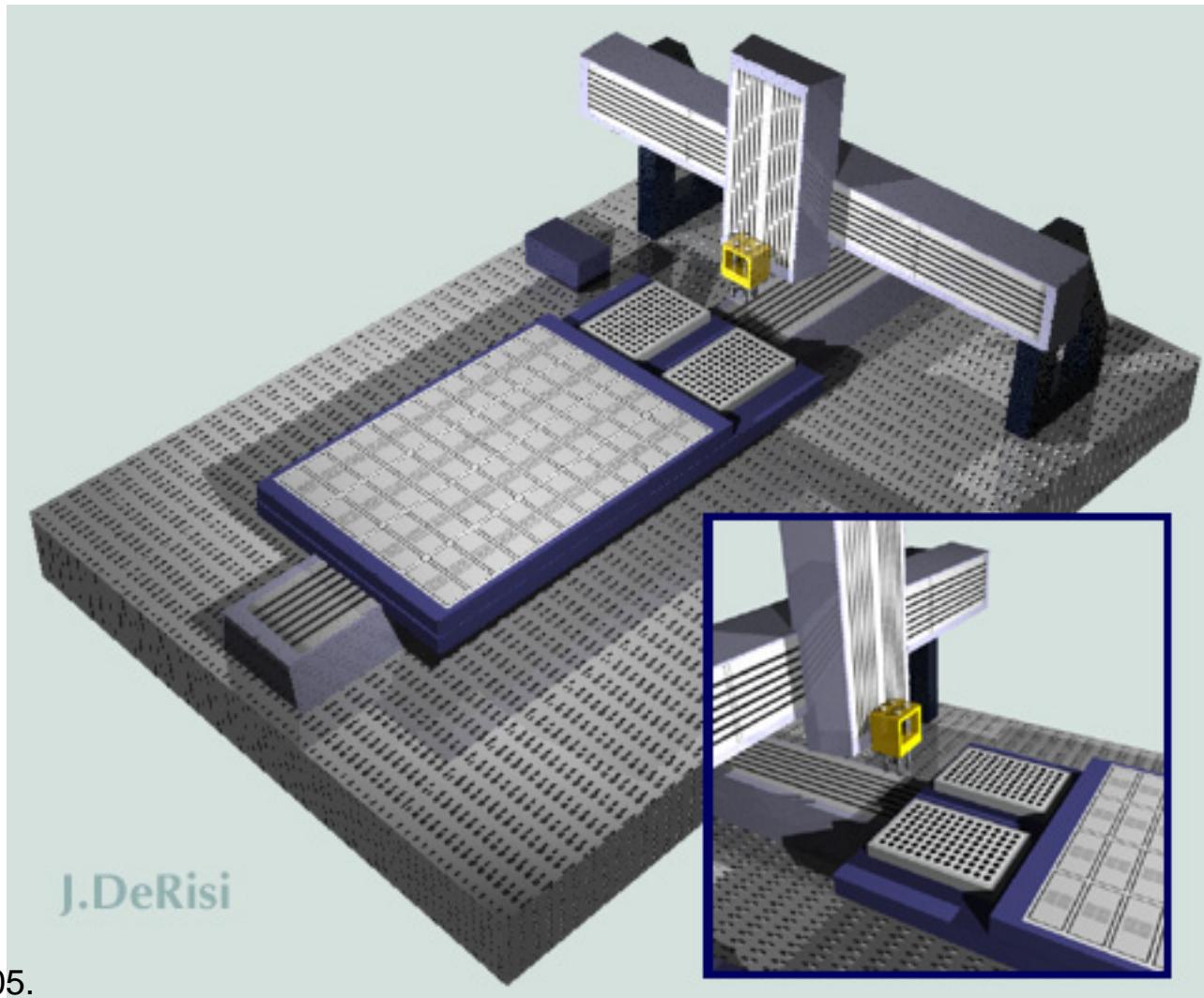
Las tecnologías difieren fundamentalmente en:

- La forma en que el **DNA es depositado en el sustrato** (spotting, lithography, Inkjet printing,...).
- **Longitud de la secuencia** del DNA que es depositada (secuencia completa o fragmentos del gen).
- El tipo de **señal que se mide de cada spot** (e.g. fluorescencia)

La tecnología del cDNA microarray

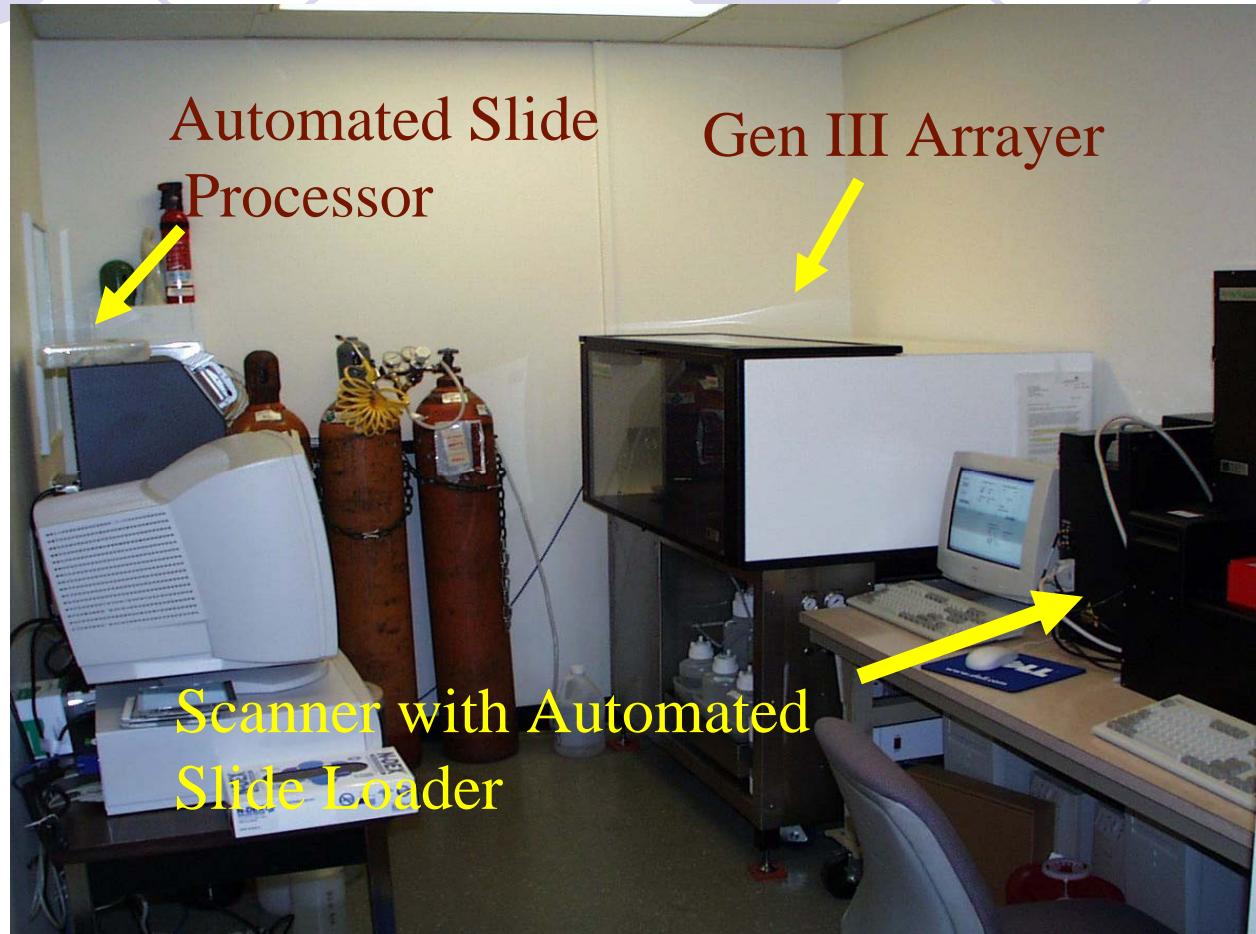


Arrayer (Robot):



Aliter, Junio 2005.

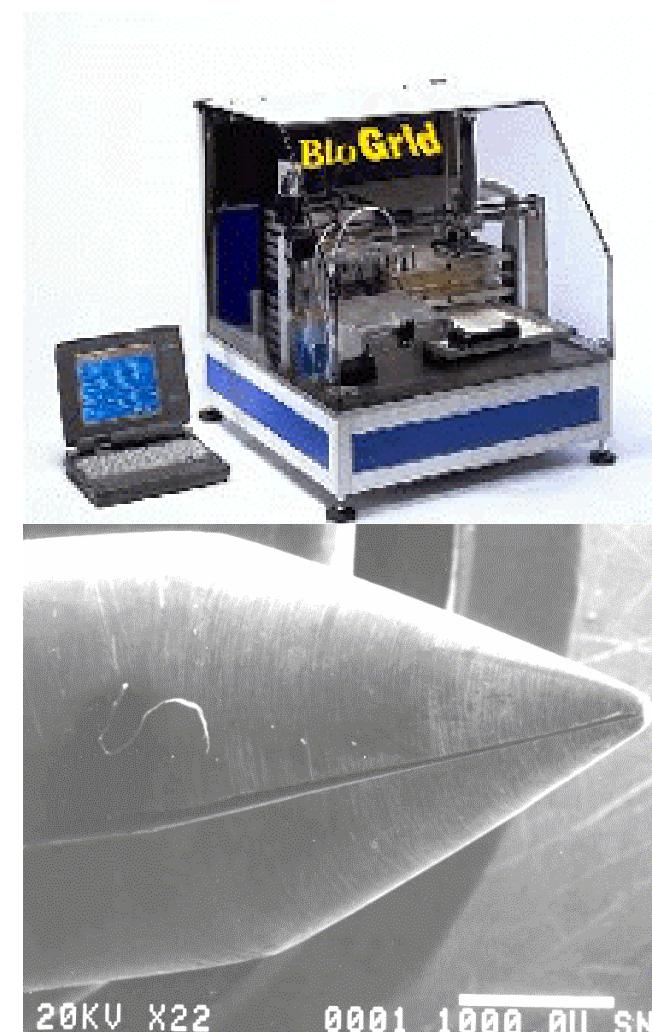
Laboratorio de Microarrays



Microarray Gridder

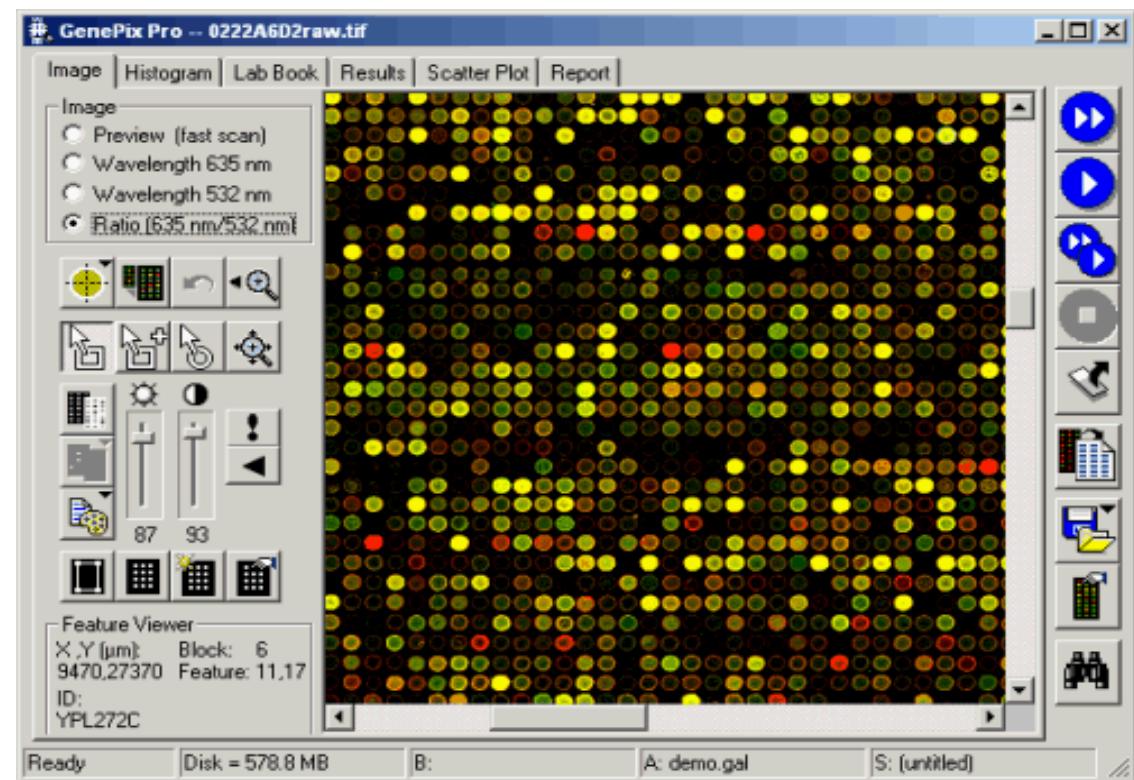


Aliter, Junio 2005.

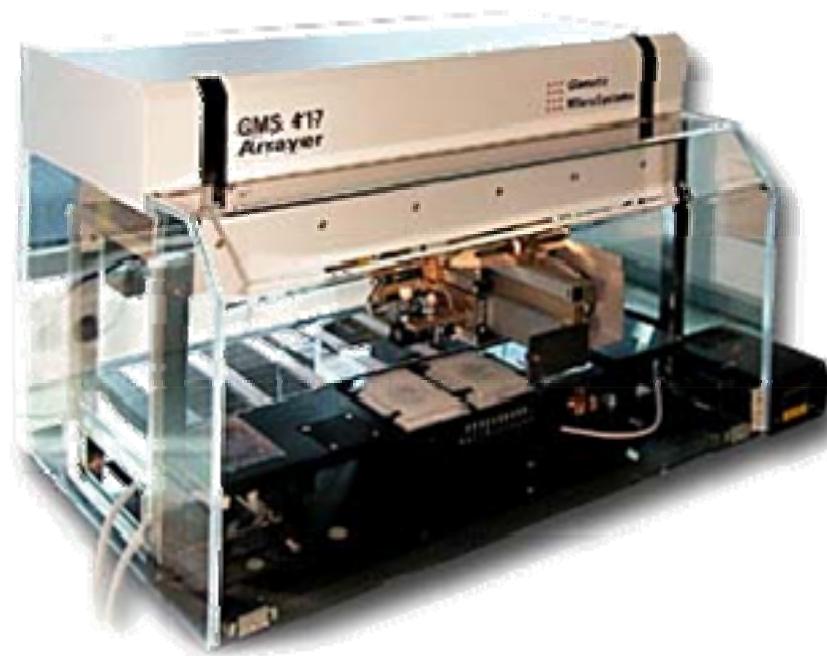


20KV X22 0001 1000.00 nm

Scanner

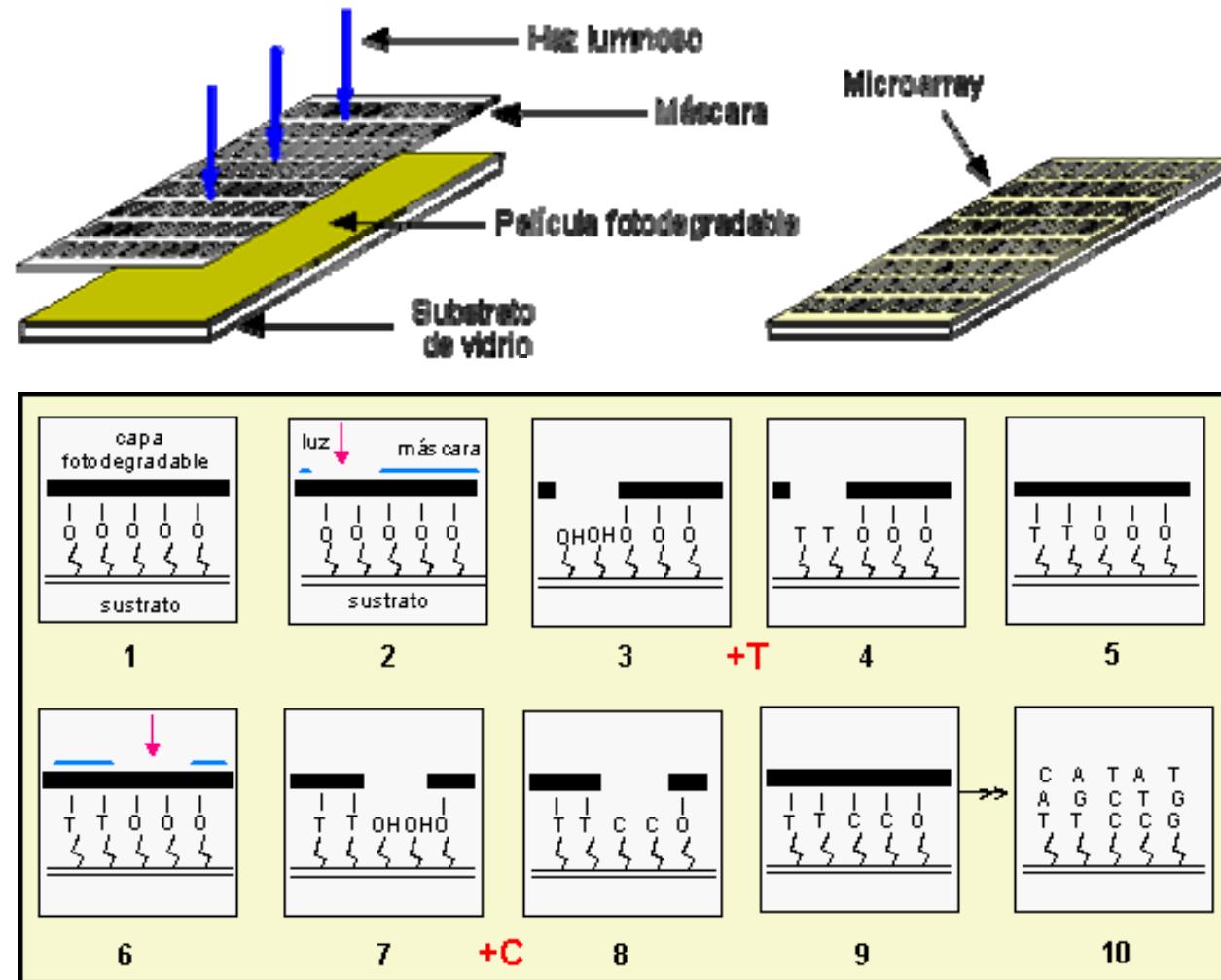


Tecnología Affymetrix

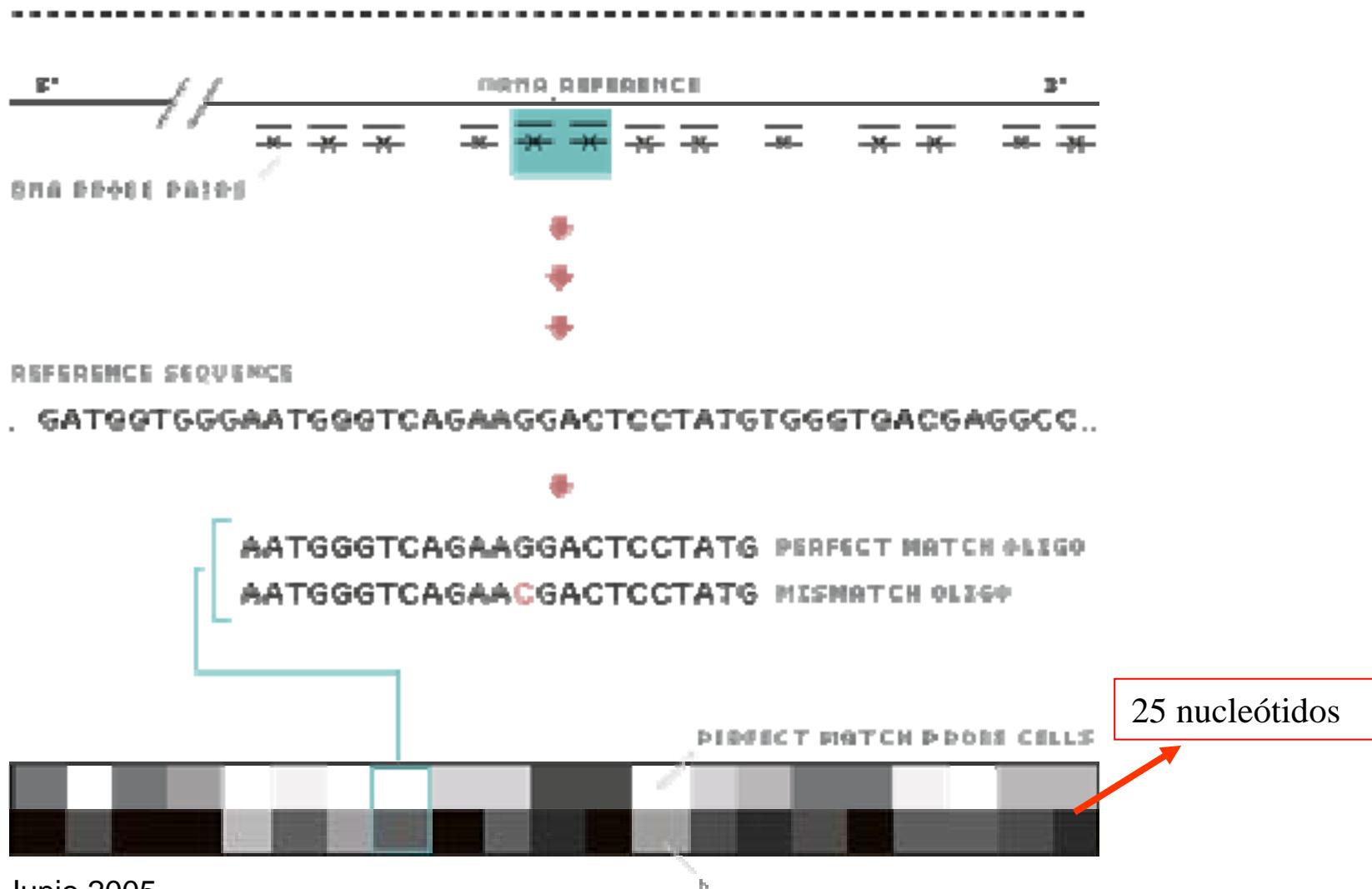


Aliter, Junio 2005.

Light directed oligonucleotide synthesis:



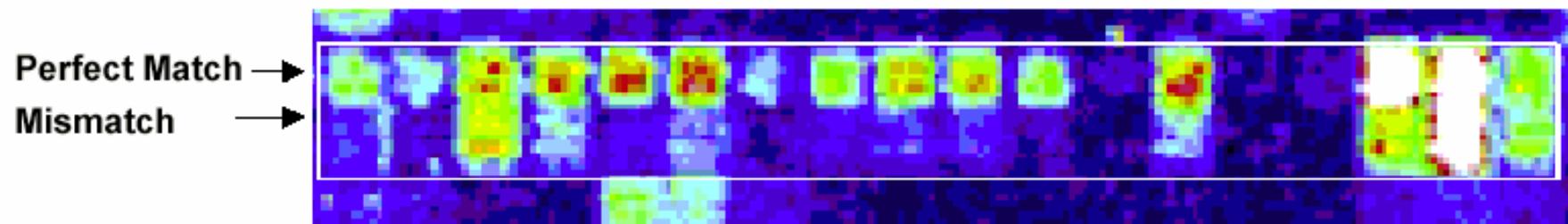
Multiples “probe pairs” por gen:



Características del proceso de Affymetrix

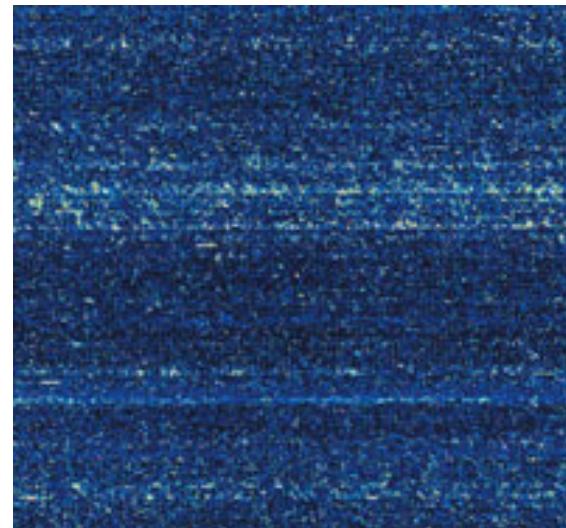
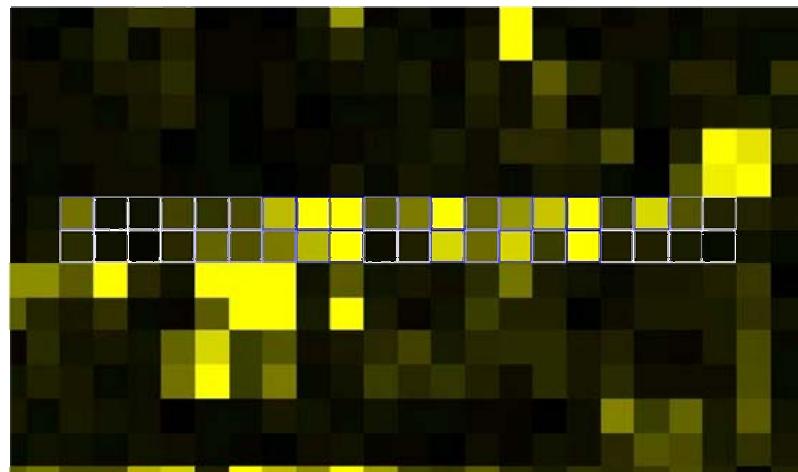
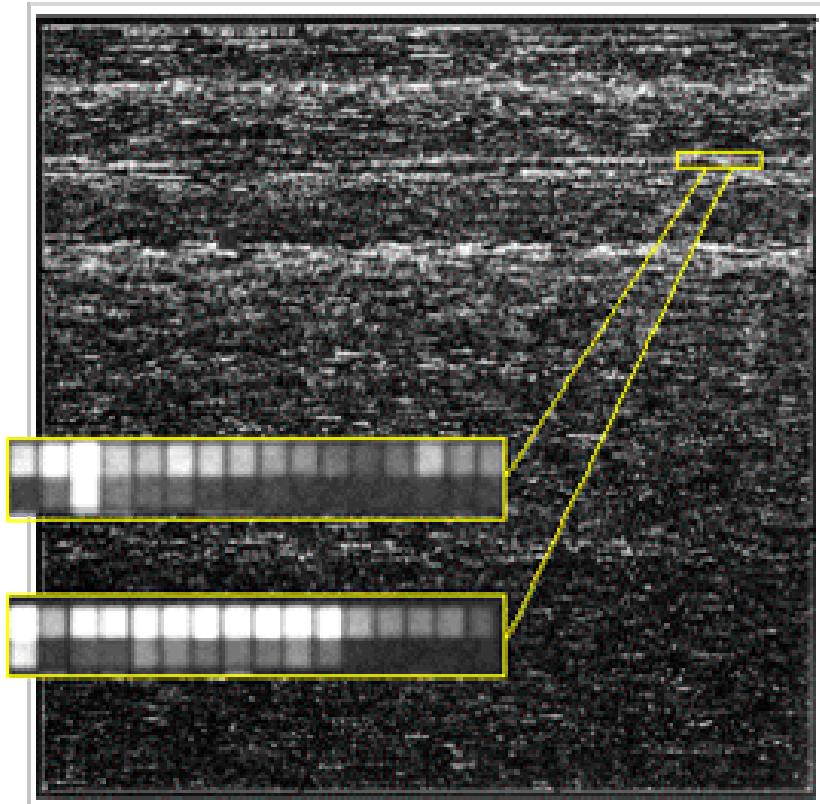
- Cada “probe pair” contiene millones de copias de una secuencia de oligonucleotidos.
- Existe una reducción significativa del número de errores que ocurren por hibridación cruzada.
- Es un método mas “cuantitativo”.
- Coste muy alto

A Probe Set (DNA Chip)



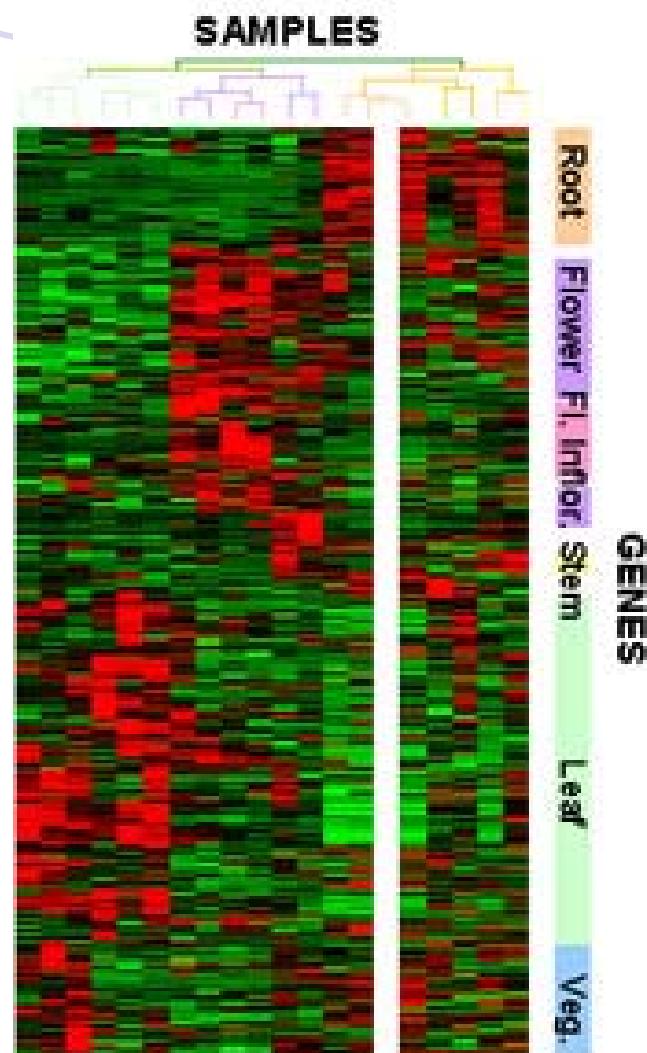
Perfect Match AGGCTATCGCACTCCAGTGG
 AGGCTATCGTACTCCAGTGG
 |

Imagen producida



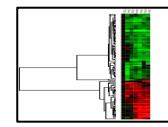
Aliter, Junio 2005.

Procesamiento de Datos



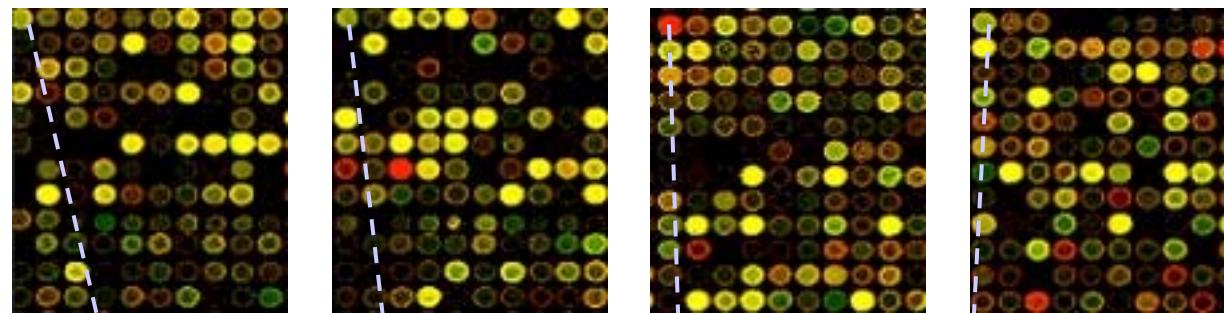
Aliter, Junio 2005.

Flujo de procesamiento

- Adquisición de datos 
- Almacenamiento 
- Preprocesamiento (Normalización, Duplicados, etc)
- Filtrado
- **Análisis (*agrupamiento, clasificación, predicción, etc*)** 
- Visualización 
- Interpretación/Anotación
- Publicación en repositorio público 



Generación de los datos Patrón de expresión

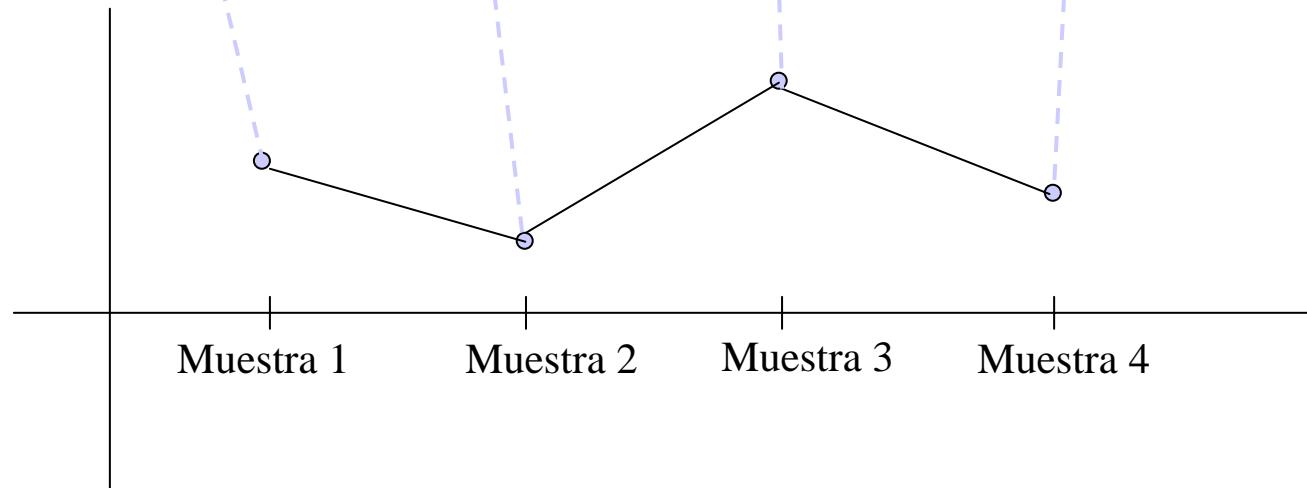


Muestra 1

Muestra 2

Muestra 3

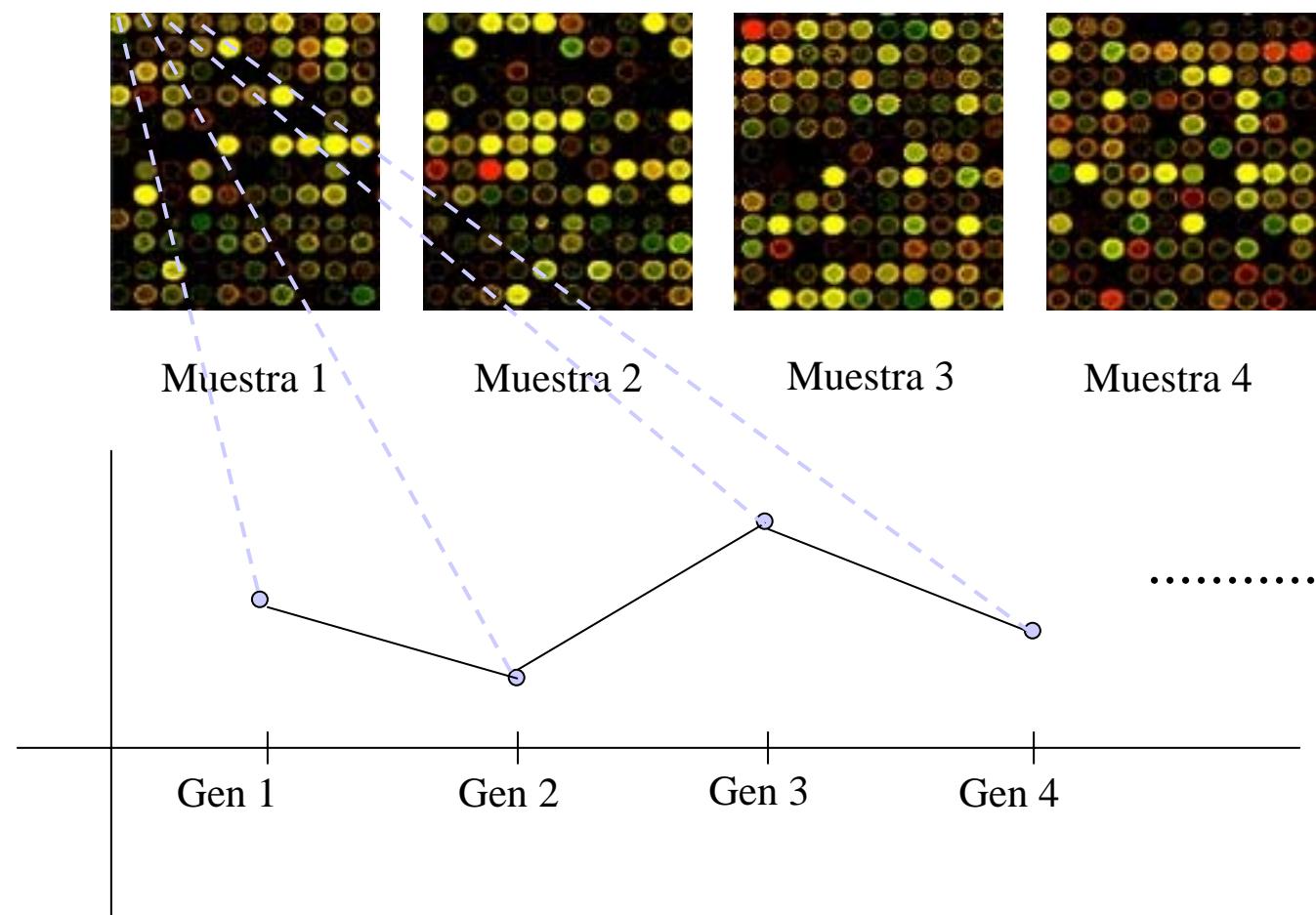
Muestra 4



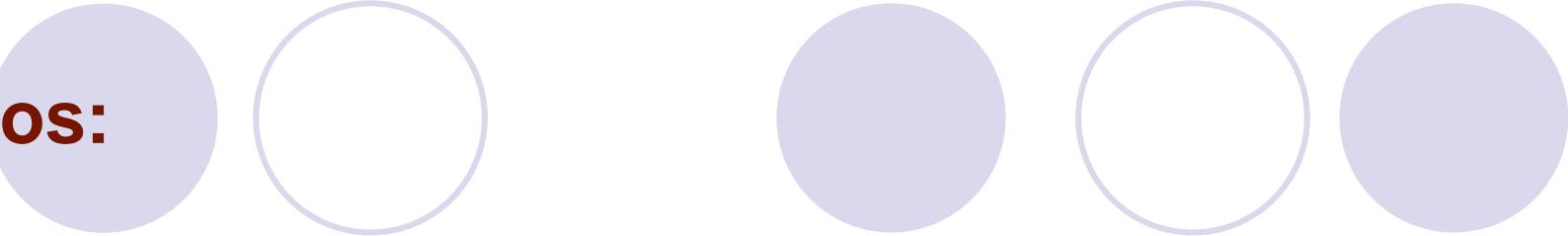
Para cada gen tenemos un perfil de expresión formado por los n experimentos.

Aliter, Junio 2005.

Generación de los datos: Por genes (Análisis fenotípico)



Para cada experimento (muestra) tenemos un perfil de expresión (huella molecular)
Aliter, Junio 2005. formado por p genes.



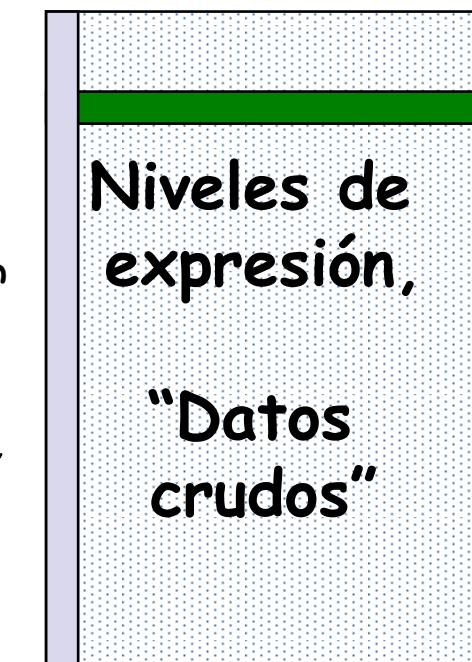
Datos:

- **Análisis de expresión:**
 n vectores de p variables
 n : número de genes (puntos en el chip)
 p : número de muestras (número de chips)
- **Análisis fenotípico:**
 n vectores de p variables
 n : número de muestras (número de chips)
 p : número de genes (puntos en el chip)

Matriz de datos:

Elementos de la matriz de datos:

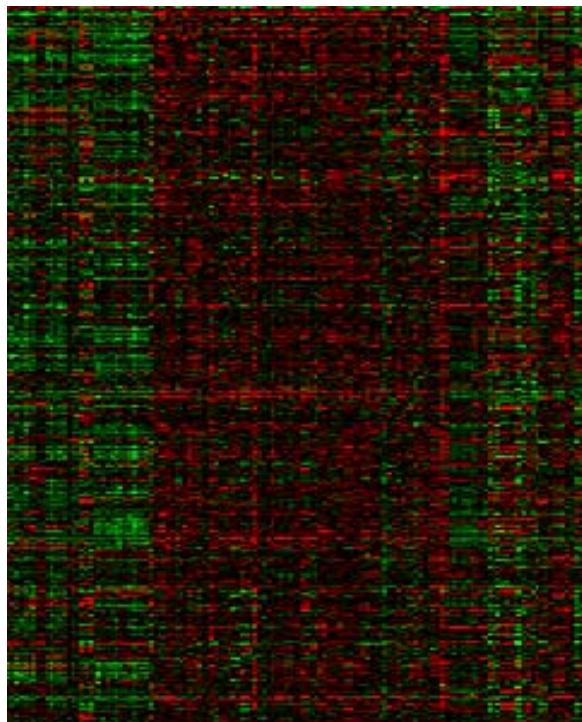
- valores relativos de expresión (ratios) condiciones →
- valores absolutos
- Distribuciones...
- Fila = patrón de expresión/
vector huella de un gen
- Columna = perfil condición /
tejido/ chip



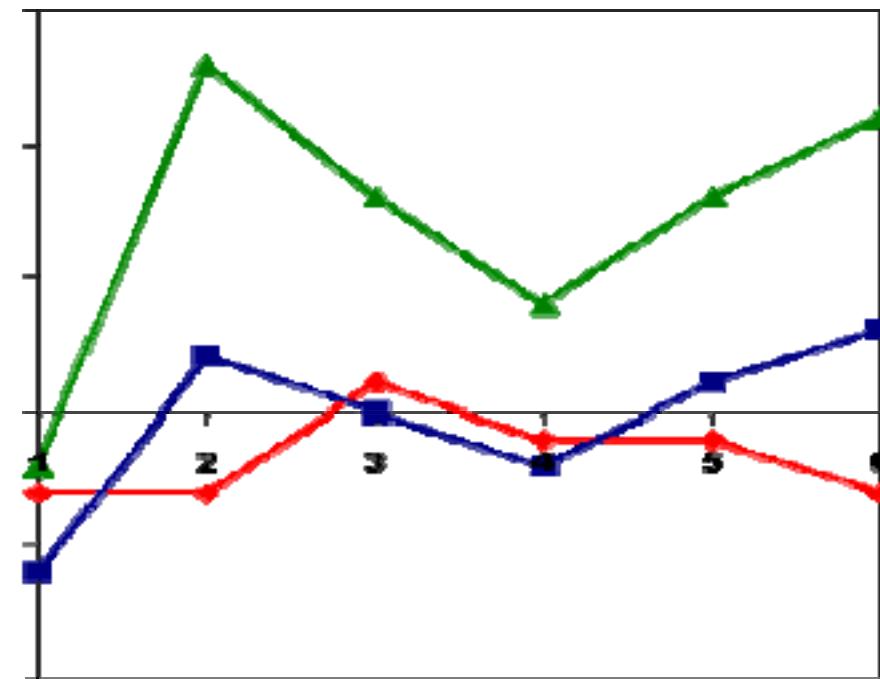
Matriz de datos:

	Gene Attribute	Exp 1	Exp 2	Exp 3	Exp4	Exp5
Exp Attributes		type I	type II	type III	type II	type III
Gene 1	foo	0.51	0.70	0.88	0.21	0.83
Gene 2	bar	0.35	0.87	0.96	0.22	0.97
Gene 3	blee	0.20	0.06	0.72	0.50	0.99
Gene 4	bas	0.06	0.17	0.37	0.16	0.42
Gene 5	groo	0.54	0.70	0.41	0.86	0.50
Gene 6	gar	0.57	0.28	0.58	0.61	0.58
Gene 7	glee	0.57	0.20	0.45	0.11	0.51
Gene 8	glas	0.52	0.68	0.21	0.43	0.08
Gene 9	gree	0.35	0.91	0.25	0.72	0.67
Gene 10	goe	0.68	0.35	0.25	0.53	0.18

Visualización:



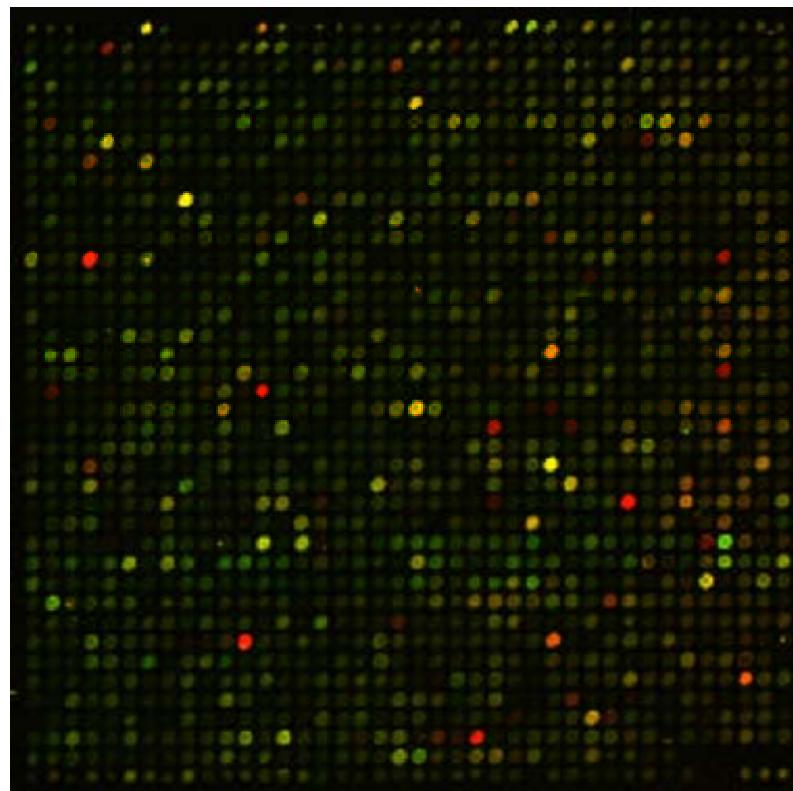
Filas: genes
Columnas: muestras
Color: Nivel de expresión



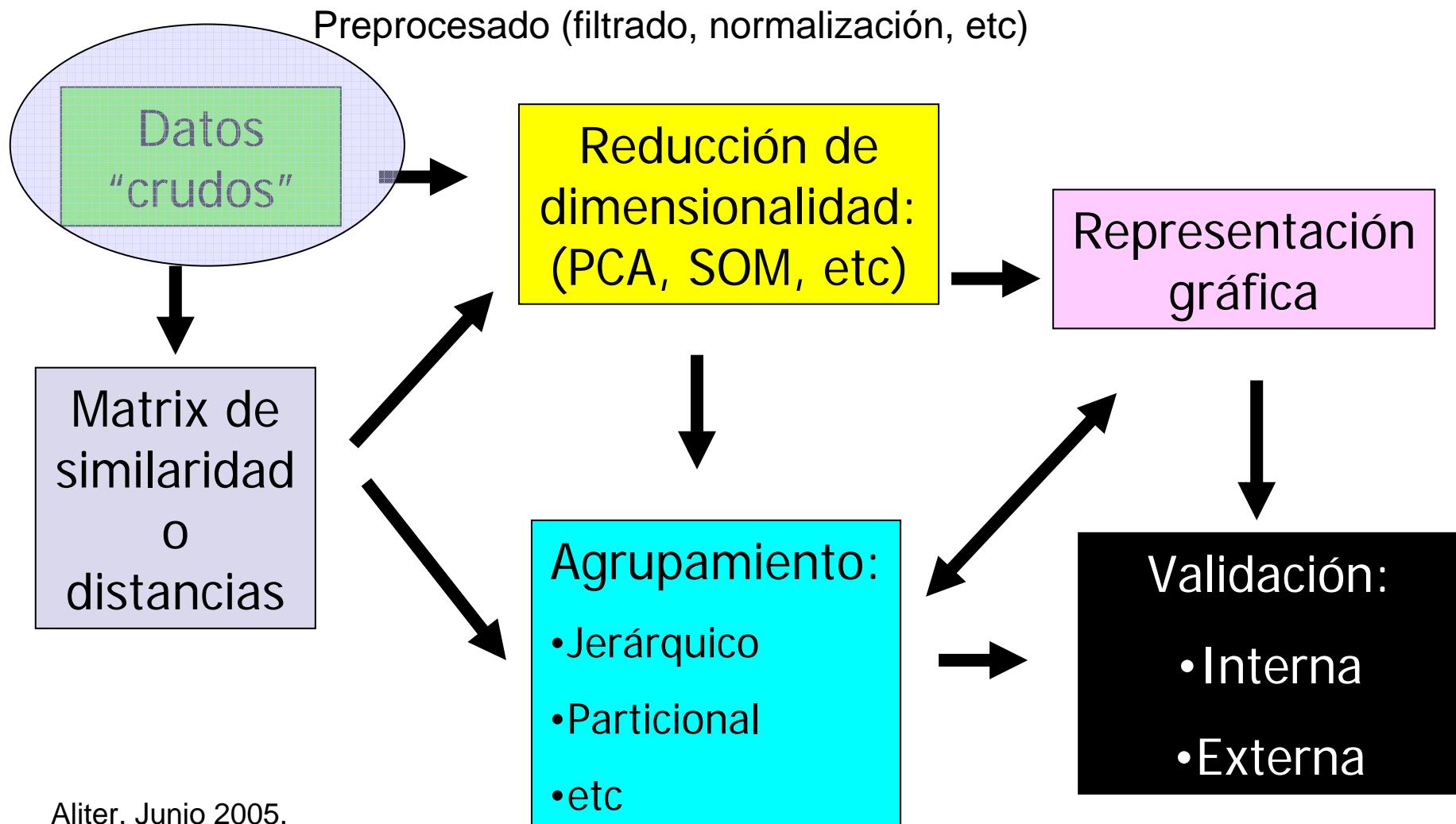
Eje X: muestras
Eje Y: nivel de expresión
Cada gráfica: Un gen

Interpretación de la imagen:

- = more abundant in cell type A
- = more abundant in cell type B
- = equally abundant in both cell types

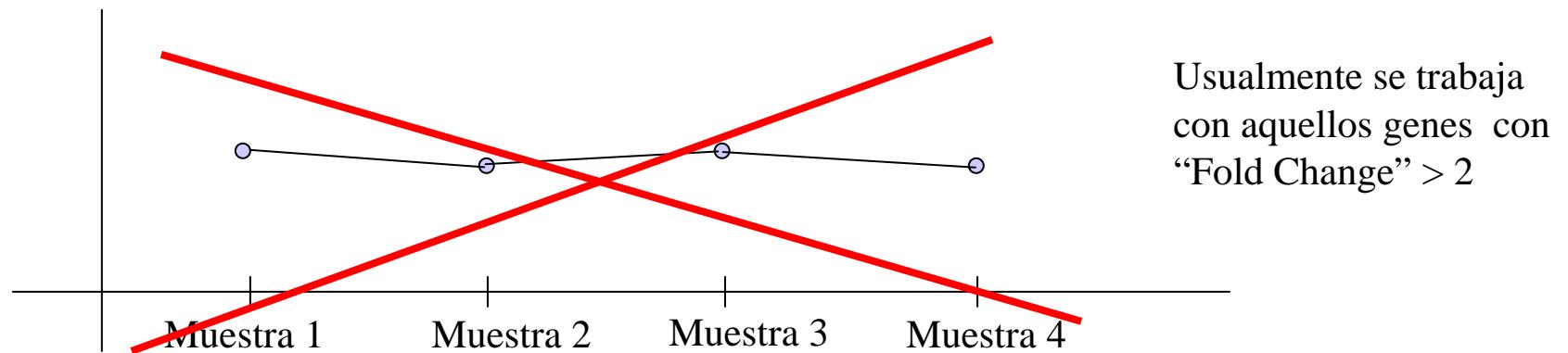


Aprendizaje no supervisado

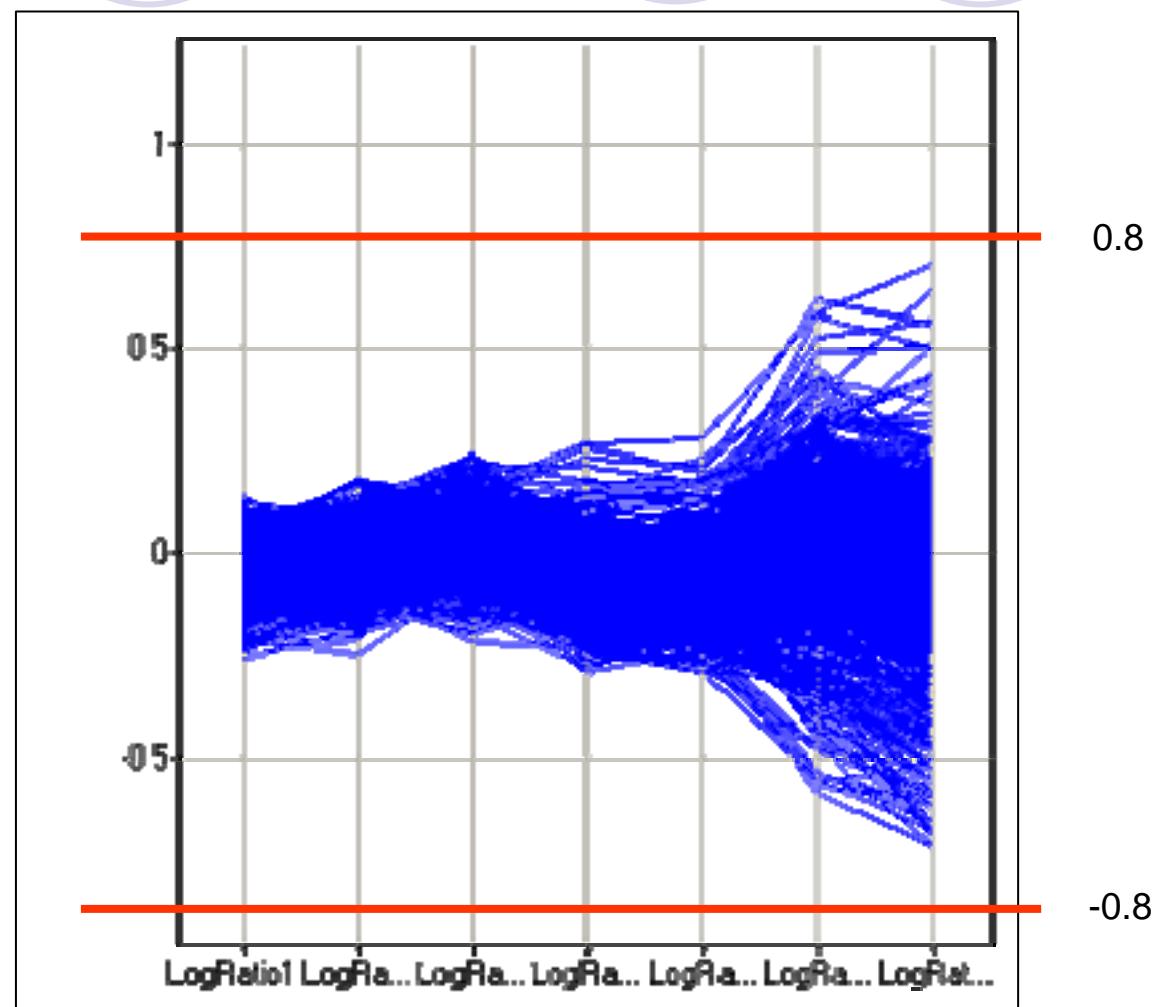


Filtrado:

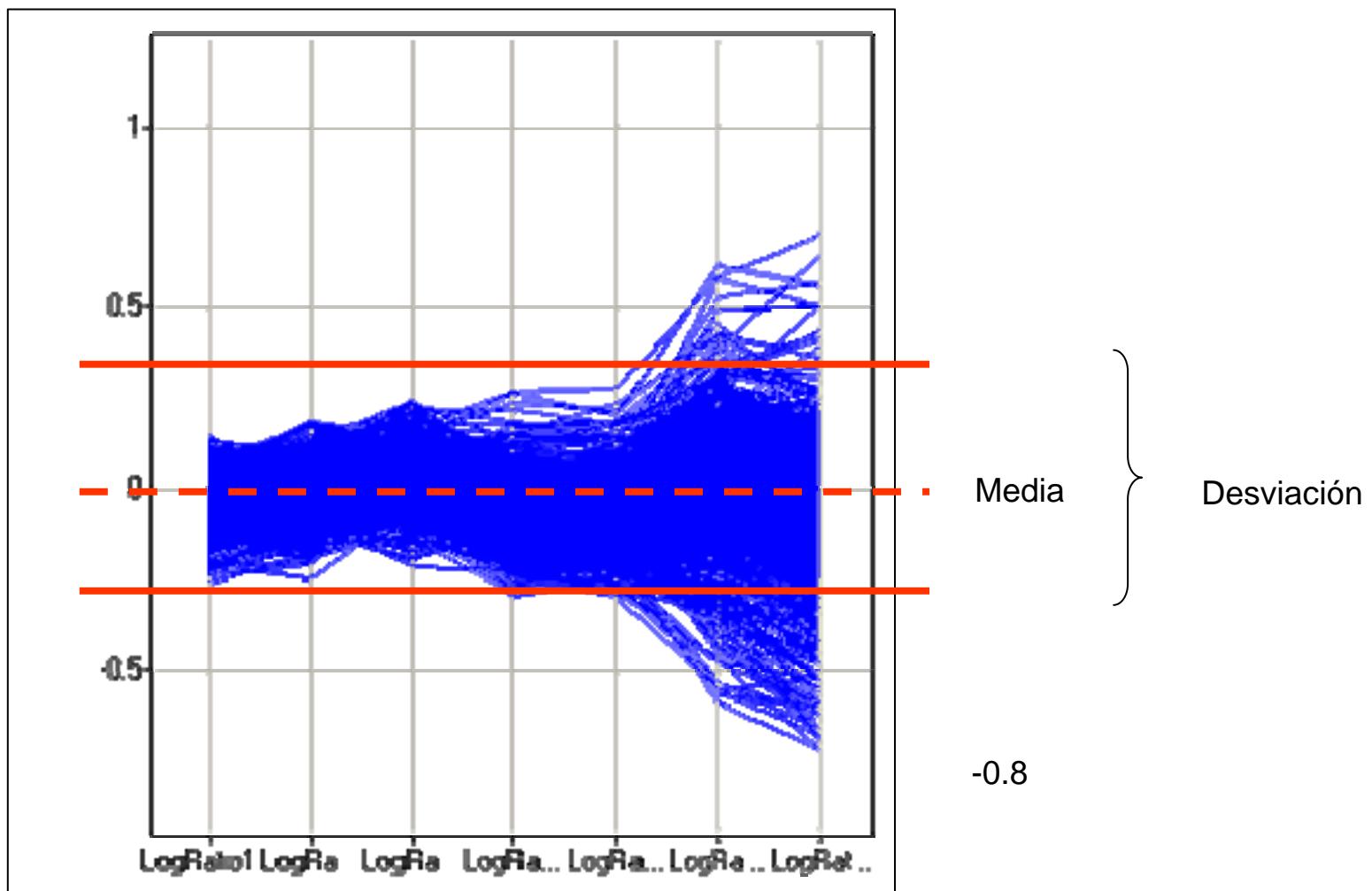
- Los datos de DNA microarray generalmente hay que pre-procesarlos antes de trabajar con ellos:
 - No todos los genes en un chip nos interesan, solo aquellos que hayan variado al menos en una condición experimental.



Umbral



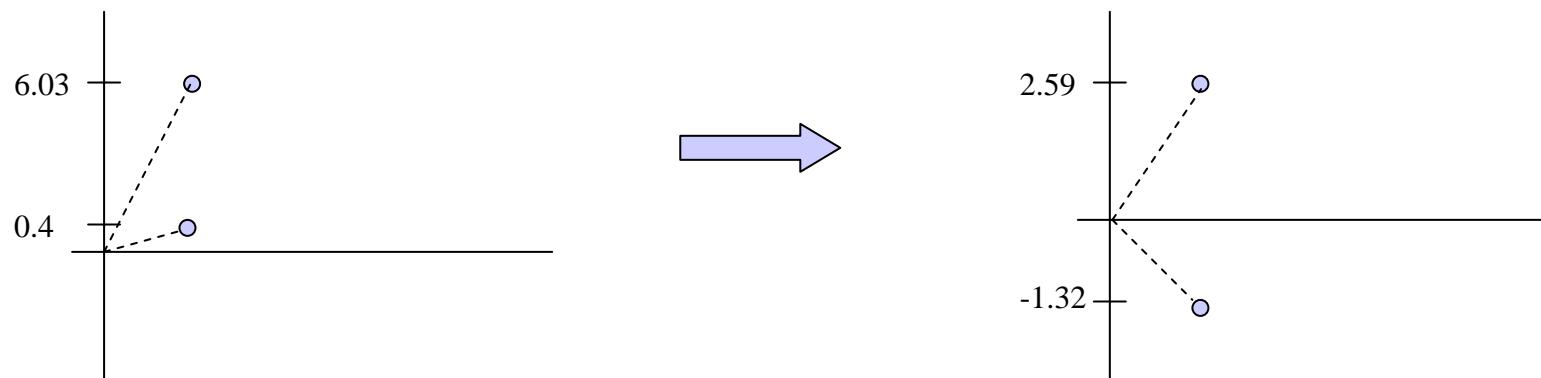
Desviación



Transformación logarítmica:

Los datos de expresión generalmente muestran distribuciones asimétricas respecto a la expresión o inhibición, lo cual dificulta el uso de medidas de distancias para establecer diferencias entre ellos. Para compensar estas diferencias, se utiliza generalmente la transformación logarítmica.

Por ejemplo, en cDNA, genes expresados ocupan la escala de 1 a infinito (o al menos 1000-fold), pero los genes inhibidos ocupan solamente la escala de 0 a 1. La transformación logarítmica pone la escala simétrica alrededor del cero.



Transformación logarítmica:

Muestra/Control:

$$100/1 = 100$$

$$10/1 = 10$$

$$1/1 = 1$$

$$1/10 = 0.1$$

$$1/100 = 0.01$$

Logaritmo:

2

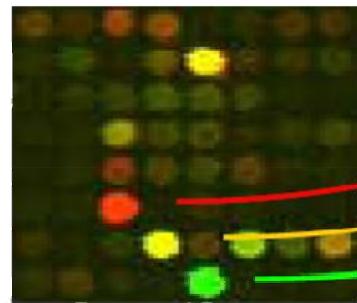
1

0

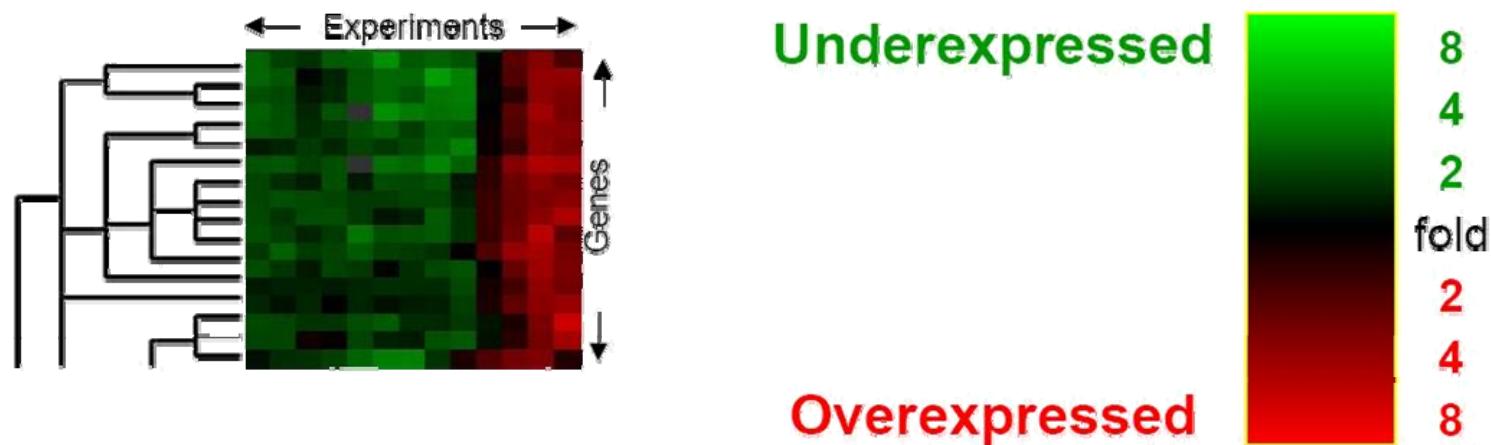
-1

-2

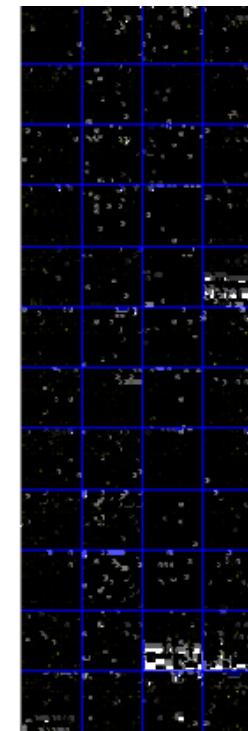
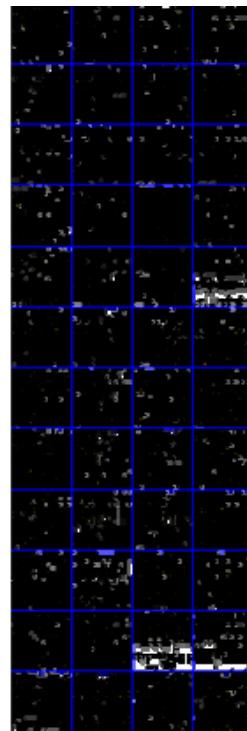
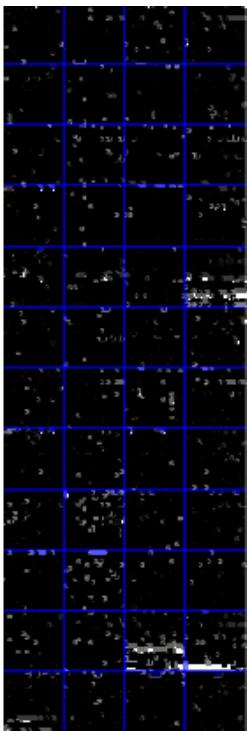
Representación



Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2 \left(\frac{Cy5}{Cy3} \right)$	
200	10000	50.00	5.64	Red
4600	4800	1.00	0.00	Black
9000	300	0.03	-4.91	Green



Valores incompletos (Missing values)

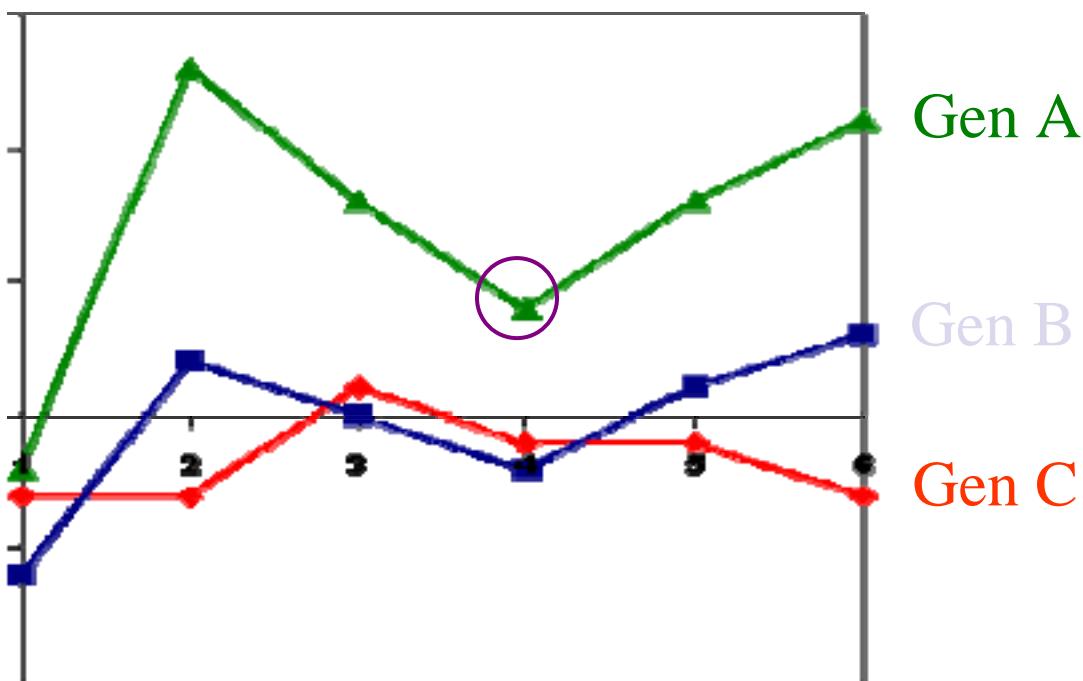


Tabla

Index	Name	ID	Log2Replicate1	Log2Replicate2	Log2Replicate3
13902	0610005A07Rik BG072517	mAB0628	0.332	0.447	0.477
5971	0610006A03Rik BG084074	mAA9100	0.175	0.021	0.095
10740	0610006H08Rik BG074701	mAB3178			
7981	0610006H10Rik BG064737	mAA2144	-0.188	-0.039	-0.385
13179	0610006H10Rik BG064737	mAA2144	-0.12	-0.02	-0.072
4961	0610006I08Rik BG076213	mAB4974	0.303	0.53	0.358
7442	0610006K04Rik BG064645	mAA2043	0.307	0.399	0.672
13718	0610006K04Rik BG064645	mAA2043	-0.123	0.112	
7225	0610006O17Rik BG086123	mAB1625	-0.004	0.228	0.541
10892	0610007A03Rik AU023429	mAA6410	0.545	0.508	0.323
16983	0610007H07Rik BG085143	mAB0381	0.286	0.282	
6906	0610007L03Rik BG069702	mAA7510			
12509	0610007L05Rik BG087267	mAB3112	0.215	0.192	0.318
8006	0610007N03Rik BG077453	mAA1400	-0.053	-0.042	-0.157
13206	0610007N03Rik BG077453	mAA1400	0.043	0.034	0.021
7235	0610007N19Rik BG073658	mAB2057	0.432	0.339	1.176
5641	0610007T007Rik BG079781	mAA4228	0.381	0.497	-0.519
6608	0610007P06Rik BG063045	mAA0046	0.453	0.475	0.224
14640	0610007P06Rik BG063045	mAA0046			
20029	0610007P06Rik BG086880	mAB2552	0.314	0.349	0.495
8593	0610008C08Rik BG071900	mAA9933			
4034	0610008F14Rik BG076975	mAA0805	0.445	0.614	0.041
17162	0610008F14Rik BG076975	mAA0805	0.46	0.672	-0.003
18305	0610008K04Rik BG085113	mAB0350	0.04	0.014	0.276
6801	0610008N23Rik BG086860	mAB2526	0.087	0.153	-0.007
5974	0610009C03Rik BG071176	mAA9082	0.368	0.198	0.188
9383	0610009D07Rik BG076503	mAB5243	0.392	0.363	0.036
11483	0610009D07Rik BG076503	mAB5243	0.292	0.313	0.098
2643	0610009D10Rik BG063439	mAA0605		0.479	
18621	0610009D10Rik BG063439	mAA0605	0.259	0.372	-0.099
2588	0610009D10Rik BG077769	mAA1787	-0.223	-0.157	-0.496
18598	0610009D10Rik BG077769	mAA1787	-0.203	-0.032	-0.667
9878	0610009D16Rik BG086417	mAB1991	0.217	0.065	
7466	0610009E20Rik BG077106	mAA0945	0.539	0.59	0.41
13772	0610009E20Rik BG077106	mAA0945	0.25	0.362	0.194
9084	0610009H04Rik BG083210	mAA8139	0.178	0.246	0.193
8568	0610009H04Rik BG087149	mAB2975	0.126	0.187	
13812	0610009J22Rik BG075746	mAB4420	-0.138	-0.029	0.079
1063	0610009M14Rik BG074040	mAB2401			
8888	0610009M14Rik BG077537	mAA1508	-0.062	0.083	-0.608
12324	0610009M14Rik BG077537	mAA1508	-0.128	-0.074	0.138
4911	0610009N12Rik BG065259	mAA2695	0.199	0.251	-0.047
16239	0610009N12Rik BG065259	mAA2695	0.258	0.445	0.149

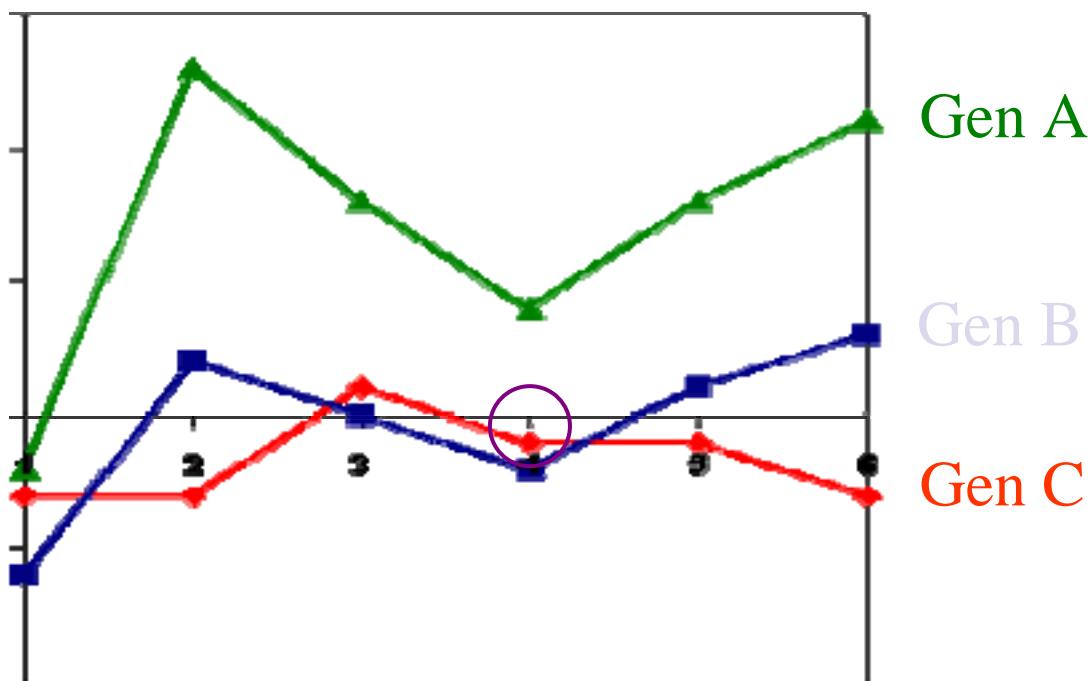
Métodos:

- Fijar a un valor predeterminado: normalmente 0
- Sustituir el valor perdido por la media de toda la columna (experimento)
- Sustituir el valor perdido por la media de toda la fila (gen)
- Interpolación local pesada



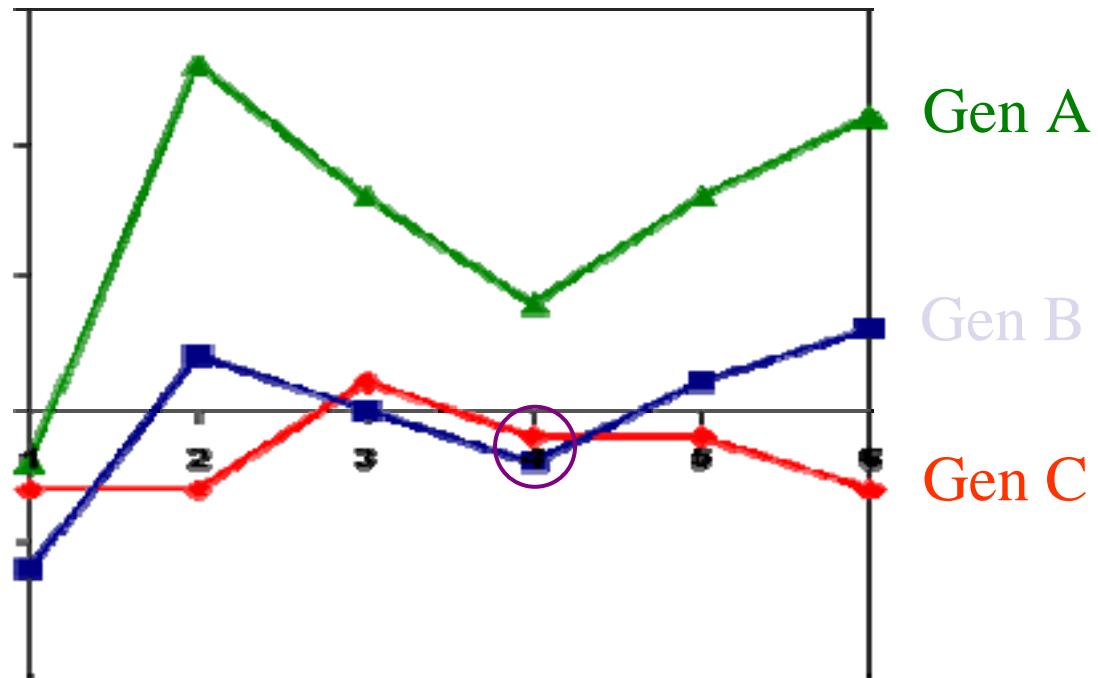
Métodos:

- Fijar a un valor predeterminado: normalmente 0
- Sustituir el valor perdido por la media de toda la columna (experimento)
- Sustituir el valor perdido por la media de toda la fila (gen)
- Interpolación local pesada



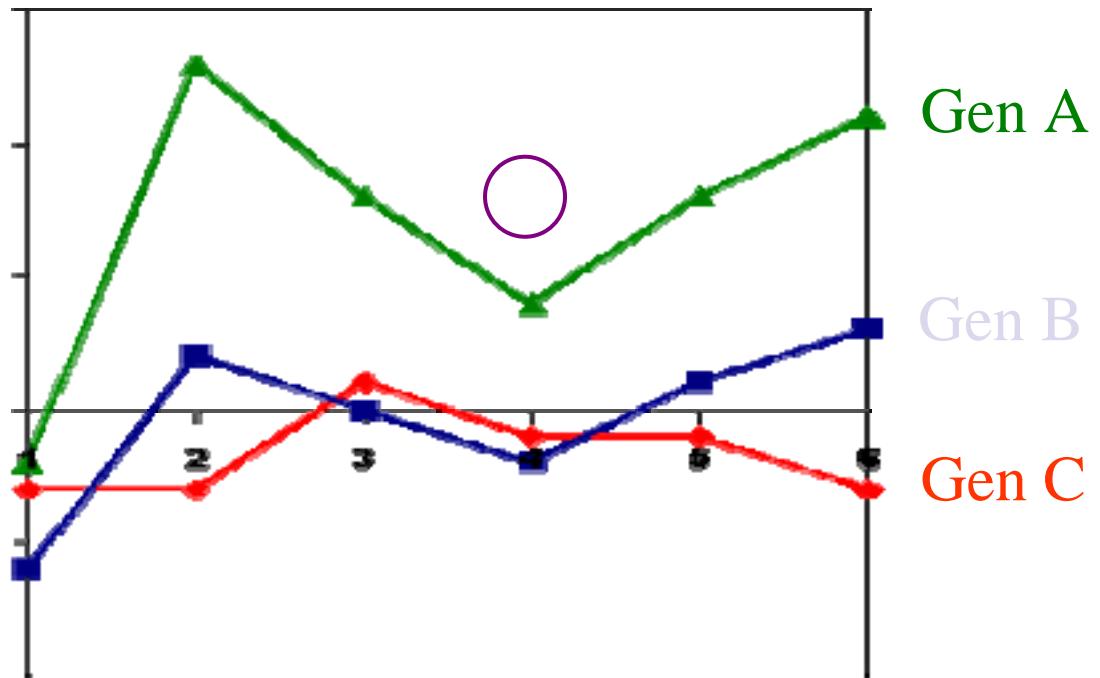
Métodos:

- Fijar a un valor predeterminado: normalmente 0
- Sustituir el valor perdido por la media de toda la columna (experimento)
- Sustituir el valor perdido por la media de toda la fila (gen)
- Interpolación local pesada

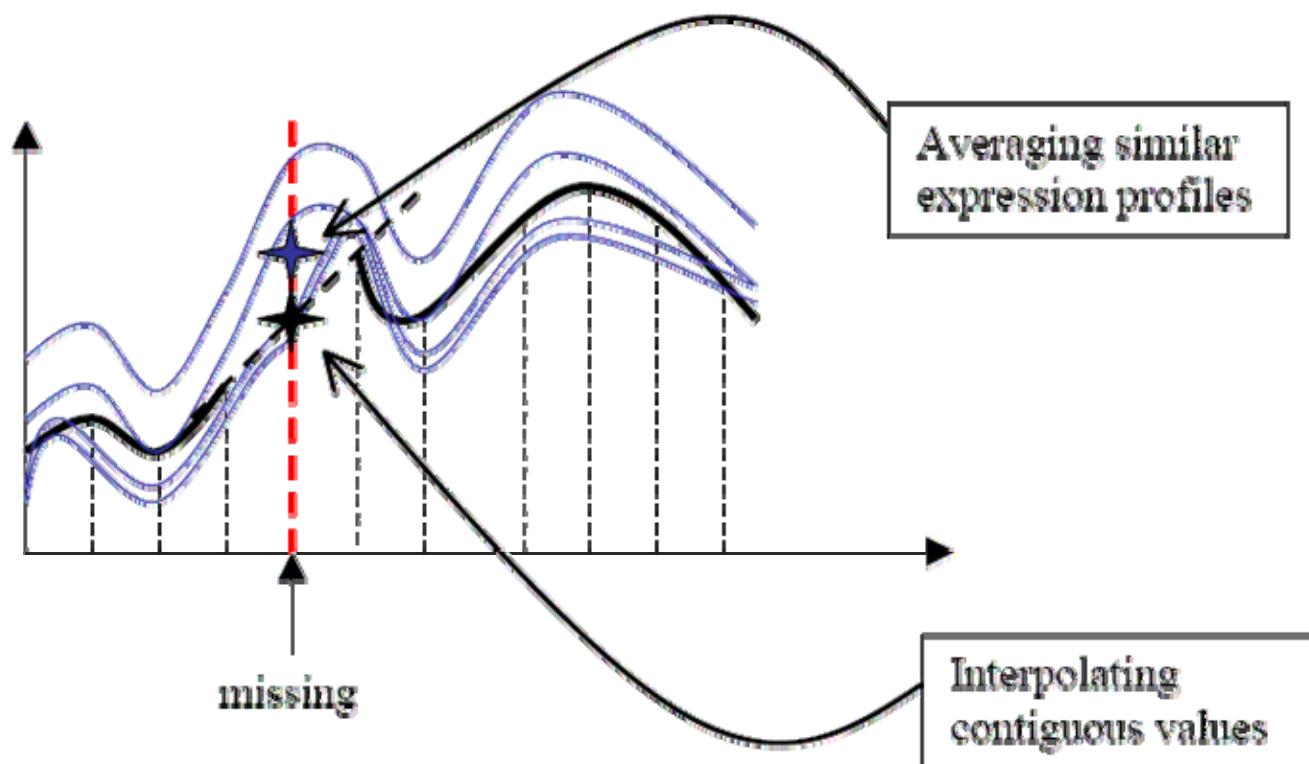


Métodos:

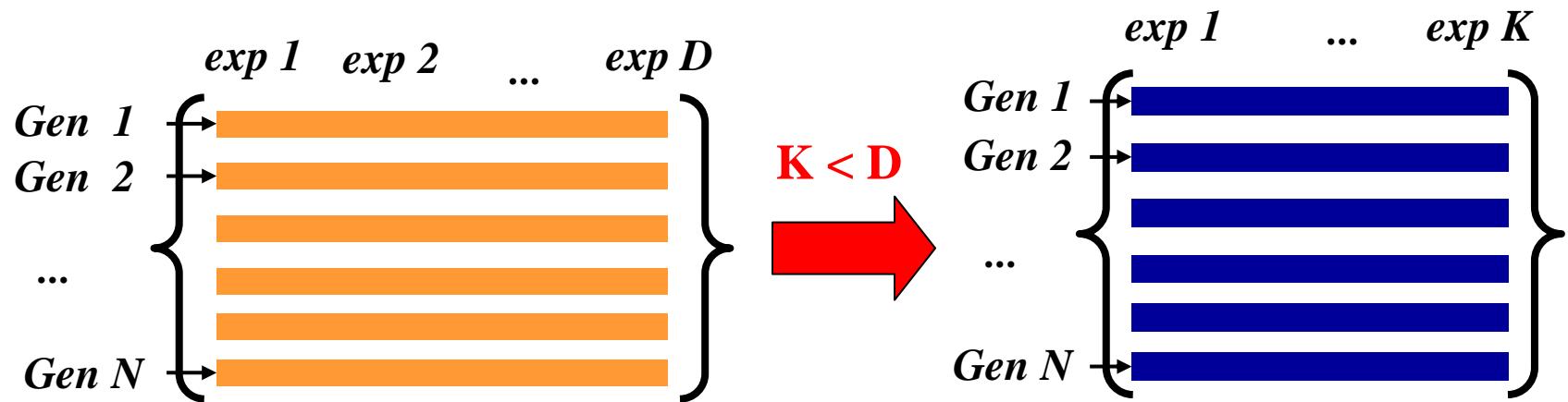
- Fijar a un valor predeterminado: normalmente 0
- Sustituir el valor perdido por la media de toda la columna (experimento)
- Sustituir el valor perdido por la media de toda la fila (gen)
- Interpolación local pesada



Interpolación pesada por k vecinos mas cercanos



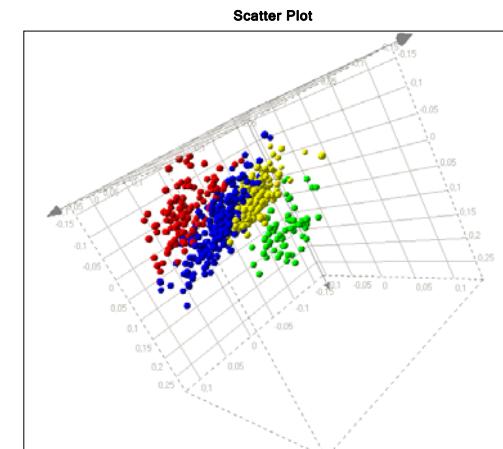
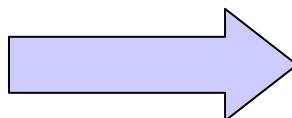
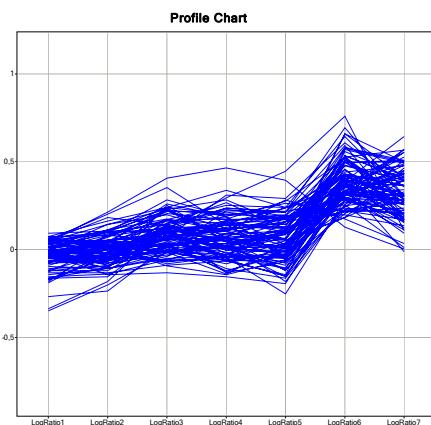
Reducción de dimensionalidad



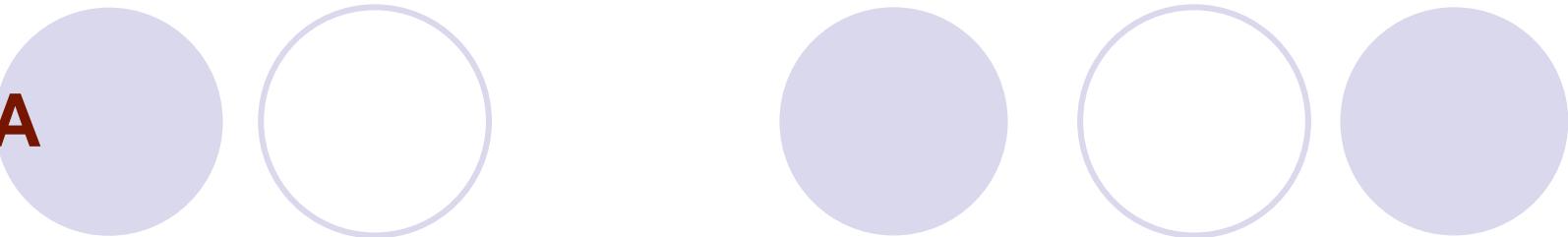
La idea es reducir el número de variables (descriptores) de los datos. En el caso de que los datos sean expresión génica, los genes se convertirán en un conjunto de “pseudo-genes”, los cuales contendrán k nuevos experimentos, siendo $k \ll$ número de experimentos originales.

Principal Component Analysis (PCA)

- Es una transformación ortogonal del sistema de coordenadas en el cual están representados los datos.
- Los nuevos valores de coordenadas mediante los cuales representamos los datos, se les llaman componentes principales.
- Los componentes principales no están correlacionados.
- Usualmente los primeros “nuevos” ejes correspondientes a los primeros componentes principales contienen gran parte de la información de los datos, así que el resto de componentes pueden ser eliminados.



PCA

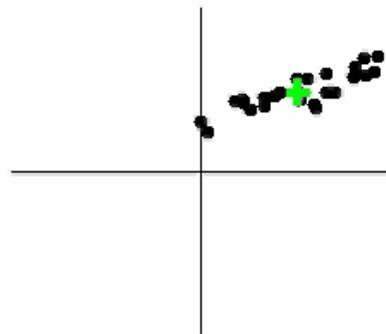


- Dado un conjunto de datos en un espacio d-dimensional, PCA describe la forma y la localización de la nube de puntos en este espacio d-dimensional.
- PCA se realiza en dos pasos:
 - Traslación: traslada la nube de puntos al origen
 - Rotación: Rota los alrededor del origen
- Otra forma de verlo es aplicar la rotación y traslación a los ejes de coordenadas en vez de a los datos.

Ejemplo de PCA

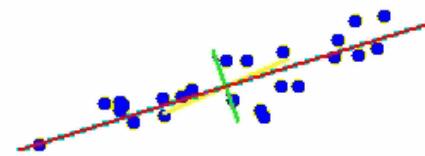
Datos “crudos”

Raw Gaussian 3a



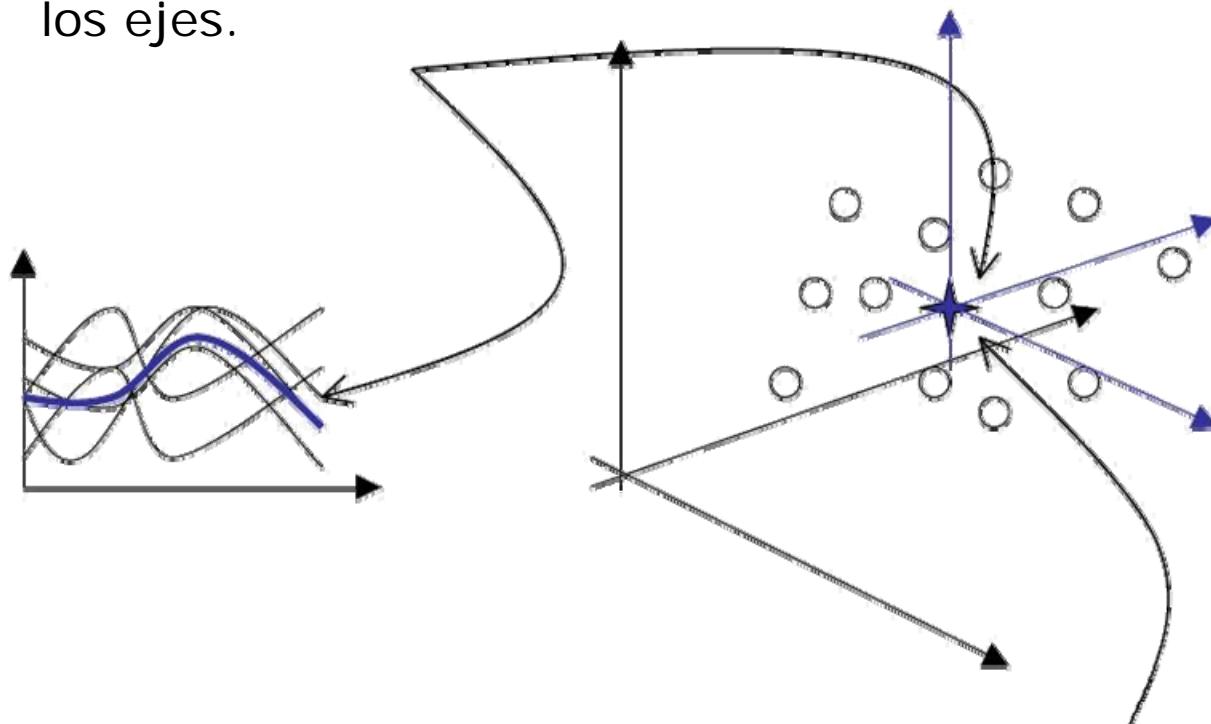
Nuevos ejes: PCA

PC Axes for Gaussian 3a

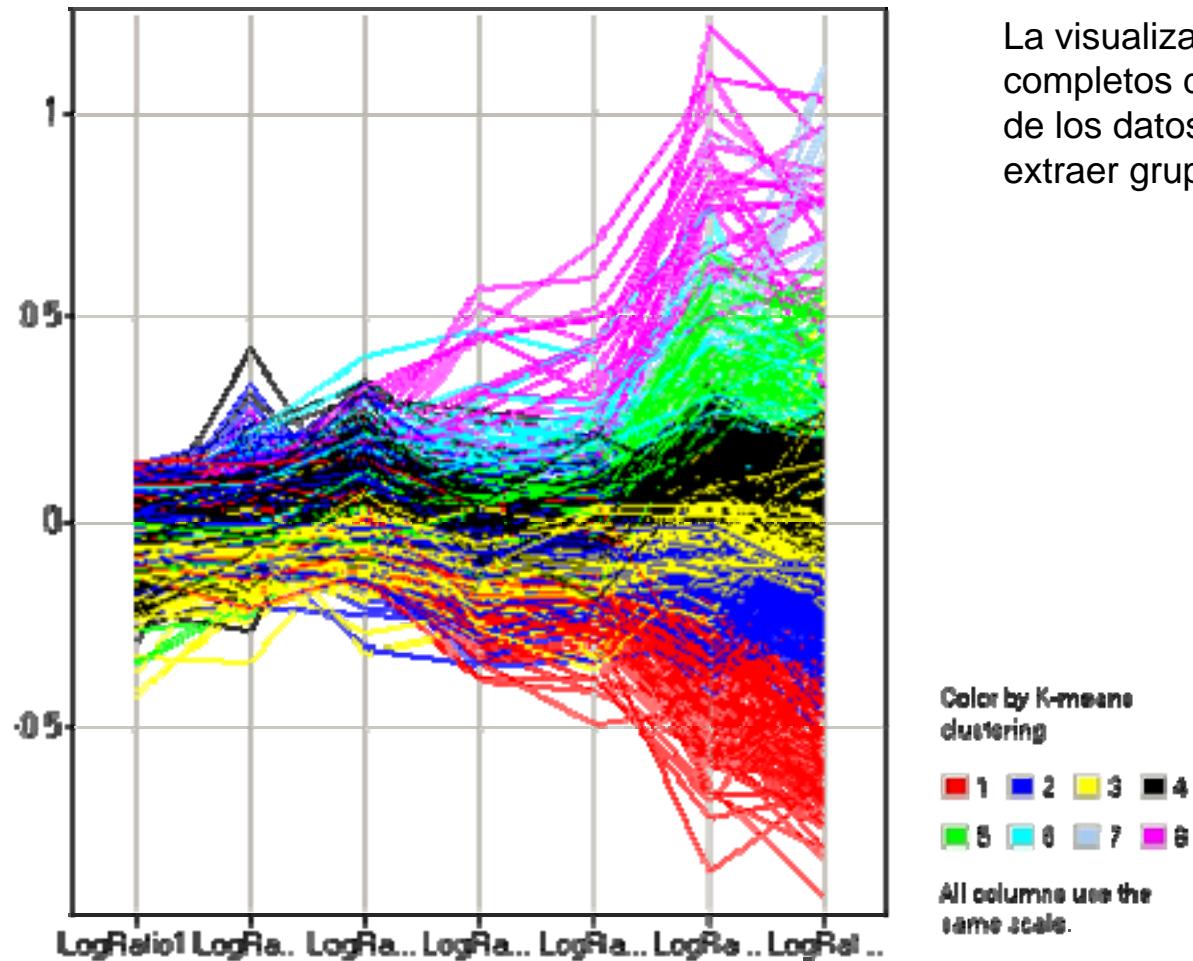


PCA en expresión génica

Transformación de
los ejes.

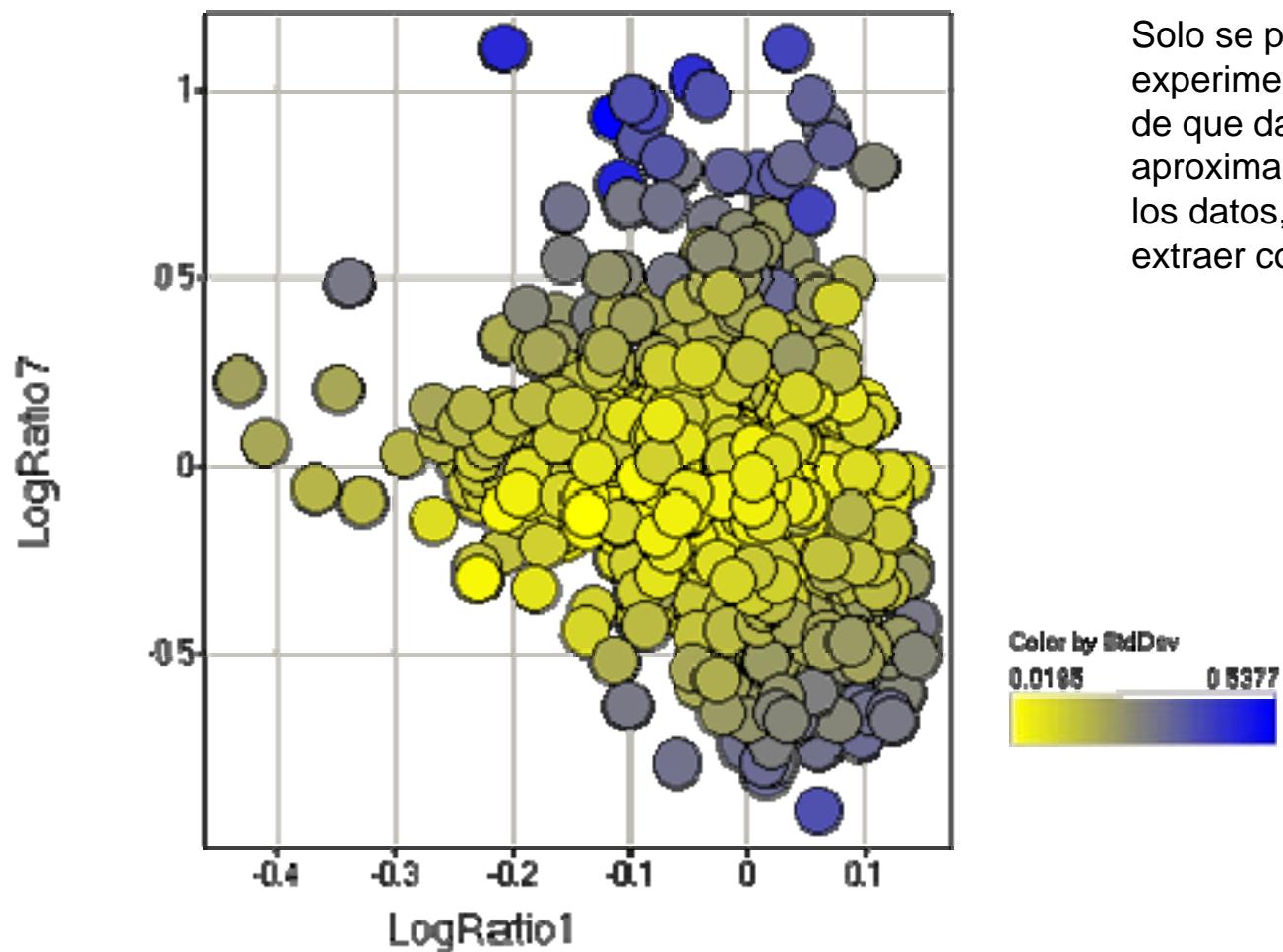


Visualización de perfiles



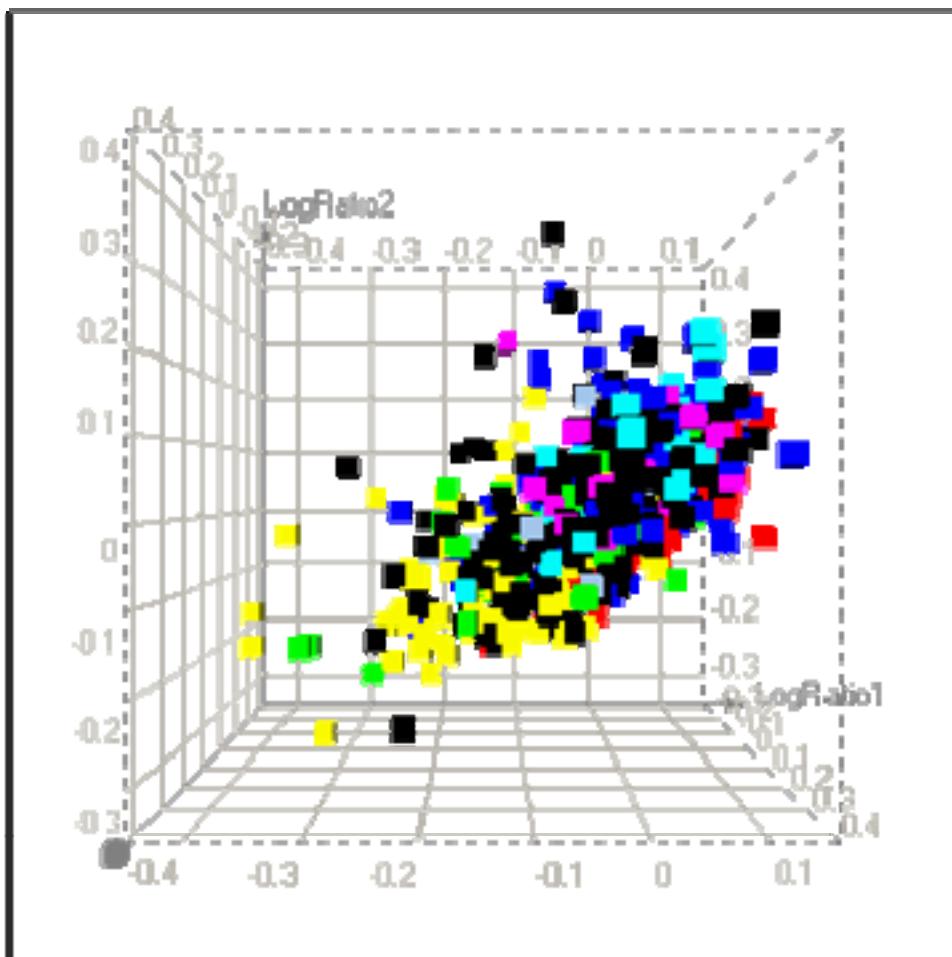
La visualización de perfiles completos da una idea de la forma de los datos, pero es muy difícil extraer grupos de manera visual.

Scatter Plot 2D



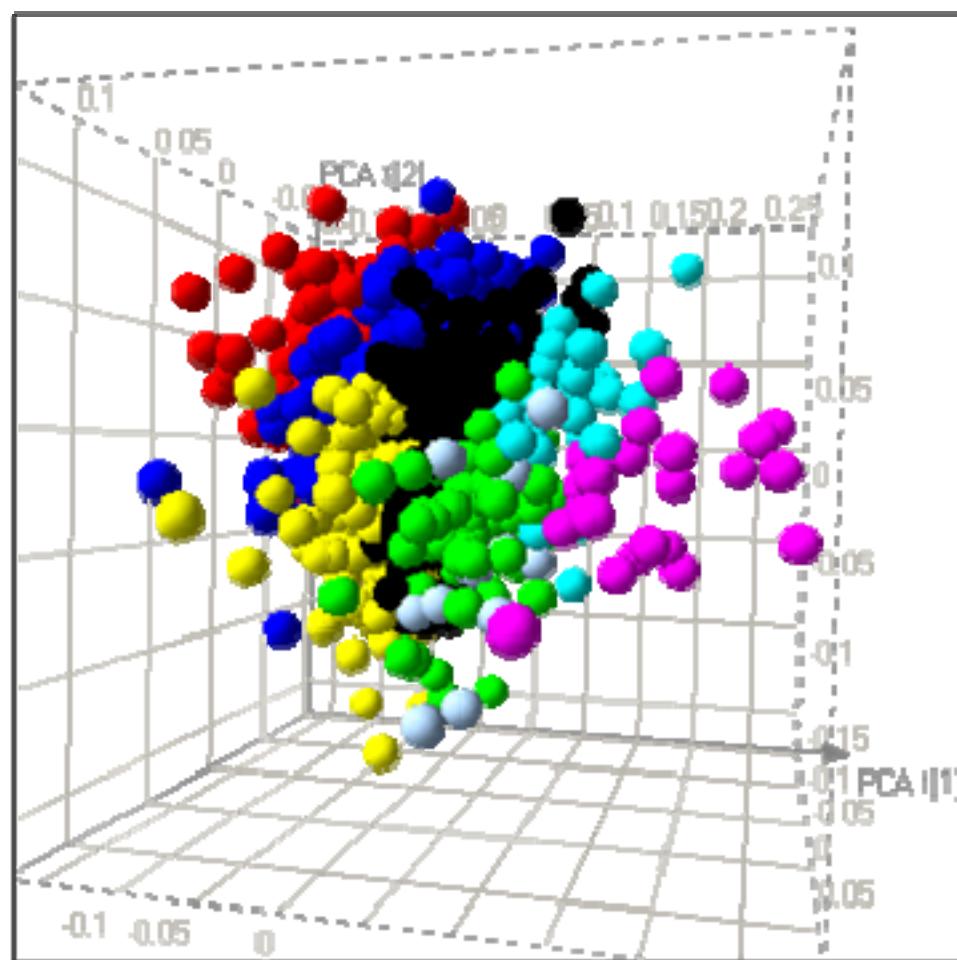
Solo se puede visualizar dos experimentos a la vez. A pesar de que da una idea aproximada de la estructura de los datos, no es suficiente para extraer conclusiones.

Scatter Plot (3D)



Solo se puede visualizar tres experimentos a la vez. A pesar de que da una idea aproximada de la estructura de los datos, no es suficiente para extraer conclusiones.

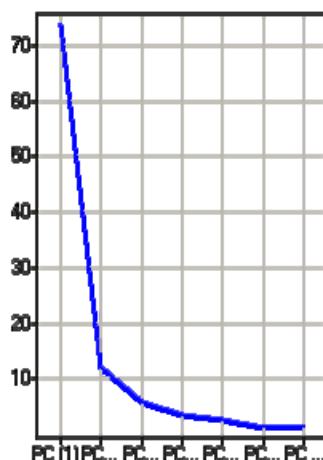
PCA (3 componentes)



Los 3 primeros componentes principales se pueden mostrar en un gráfico 3D. Al representar los ejes de mayor varianza de los datos, su representación es mas completa y fiable.

Interpretación:

Principal Component	Eigenvalue	Eigenvalue (%)	Cumulative Eigenvalue (%)
PC (1)	0.149	73.661	73.661
PC (2)	2.4e-2	12.083	85.744
PC (3)	1.2e-2	5.765	91.509
PC (4)	7.3e-3	3.600	95.109
PC (5)	5.0e-3	2.499	97.608
PC (6)	2.4e-3	1.197	98.806
PC (7)	2.4e-3	1.194	100.000



Preguntas más frecuentes:

- ¿Qué genes están o no expresados?
 - En distintas células
 - En condiciones externas diferentes
 - En diferentes estados de enfermedades
- ¿En cuánto ha cambiado sus niveles de expresión?
- ¿El cambio en la expresión de los genes está correlacionado con otros parámetros externos?

Técnicas de estudio: Estadística descriptiva

Preguntas más frecuentes:

- Los proyectos de secuenciación están produciendo gran cantidad de genes cuya función se desconoce. ¿Se puede utilizar los datos de expresión génica para “predecir” las funciones de los nuevos genes?

	Gene Function	Exp 1	Exp 2	Exp 3	Exp4	Exp5
Exp Attributes		type I	type II	type III	type II	type III
Gene 1	kinase	0.21	0.56	0.72	0.38	0.69
Gene 2	kinase	0.69	0.64	0.55	0.57	0.79
Gene 3	protease	0.01	0.74	0.49	0.60	0.38
Gene 4	protease	0.37	0.91	0.98	0.31	0.25
Gene 5	protease	0.14	0.34	0.92	0.43	0.59
Gene 6	metabolism	0.28	0.45	0.60	0.60	0.46
Gene 7	metabolism	0.66	0.60	0.17	0.15	0.61
Gene 8	?	0.28	0.86	0.24	0.71	0.99
Gene 9	?	0.13	0.30	0.55	0.84	0.86
Gene 10	?	0.85	0.85	0.46	0.77	0.78

Técnicas de estudio: Métodos de clasificación supervisados, redes neuronales, etc.

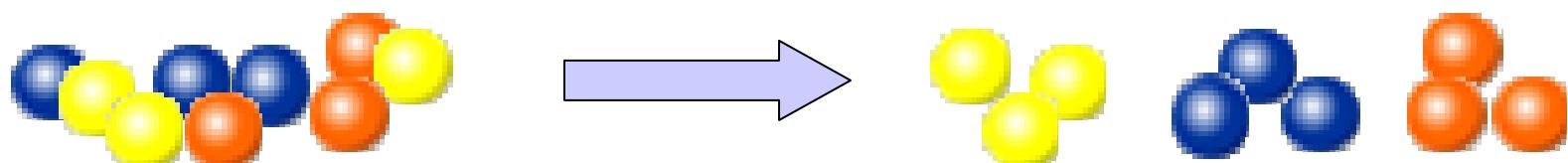
Preguntas más frecuentes:

- **¿Podemos utilizar los patrones de expresión de los genes para agrupar genes cuya función se desconoce?**
 - Clasificación funcional de genes cuya función se desconoce (patrones de expresión similares puede implicar funciones similares)
 - Clasificación molecular de muestras (por ejemplo subtipos de tumores indistinguibles morfológicamente).
 - Descifrar mecanismos de regulación mediante la identificación de grupos de genes que se co-expresen y que probablemente estén co-regulados.
 - Identificación de patrones de expresión de genes “diagnóstico” (cuya función se conoce)

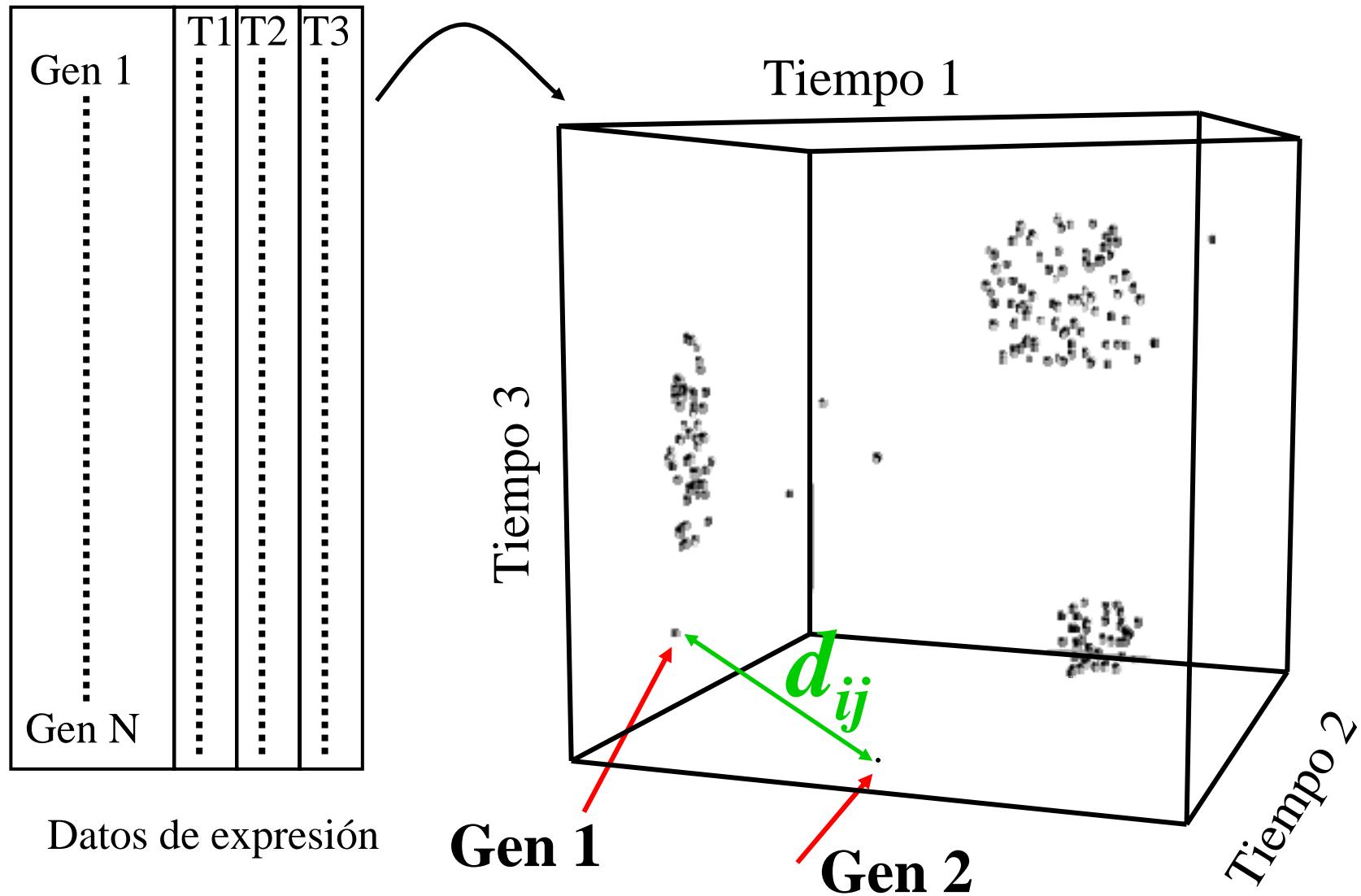
Técnicas de estudio: Análisis de agrupamiento

Análisis de agrupamiento (Cluster Analysis):

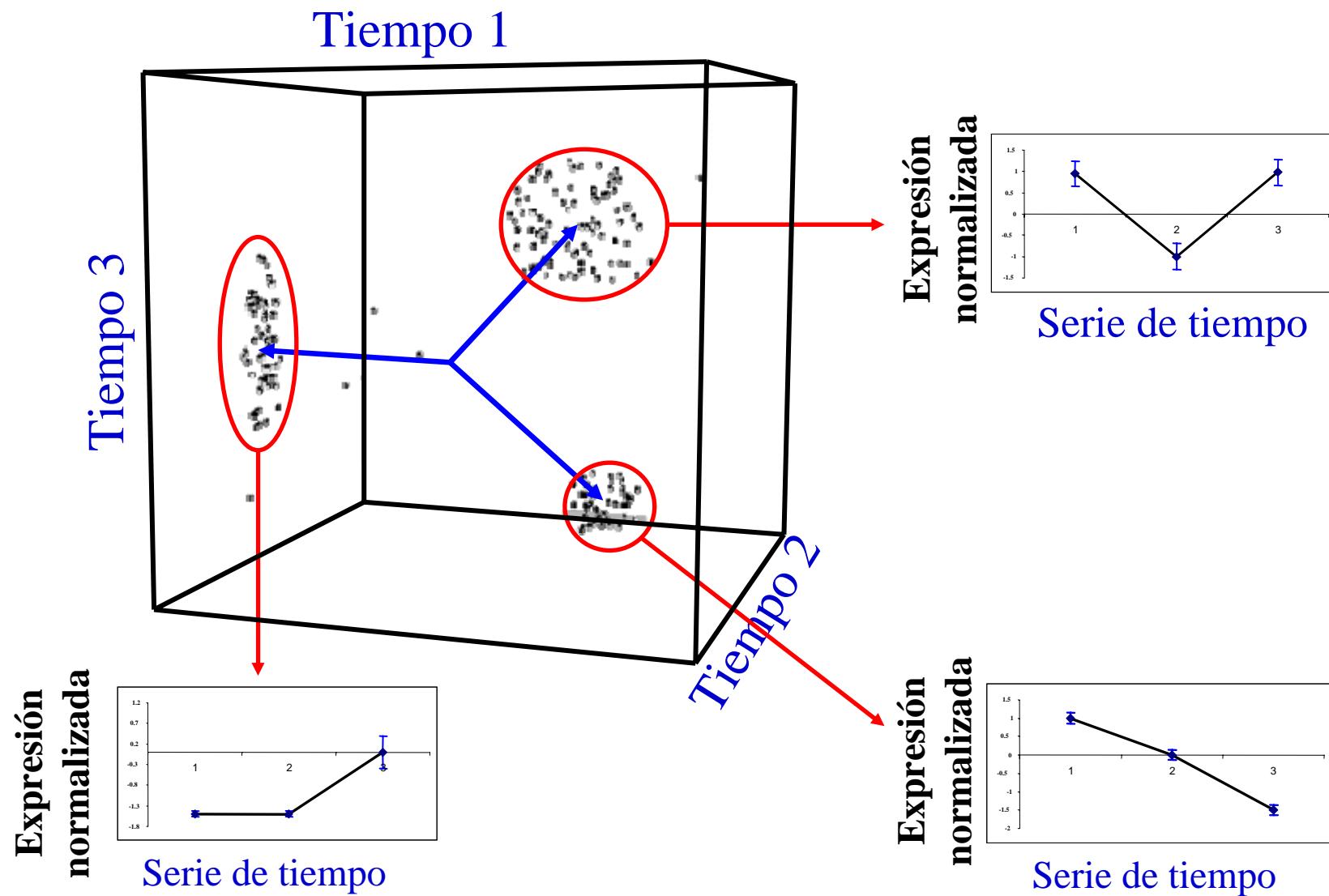
- Entrada: n datos, \mathbf{X}_i , $i=1,2,\dots,N$ en un espacio p dimensional. Por ejemplo: 1000 (n) genes en 10 (p) condiciones experimentales.
- Objetivo: Encontrar grupos ó “clusters” naturales. Los datos en un mismo grupo o cluster deben ser “más similares”
- Nota importante: ¿Cuántos grupos tenemos?



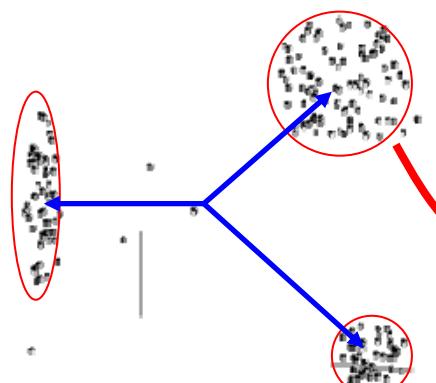
Representación de los datos de expresión:



Identificación de patrones de expresión prevalentes (grupos de genes ó clusters)



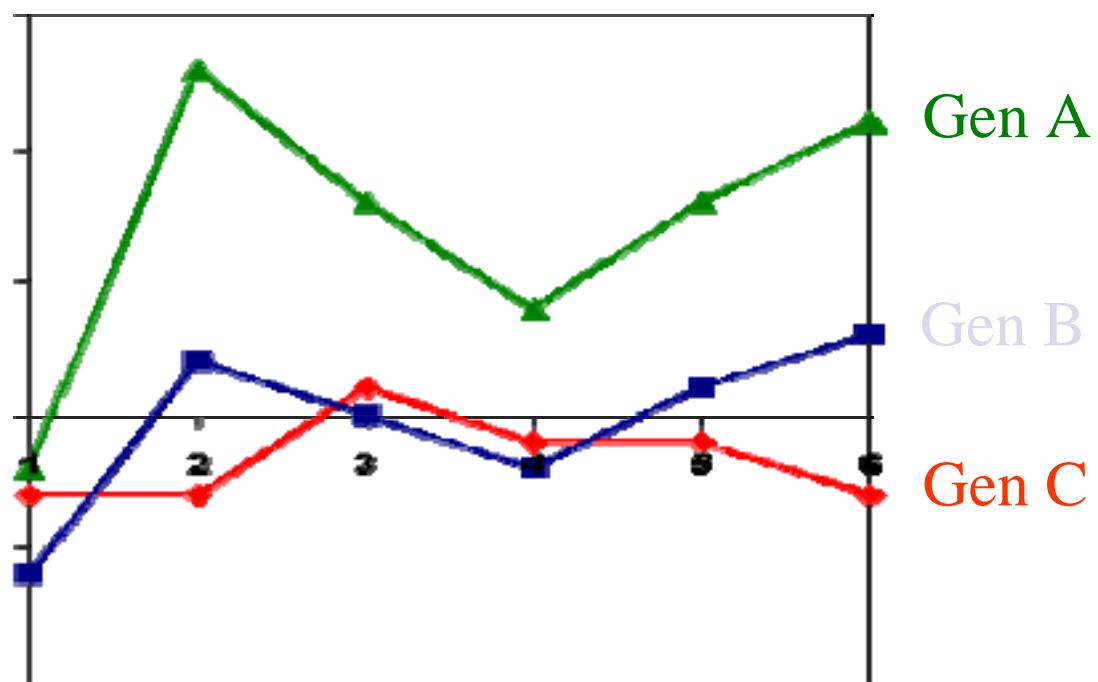
Evaluación del contenido de los grupos:



Genes

- gpm1** → Glycolysis
- HTB1** → Nuclear Organization
- RPL11A**
- RPL12B**
- RPL13A**
- RPL14A**
- RPL15A**
- RPL17A**
- RPL23A**
- TEF2** → Translation
- YDL228c**
- YDR133C**
- YDR134C**
- YDR327W**
- YDR417C**
- YKL153W**
- YPL142C** → Unknown

Distancia entre patrones de expresión: ¿Cuál es la más apropiada?



Métricas más comunes:

1, $r = 2$ (Distancia euclídea)

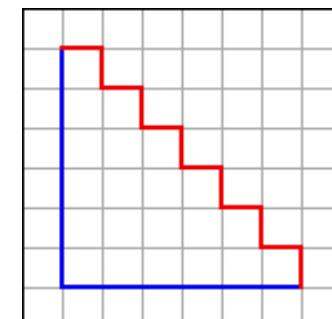
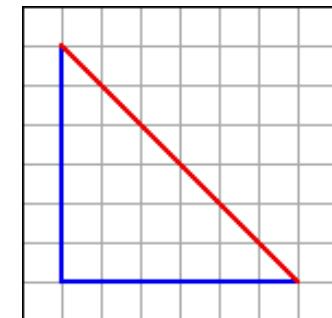
$$d(x, y) = \sqrt{2} \sum_{i=1}^p |x_i - y_i|^2$$

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

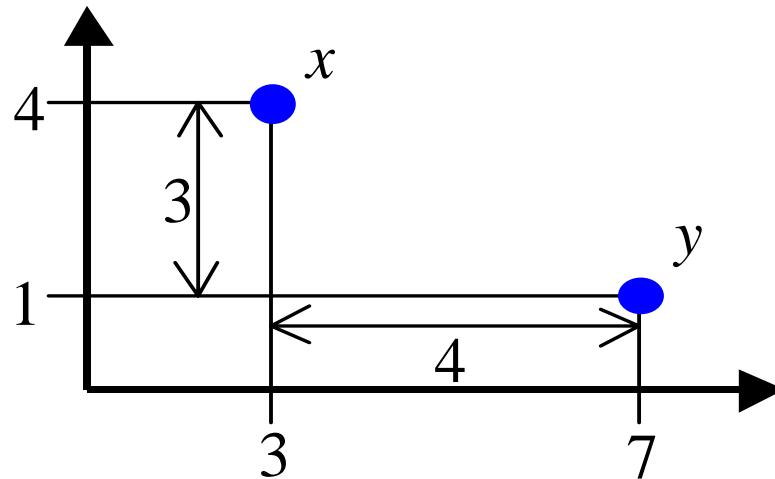
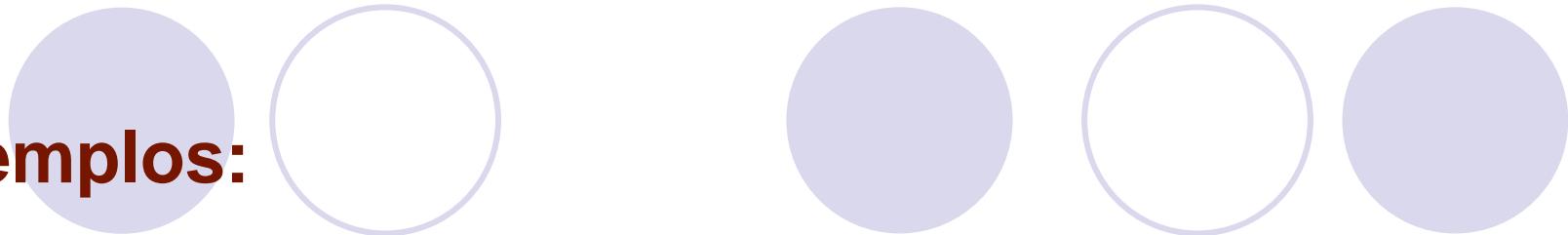
$$d(x, y) = \sum_{i=1}^p |x_i - y_i|$$

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

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



Ejemplos:



1, Euclidean distance : $\sqrt{4^2 + 3^2} = 5$.

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

3, "sup" distance : $\max\{4, 3\} = 4$.

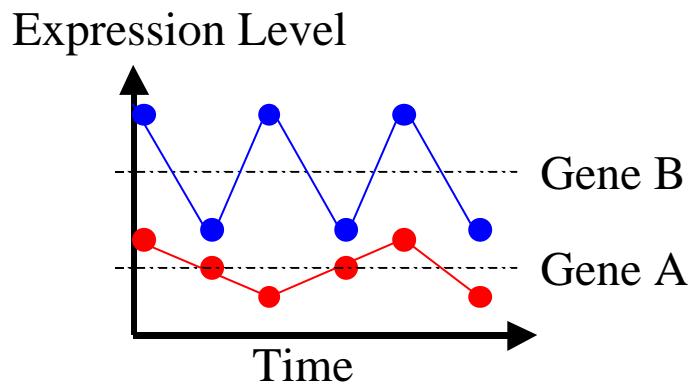
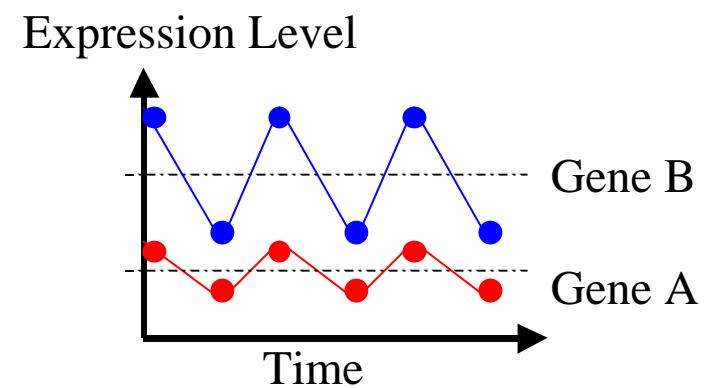
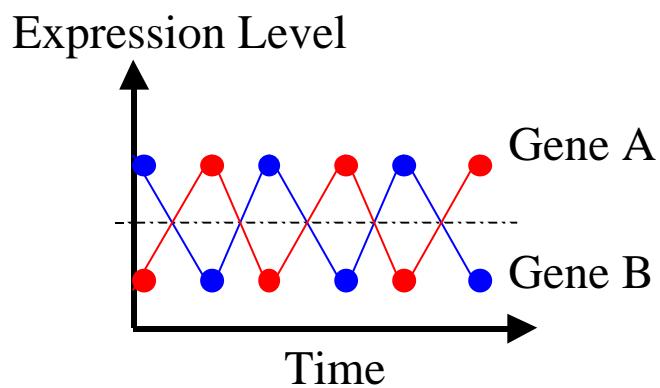
Otro método de similitud: coeficiente de correlación

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

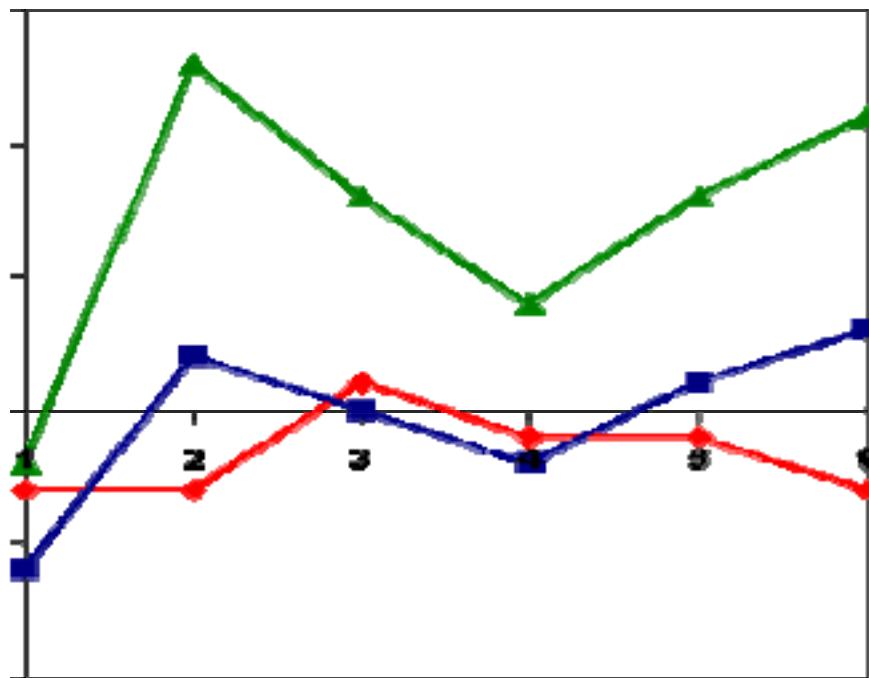
averages: $\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i$ and $\bar{y} = \frac{1}{p} \sum_{i=1}^p y_i$.

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

Ejemplos:



Distancia entre genes: ¿Euclidea ó Correlación?



$$\text{Euclidea: } d_{x,y} = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

$$\text{Correlación: } d_{x,y} = 1 - \frac{1}{p} \sum_{i=1}^p \left[\frac{(x_i - \bar{x})}{\delta x} \right] \left[\frac{(y_i - \bar{y})}{\delta y} \right]$$

Euclidea: Tiende a agrupar perfiles de acuerdo al valor absoluto de las diferencias entre los niveles de expresión. **Rojo** y **Azul**

Correlación: Tiende a agrupar perfiles de acuerdo a la tendencia de los mismos. **Verde** y **Azul**

En DNA Arrays, la distancia de correlación usualmente tiene más significado biológico que la distancia Euclidea.

Métodos de clustering:

- ✓ Jerárquicos:
 - Aglomerativos (HAC)
 - Divisivos
- ✓ Particionales
 - K-means
 - Fuzzy K-means
- ✓ Basados en modelos
- ✓ Basados en Densidad (EM)
- ✓ Basados en grid (CLIQUE)
- ✓ Basados en grafos (CLICK)
- ✓ Redes neuronales (SOM)

Se necesita calcular la **distancia** entre el **nuevo cluster** y los demás. Existen varios tipos:

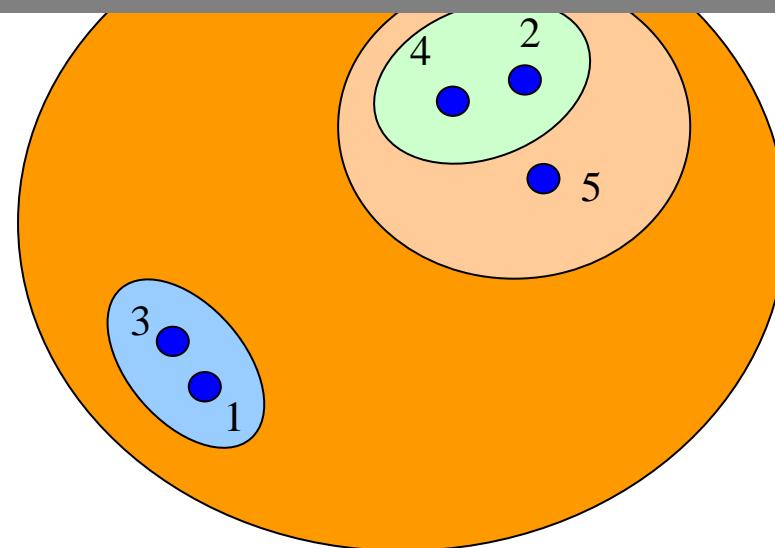
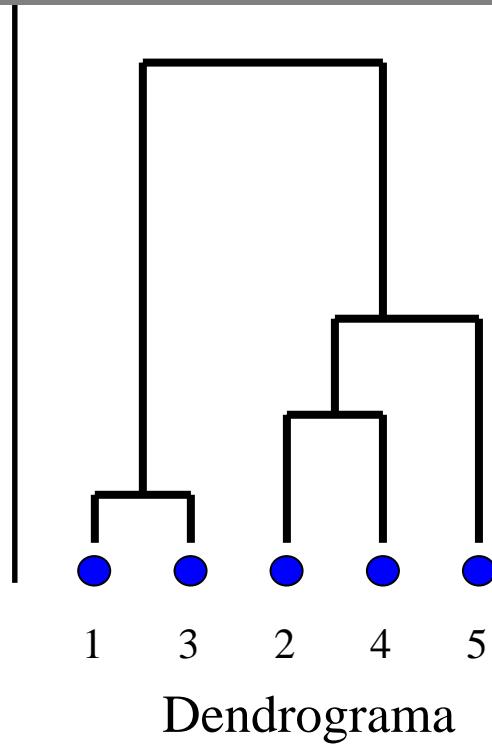
Single Linkage: distancia entre el par de puntos más cercano.

Complete Linkage: distancia entre el par de puntos más alejado.

Average Linkage: distancia promedio entre todos los pares de puntos.

Centroids: distancia entre los centros de los clusters.

Ward: Une clusters que sean mas “compactos”



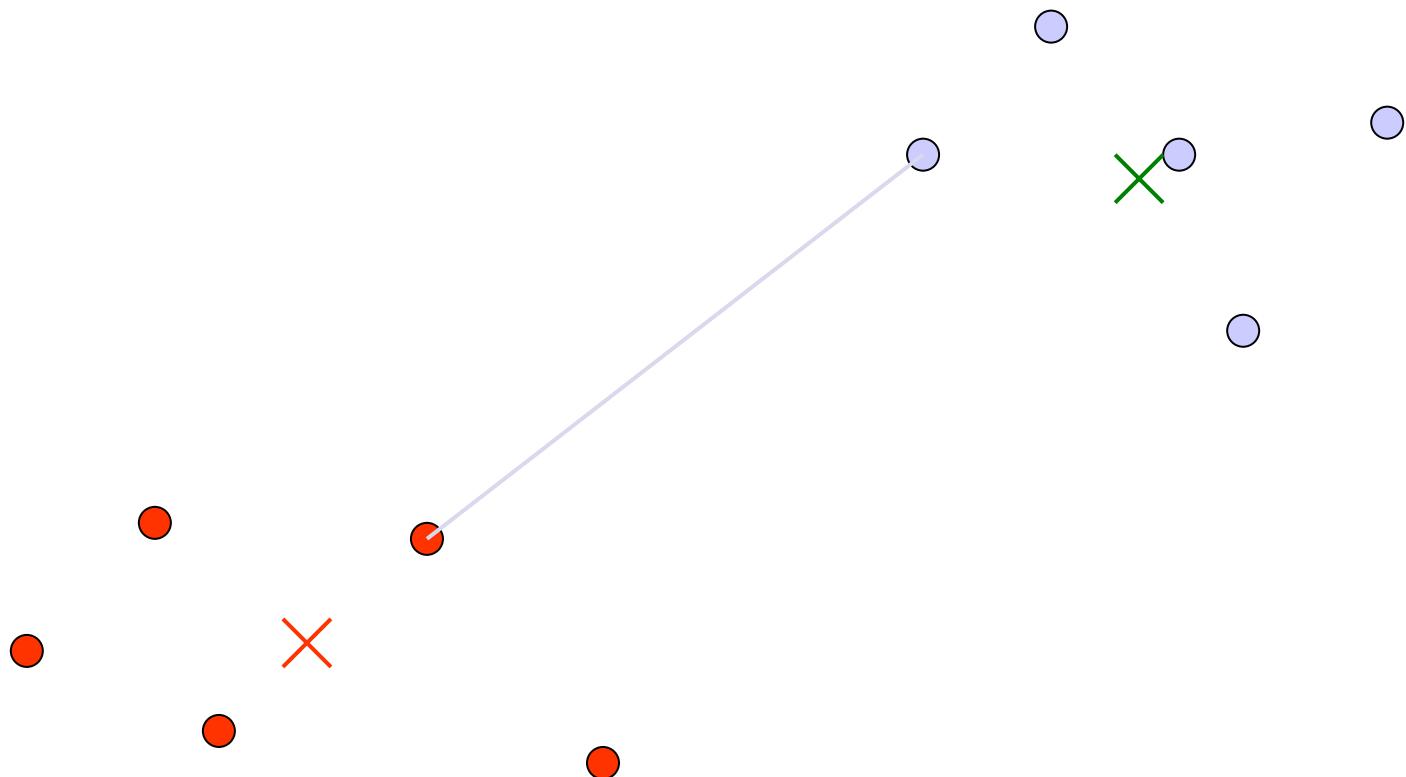
El dendrogram induce un **orden lineal** de los datos.

Single Linkage: distancia entre el par de puntos más cercano.

Tipo de clusters: alongados, tipo cadenas.

Ventajas: Simples, eficientes, propiedades teóricas conocidas

Desventajas: No son adecuados para clusters no bien separados o esféricos

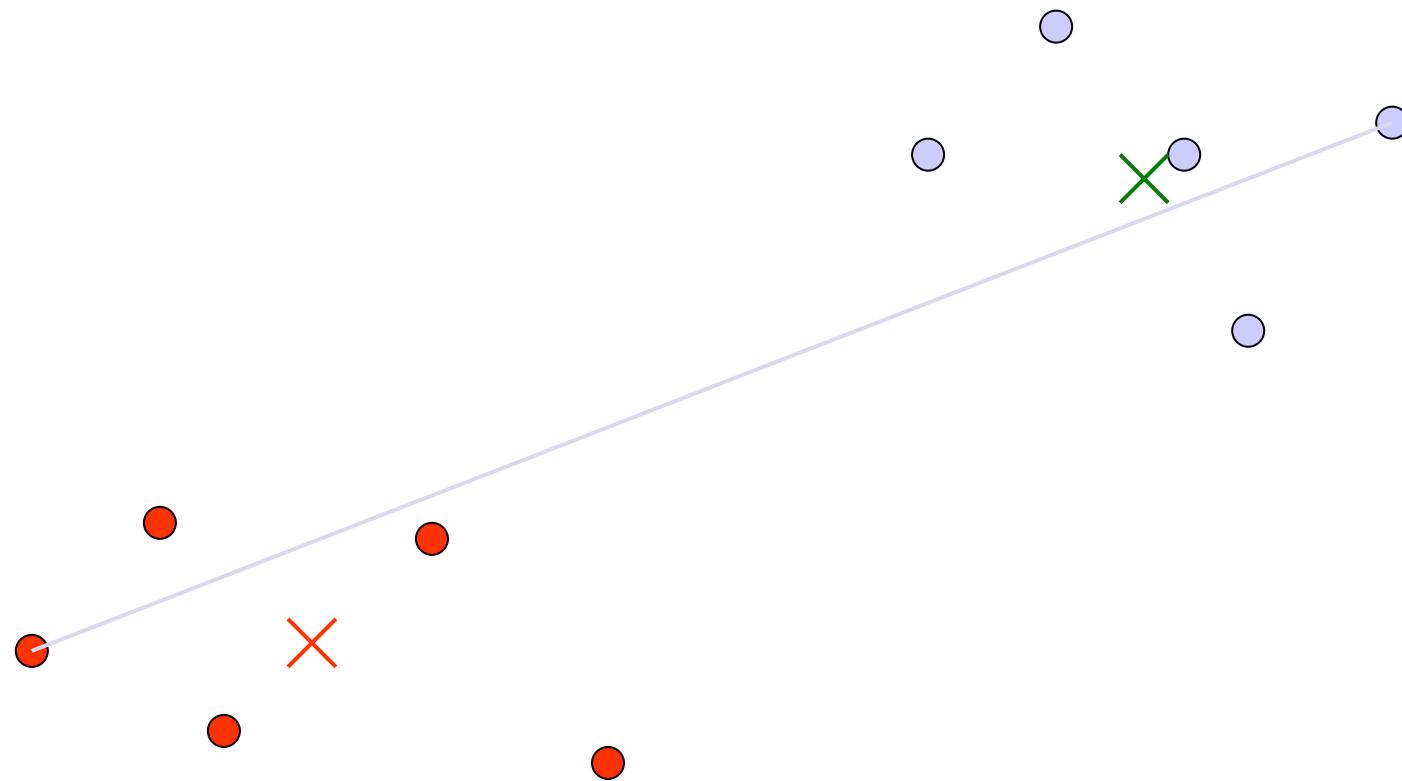


Complete Linkage: distancia entre el par de puntos más alejado.

Tipo de clusters: compactos, esféricos. Tiende a extraer “nubes de puntos”

Ventajas: Funcionan bien cuando los clusters son muy compactos.

Desventajas: No es eficiente en conjuntos grandes de datos.

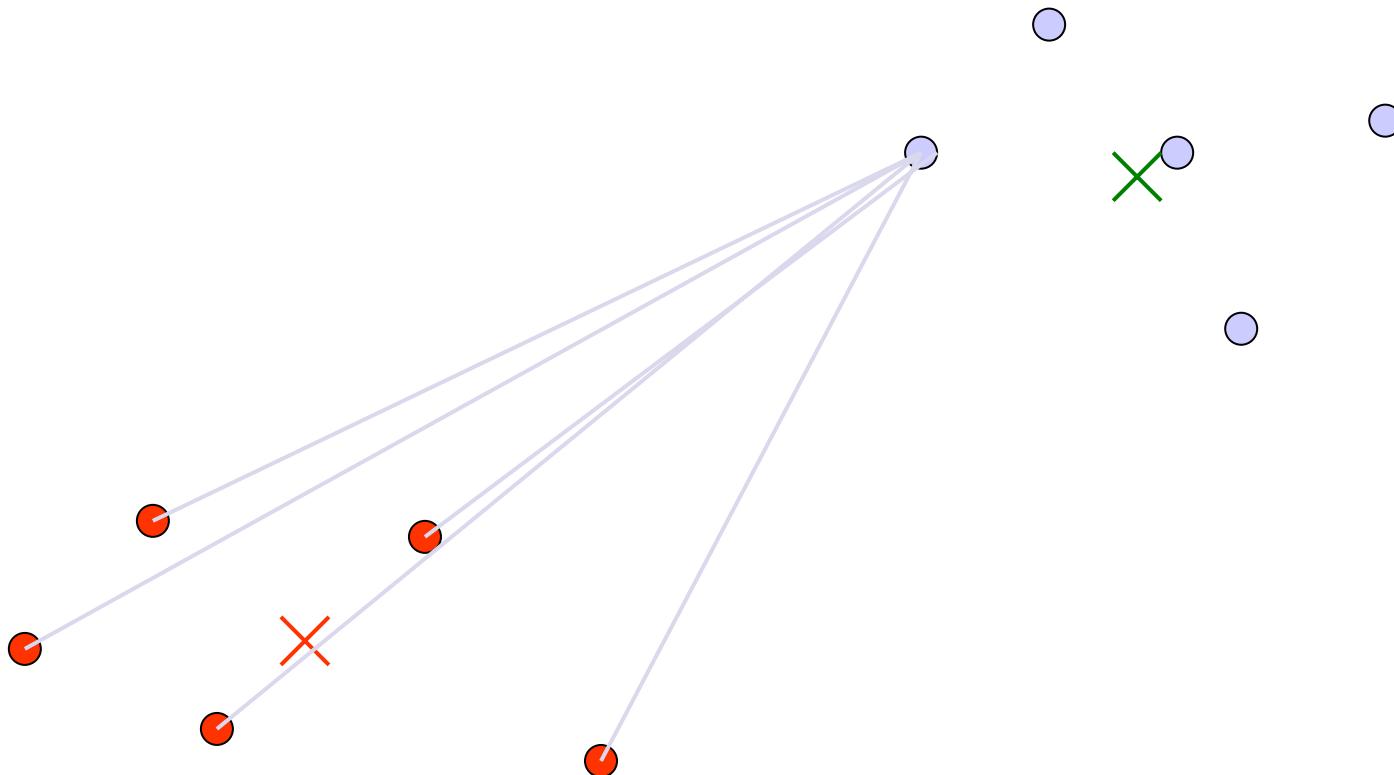


Average Linkage: distancia promedio entre todos los pares de puntos.

Tipo de clusters: Intermedio entre single y complete linkage.

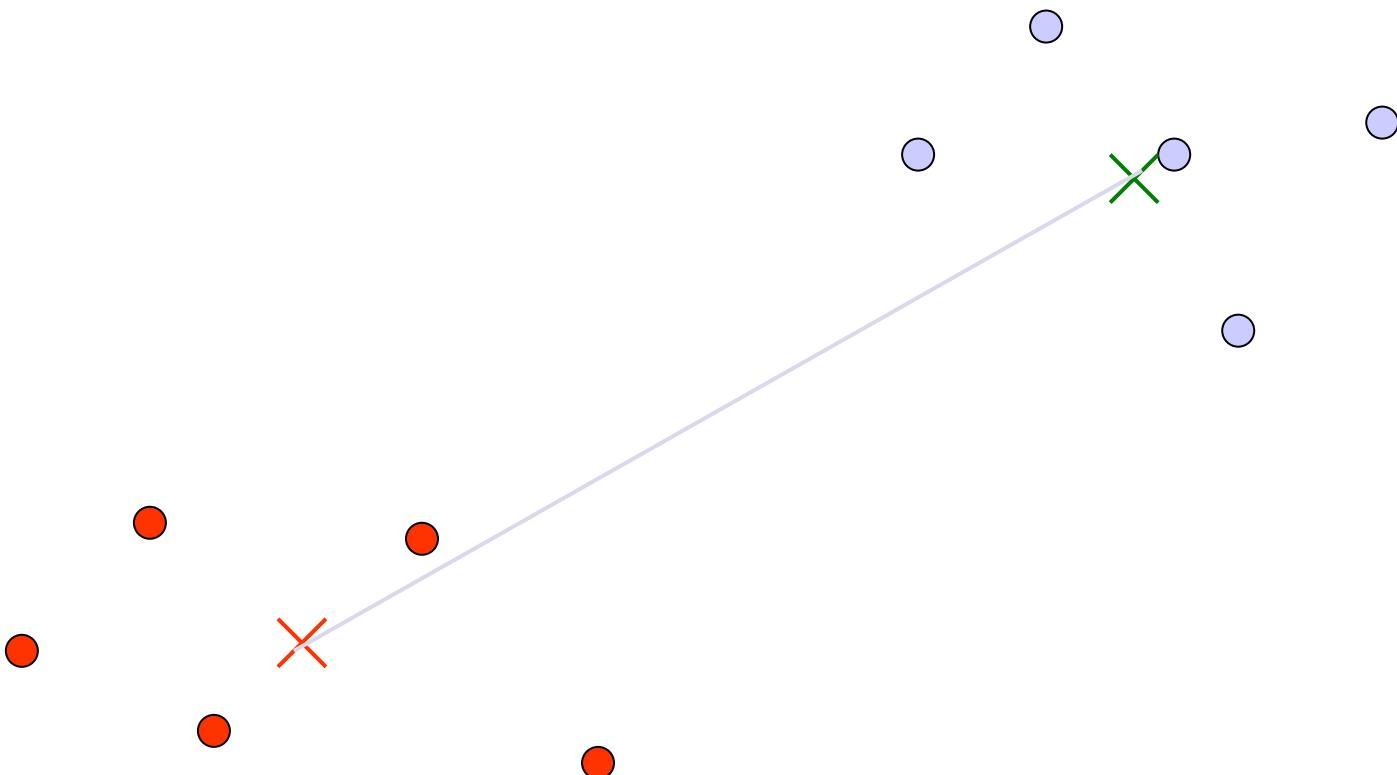
Ventajas: Funciona bien tanto para clusters alongados como esféricos.

Desventajas: Muy costoso computacionalmente.



Centroids: distancia entre los centros de los clusters.

Desventajas: Actualizaciones en los datos pueden provocar la creación de jerarquías completamente diferentes.



Efecto de los métodos de agregación:

304 STATISTICAL DECISION METHODS. SOME BACKGROUND

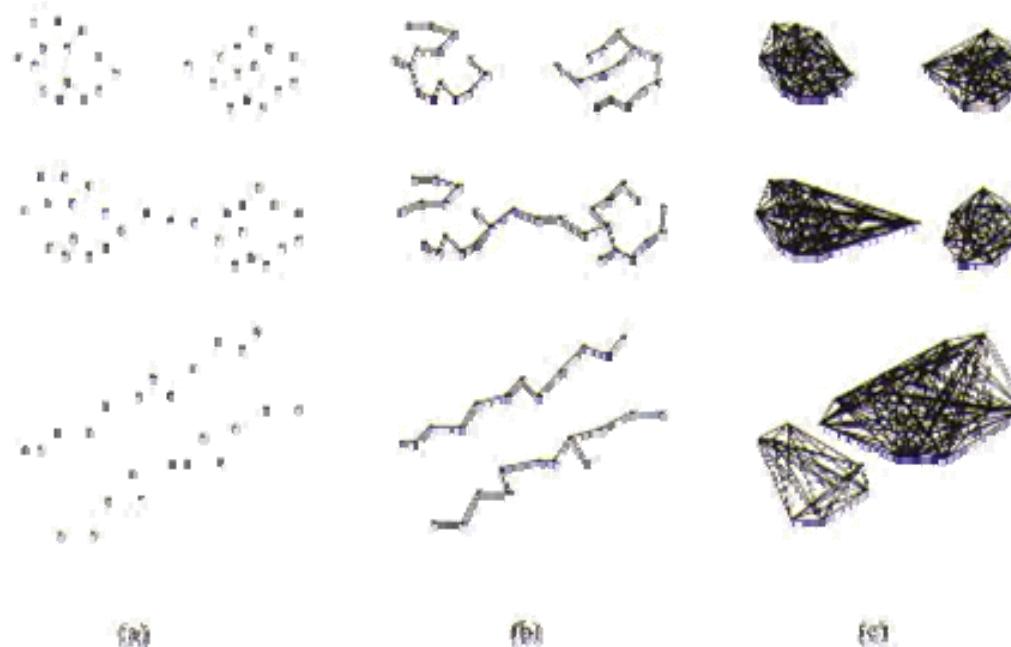
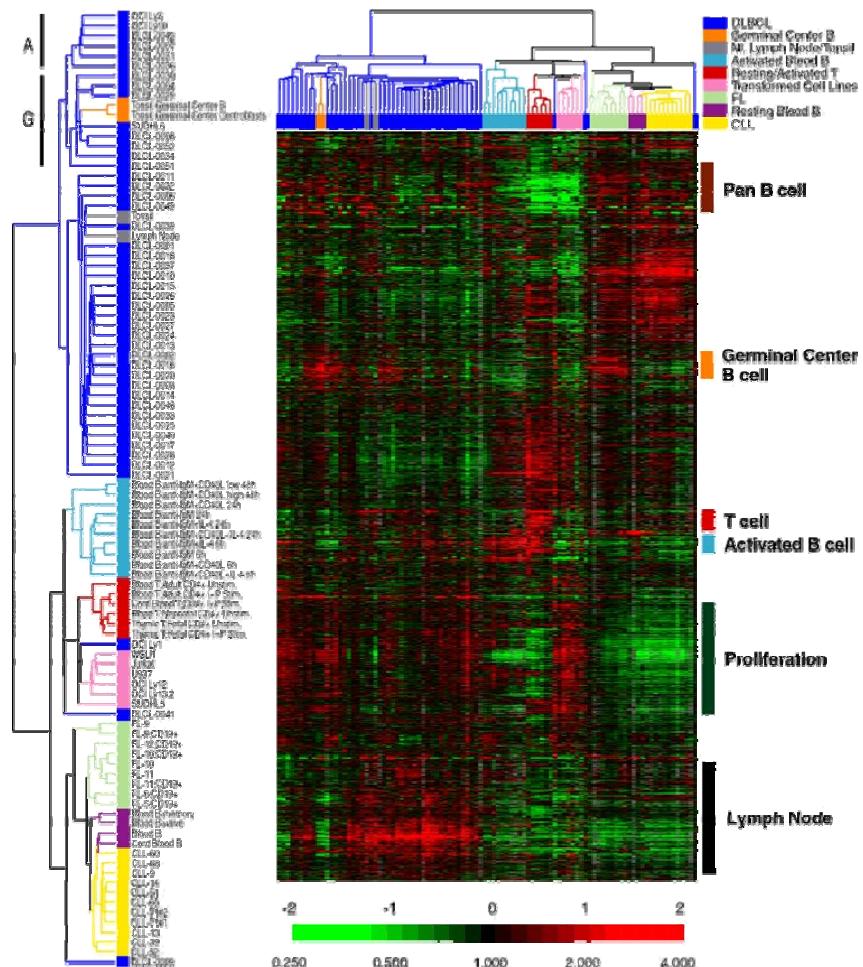


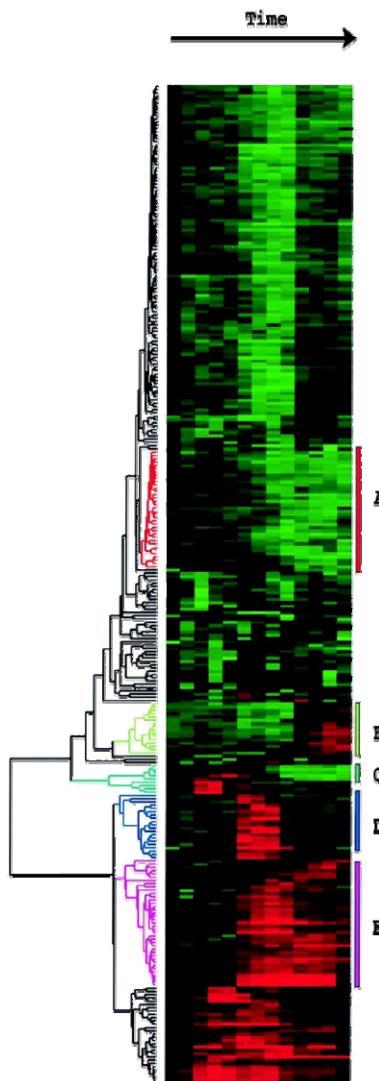
FIG. 6.19. Graphical examples of hierarchical merging. (a) Three data sets. (b) Results of single-link method. (c) Results of complete-link method. (Reproduced with permission from [Duda73]; copyright 1973 by John Wiley & Sons.)

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., et.al. (2000)
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.
Nature 403: 503-11.



Depicted are the ~1.8 million measurements of gene expression made on 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of relatedness of gene expression in each samples. The dendrogram is color coded based on the category of mRNA sample studied (see upper right key). Each row represents a separate cDNA clone on the microarray and each column a separate mRNA sample. The scale extends from fluorescence ratios of 0.25 to 4 (-2 to +2 in log base 2 units). Grey indicates missing or excluded data.

Eisen, M., Spellman, P.T., Botstein, D. & Brown, P.O. (1998). Cluster Analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14867.



Single time course data of a canonical model of the growth response in human cells: clustered data from serum simulation of primary human fibroblasts. Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hours. Serum was added back and samples taken at time , 0, 15 min, 30 min, 1h, 2h, 3h, 4h, 8h, 12h, 16h, 20h and 24h. Five clusters were identify containing known genes involved in:

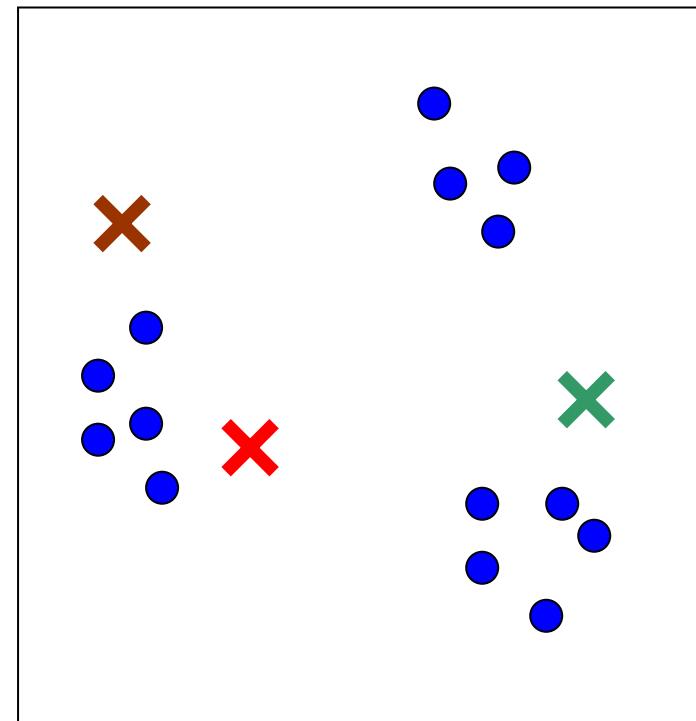
- (A) cholesterol biosynthesis
- (B) The cell cycle
- (C) The immediate-early response
- (D) Signaling and Angiogenesis
- (E) Wound healing and tissue remodeling.

Agglomerative Hierarchical Clustering:

- Los resultados dependen mucho del método de distancia escogido:
 - Single Linkage: clusters alongados.
 - Complete Linkage: clusters esféricos.
- Proceso iterativo muy costoso.
- Fueron inicialmente diseñados para trabajar con datos que poseen estructura jerárquica, lo cual no garantiza que funcionen bien para todo tipo de datos.
- La naturaleza determinista del método y la imposibilidad de la reevaluación del clustering a posteriori puede causar que la agrupación sea basada en decisiones locales más que en globales.
- NO es robusto en presencia de ruido.
- La decisión del número de clusters es subjetiva.

Clustering particional: K-means

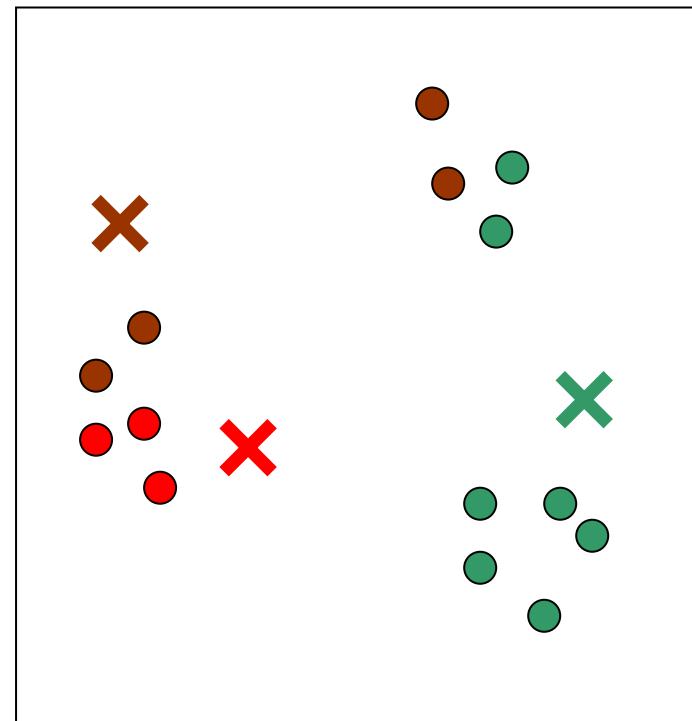
- *Comenzar con una posición aleatoria para los centros de los clusters.*
- Iterar hasta que los centroides se estabilicen.
 - Asignar puntos a los centroides.
 - Mover los centroides hacia los “centros” de los puntos asignados.



Iteración = 0

Clustering particional: K-means

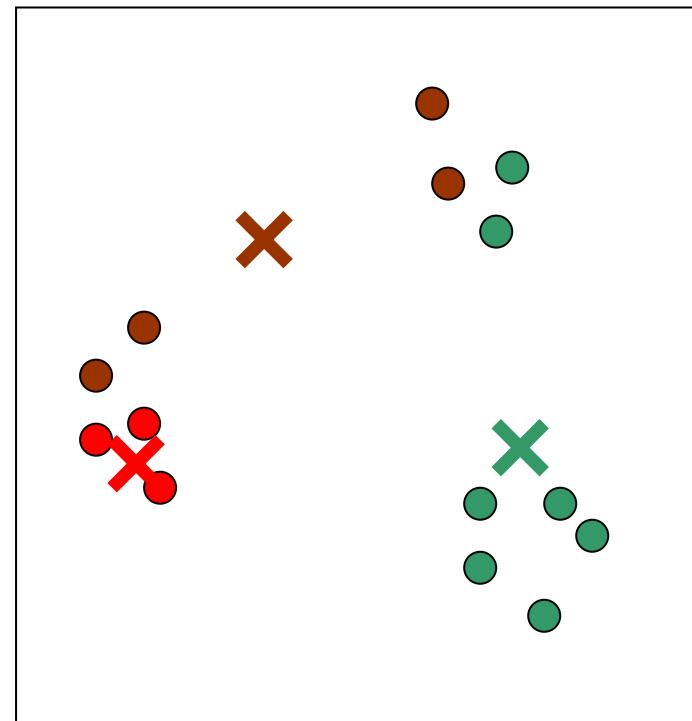
- Comenzar con una posición aleatoria para los centros de los clusters.
- Iterar hasta que los centroides se estabilicen.
 - *Asignar puntos a los centroides.*
 - Mover los centroides hacia los “centros” de los puntos asignados.



Iteración = 1

Clustering particional: K-means

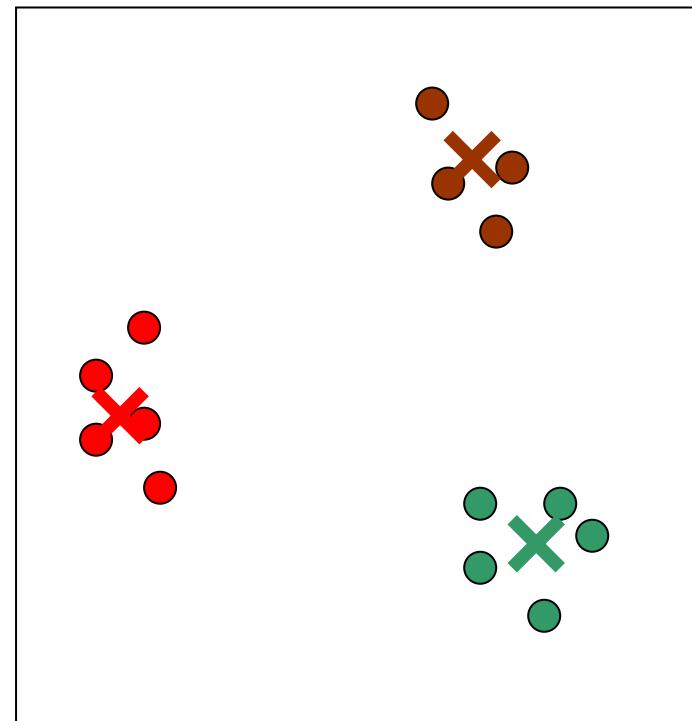
- Comenzar con una posición aleatoria para los centros de los clusters.
- Iterar hasta que los centroides se estabilicen.
 - Asignar puntos a los centroides.
 - *Mover los centroides hacia los “centros“ de los puntos asignados.*



Iteración = 1

Clustering particional: K-means

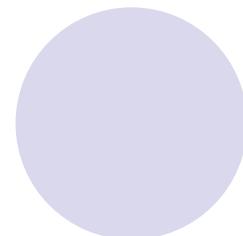
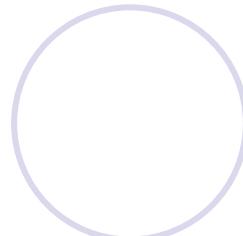
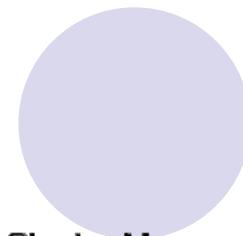
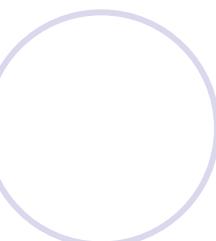
- Comenzar con una posición aleatoria para los centros de los clusters.
- ***Iterar hasta que los centroides se estabilicen.***
 - Asignar puntos a los centroides.
 - Mover los centroides hacia los “centros” de los puntos asignados.



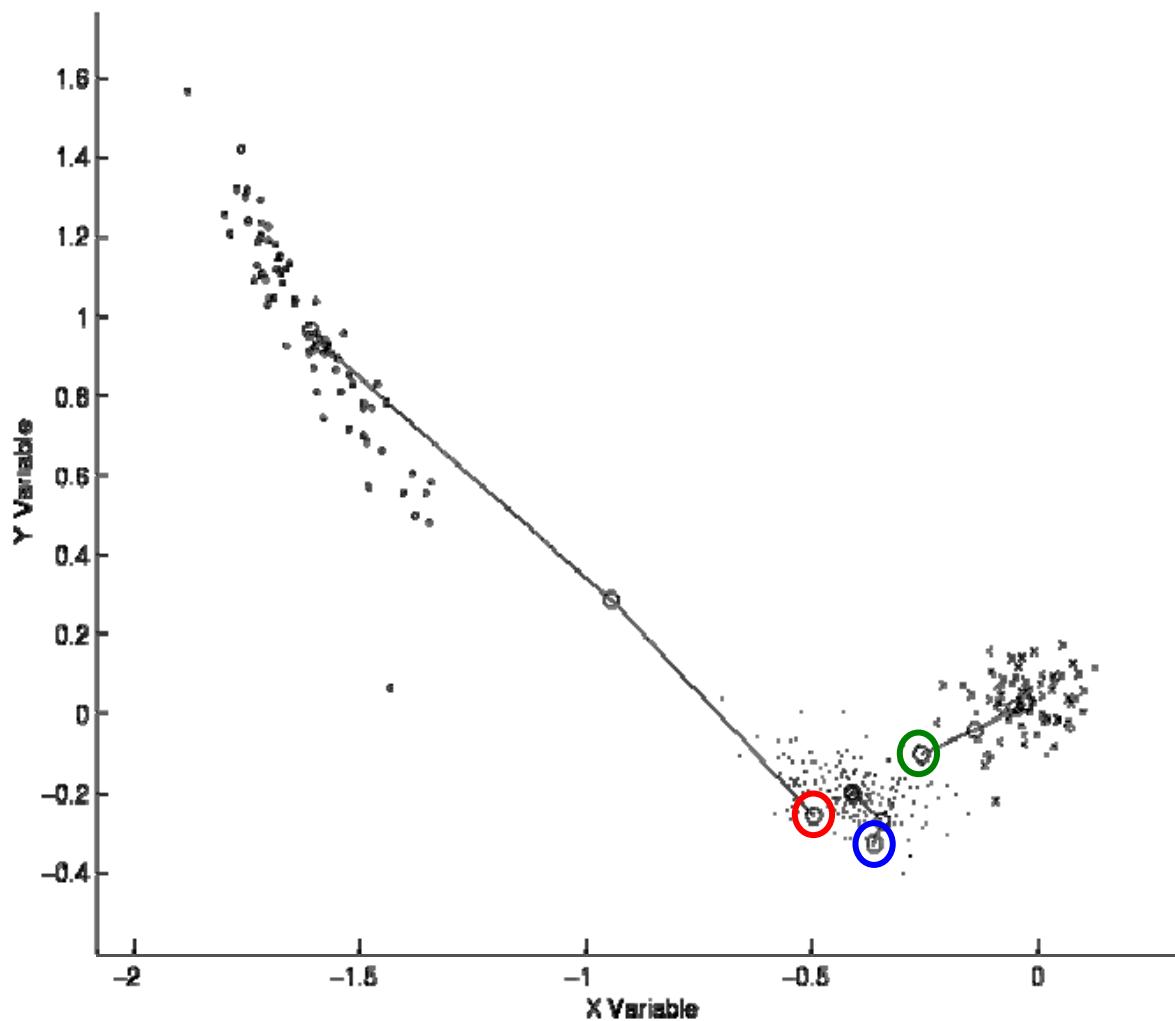
Iteración = 2

Ejemplo:

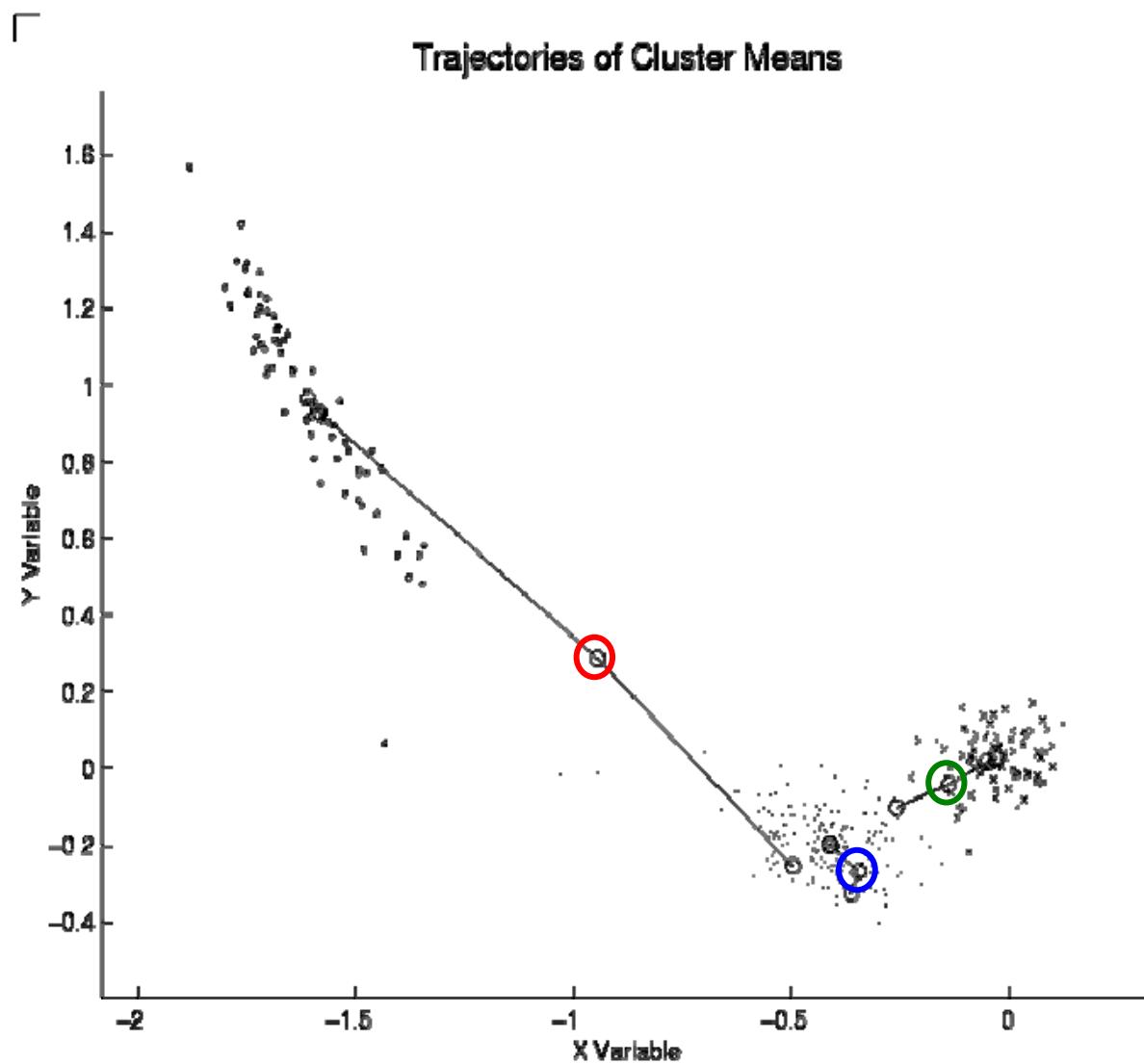
Γ



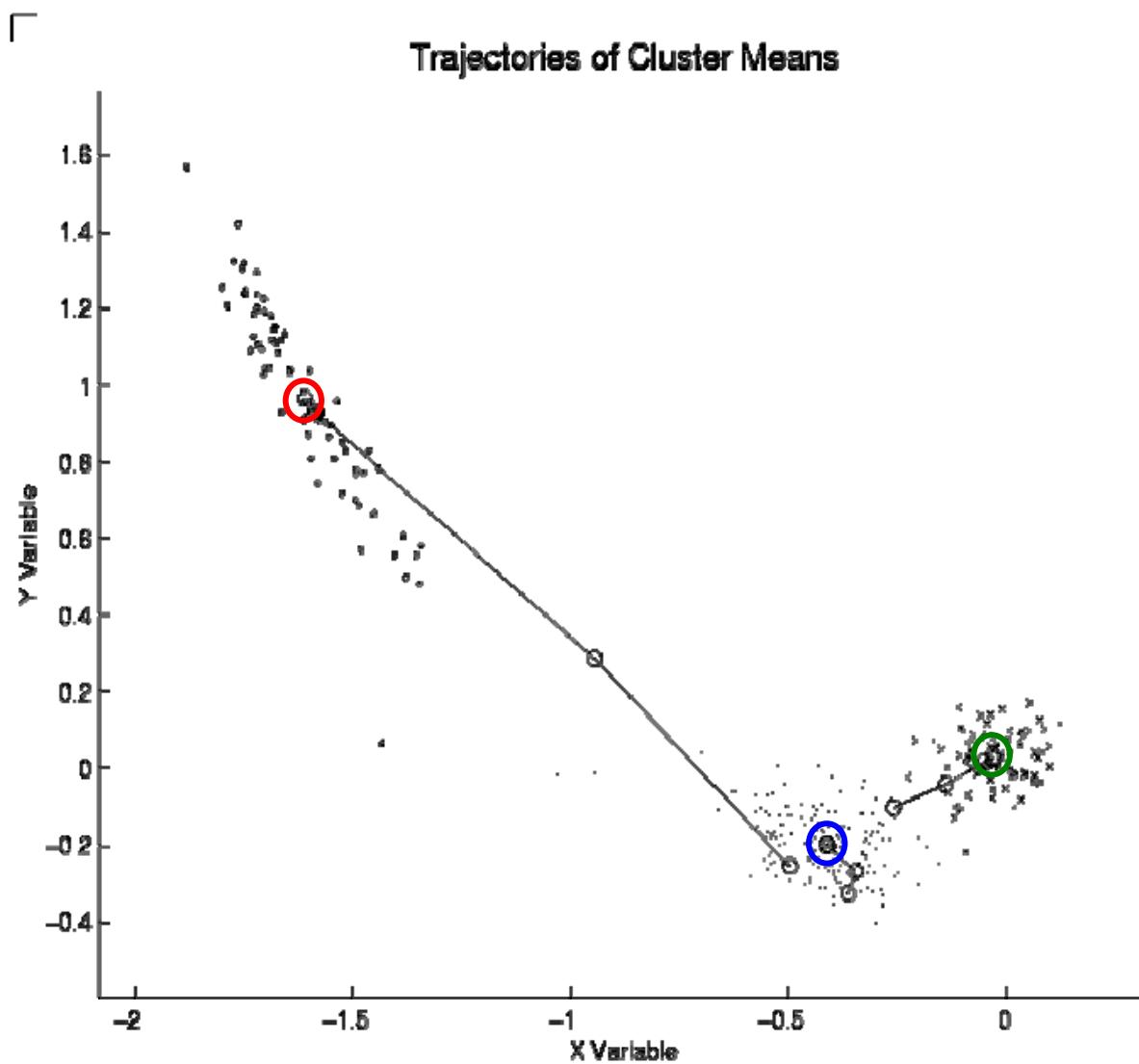
Trajectories of Cluster Means



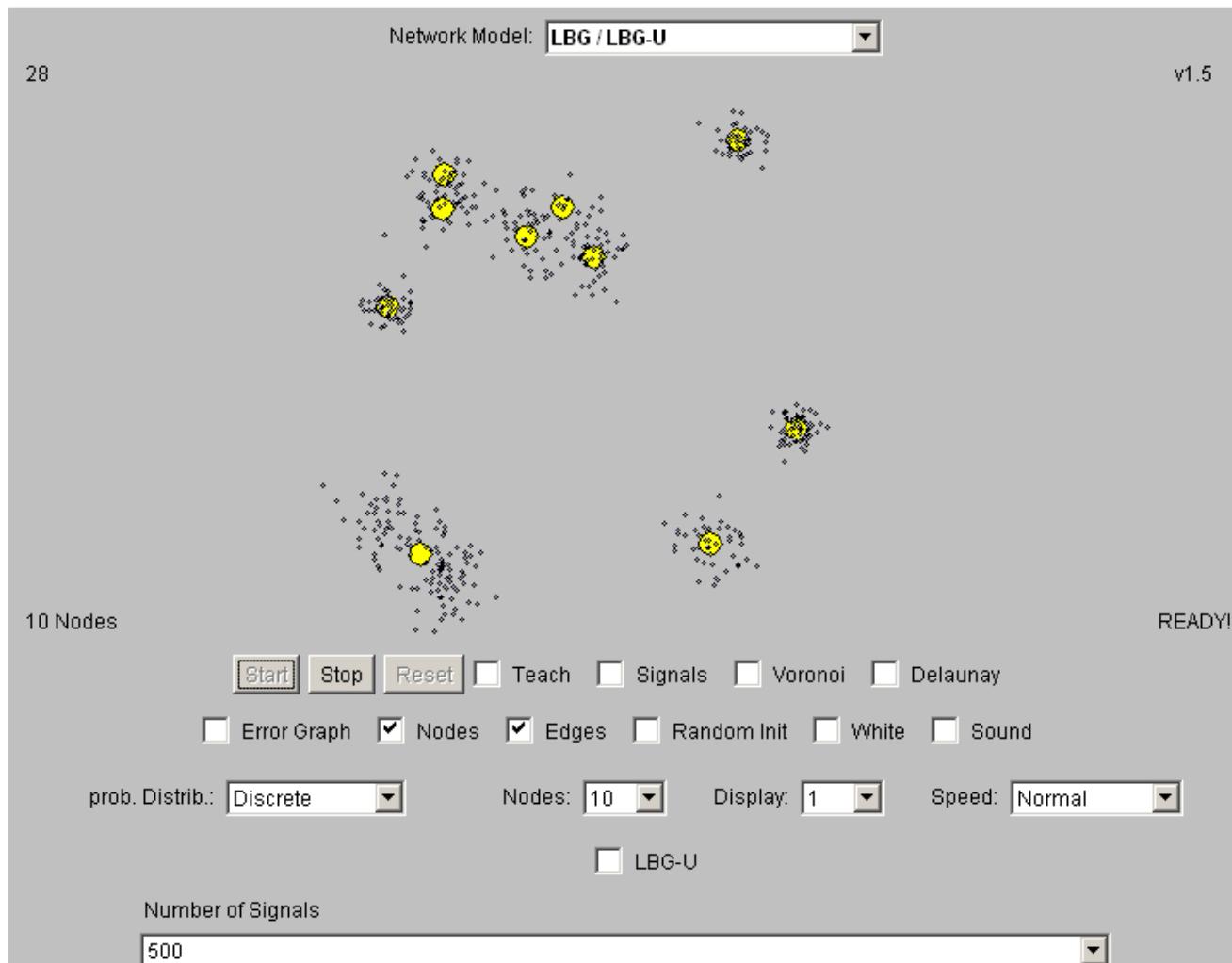
Ejemplo:



Ejemplo:



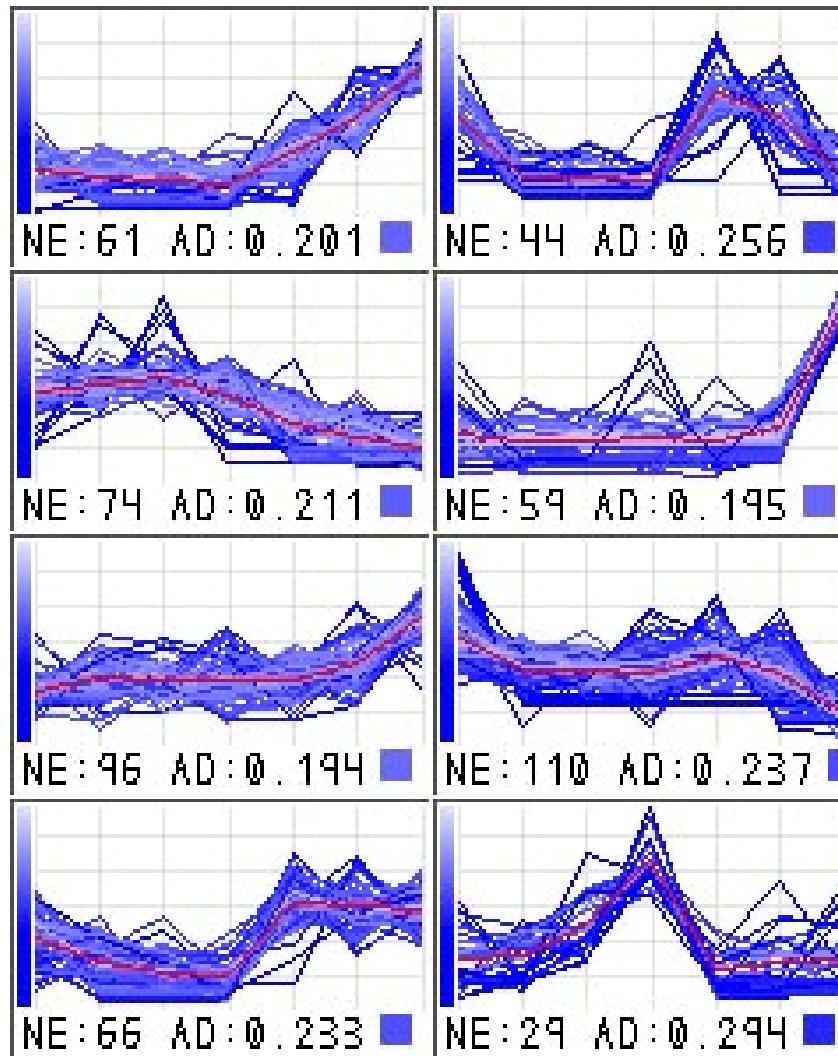
Demo de K-means:



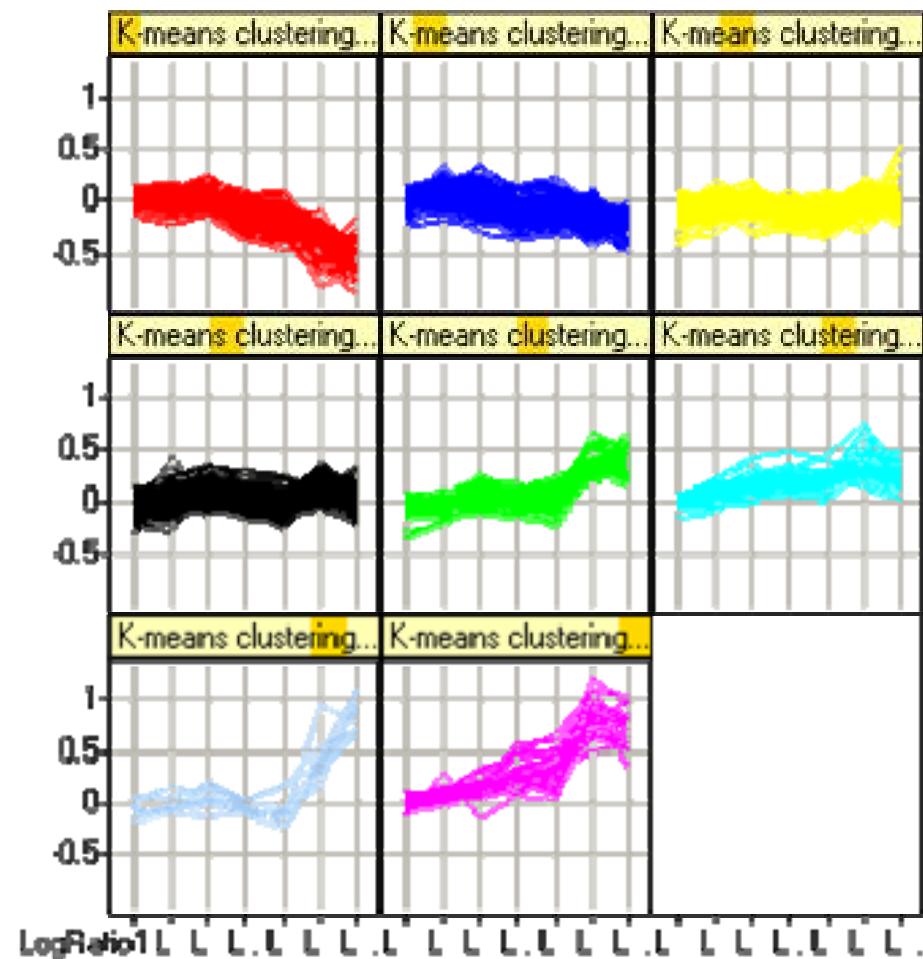
Clustering particional: K-means

- Los resultados dependen de las posiciones **iniciales** de los centroides.
- Algoritmo **rápido**: solo calcula las distancias de los puntos de datos a los centroides.
- El número de clusters hay que decidirlo de antemano (**gran desventaja!**)

Ejemplo de K-Means:



K-Means Clustering



8 Grupos obtenidos utilizando k-means con distancia de correlación.

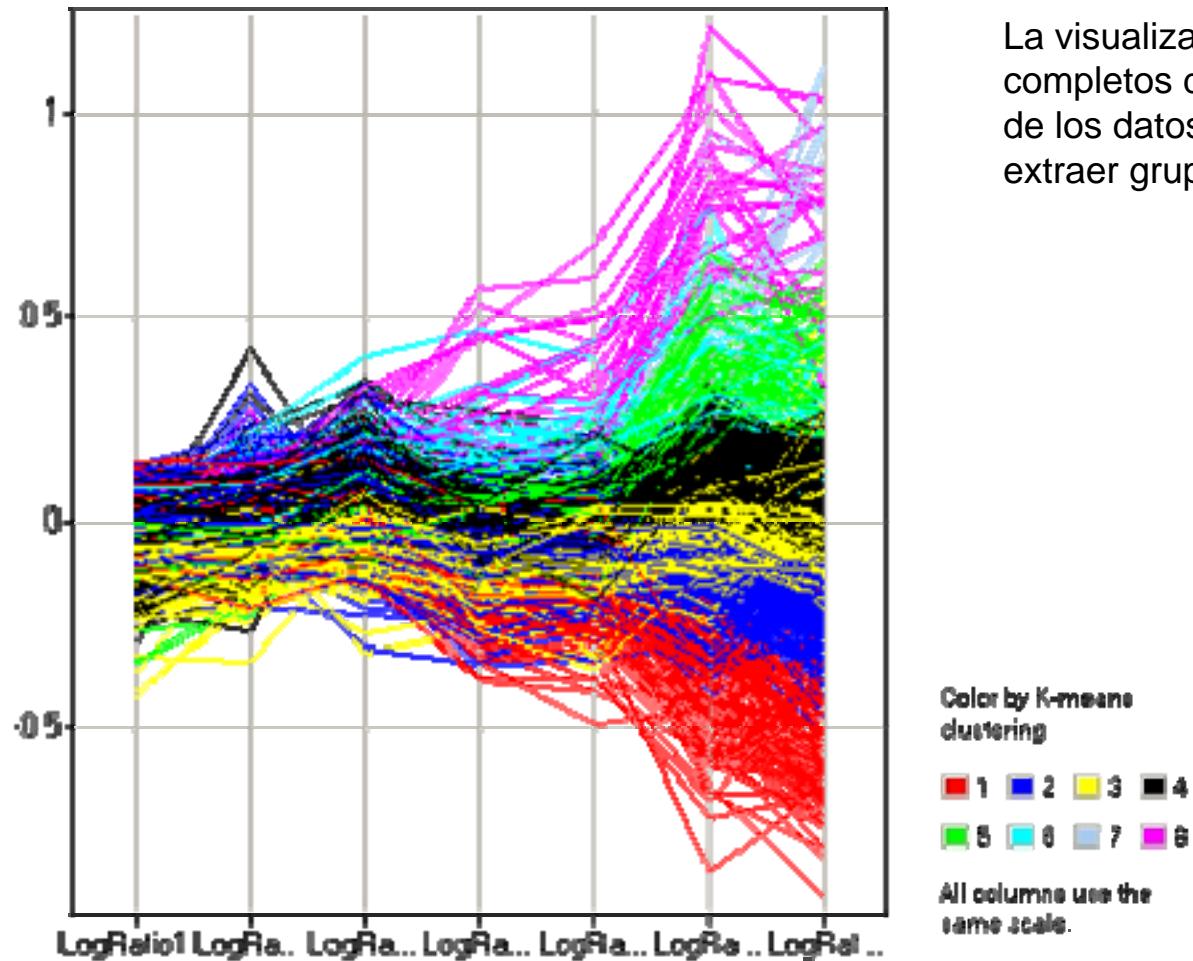
Color by K-means clustering

■ 1 ■ 2 ■ 3 ■ 4

■ 5 ■ 6 ■ 7 ■ 8

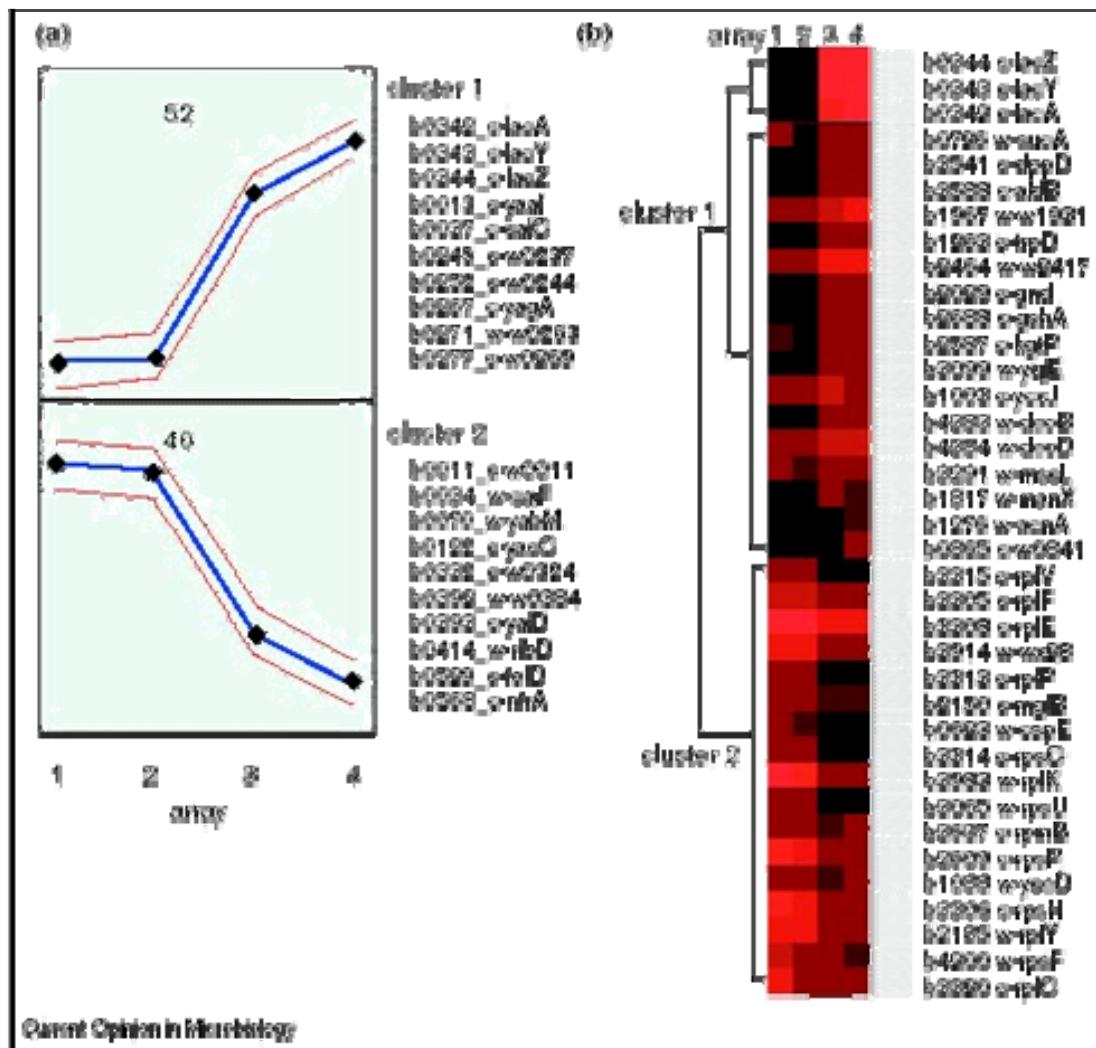
All columns use the same scale.

Visualización de perfiles



La visualización de perfiles completos da una idea de la forma de los datos, pero es muy difícil extraer grupos de manera visual.

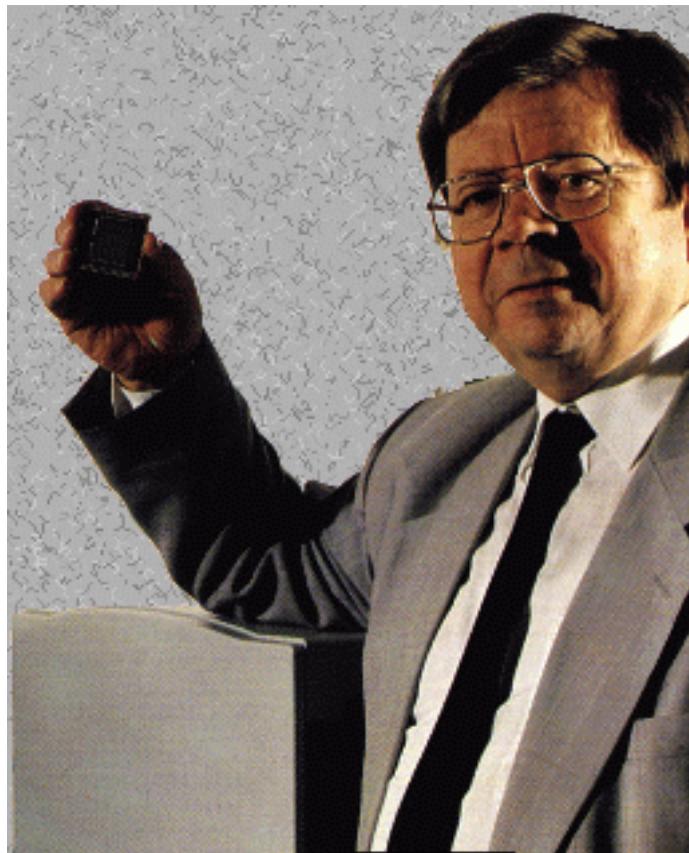
K-Means vs. HCA:



Redes Neuronales: Self-Organizing Maps (SOM)

- Es un modelo de red neuronal que simula la auto organización de las neuronas que se produce en el cortex del cerebro humano cuando le es presentado un estímulo.
- Tiene la capacidad de crear un conjunto mucho menos de datos que son fieles “representantes” de los datos originales.
- Conocidos como Mapas Auto-organizativos de Kohonen

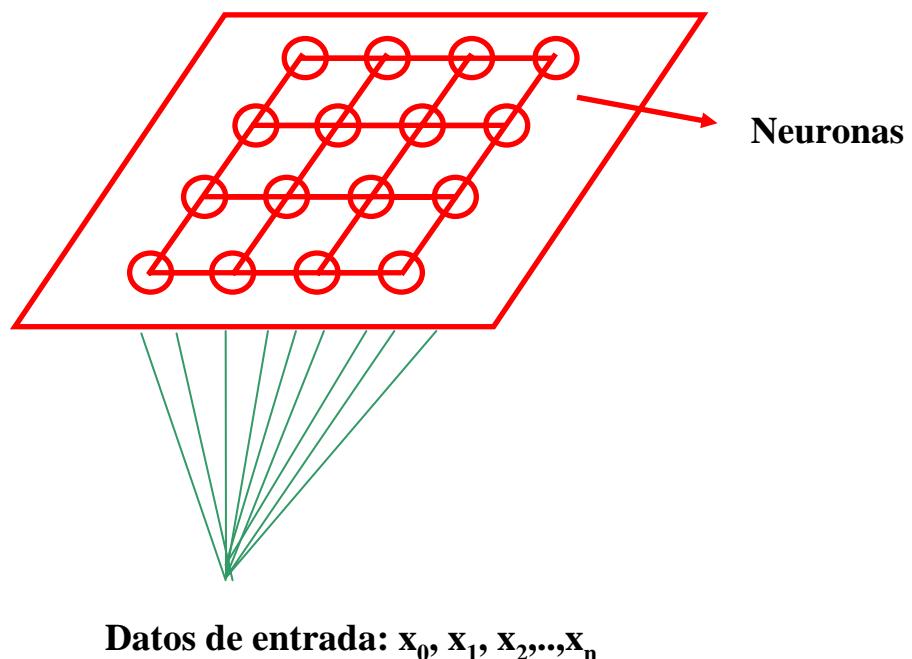
Teuvo Kohonen



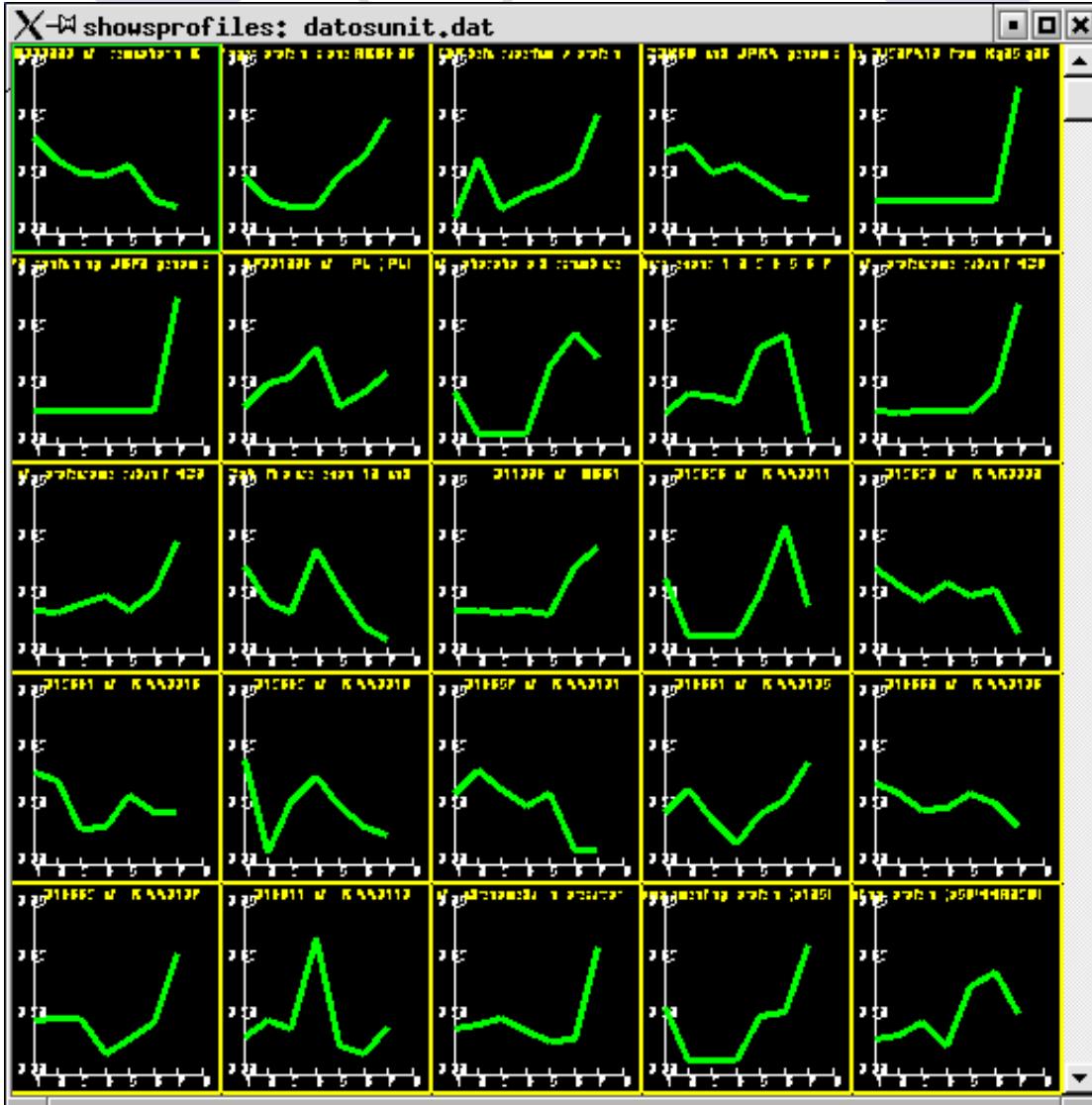
***Dr. Eng., Emeritus Professor of the Academy of Finland;
Academician***

Estructura del SOM:

Self-Organizing Map



DNA Arrays. Datos originales:

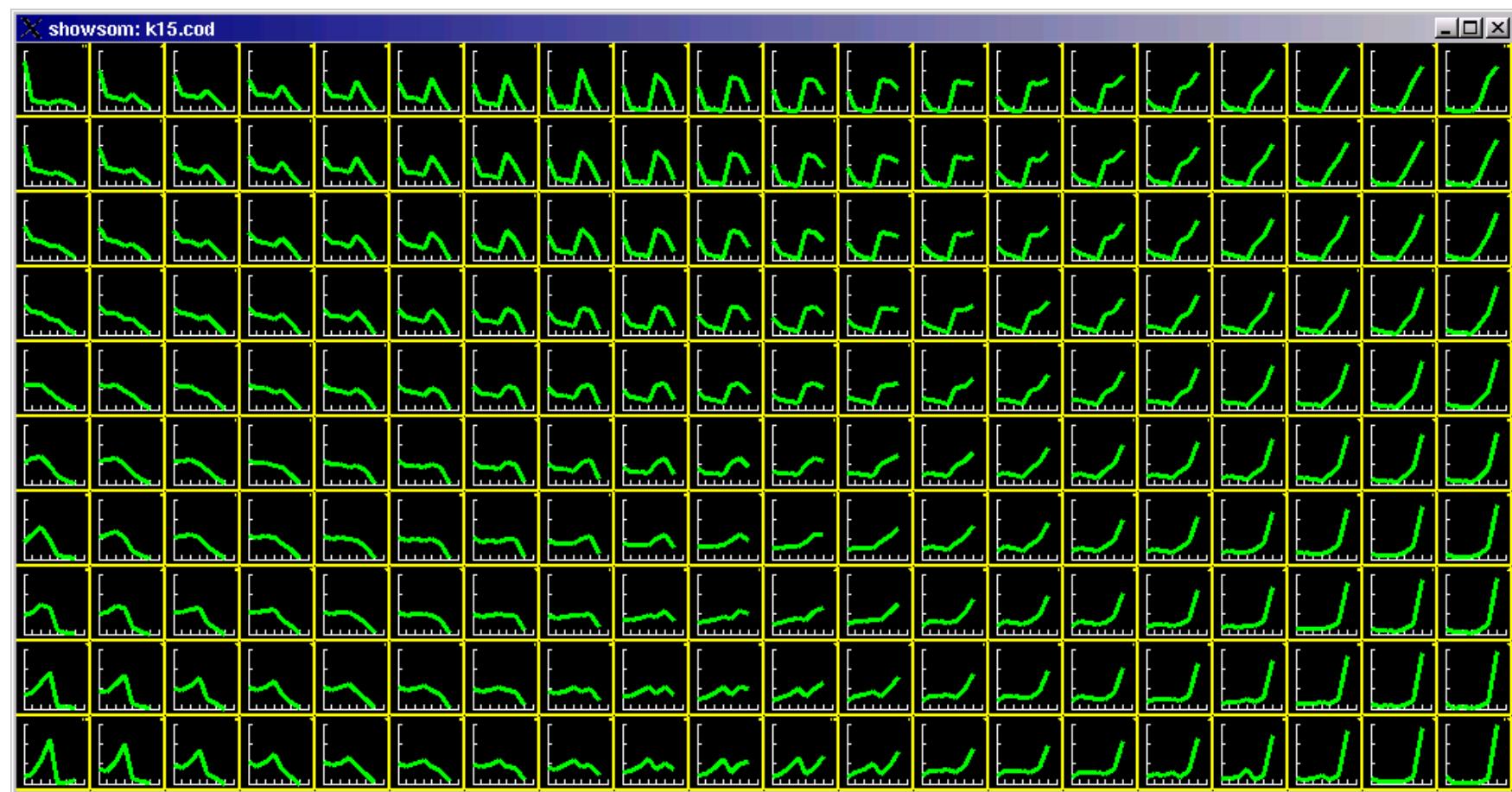


Gene expression behaviour of ultraviolet response on the skin (keratinocytes cells)

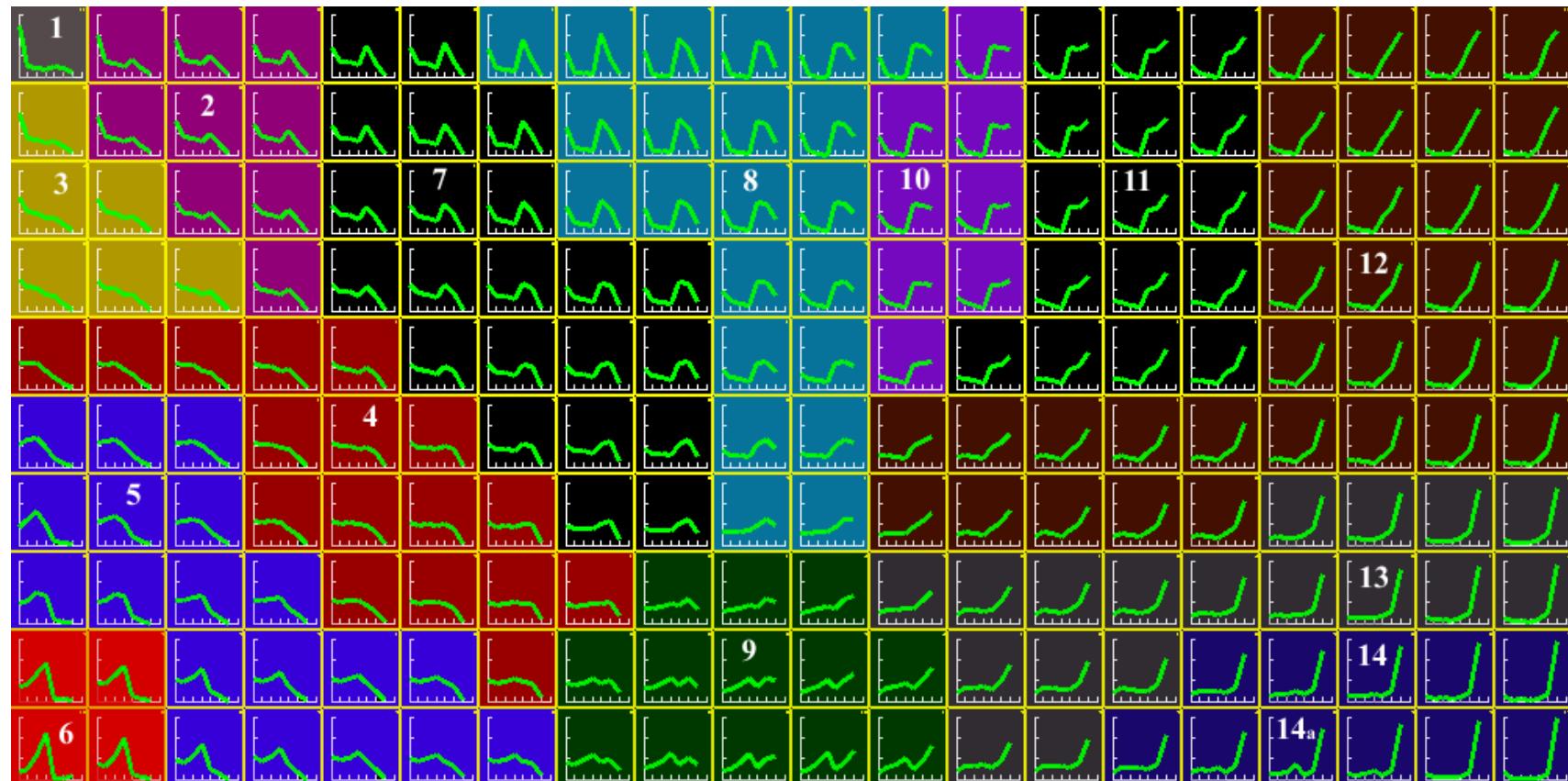
Experimental points:

- 1 Control
 - 2 10 mJ/cm² 4h
 - 3 20 mJ/cm² 4h
 - 4 40 mJ/cm² 4h,
 - 5 10 mJ/cm² 24h,
 - 6 20 mJ/cm² 24h
 - 7 40 mJ/cm² 24h.

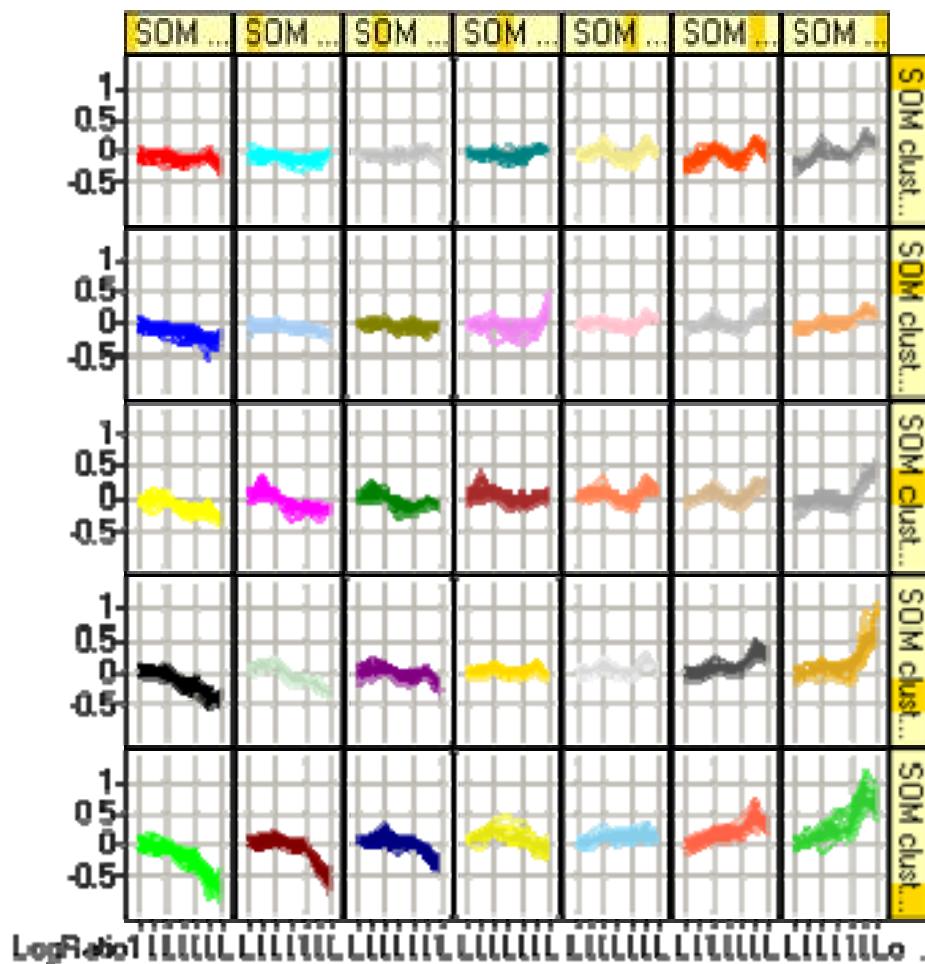
SOM: 20x10 neuronas



SOM agrupado:



Self-Organizing Maps



Generated by Self-Organizing Maps

Grid size (width x height): 7 x 5

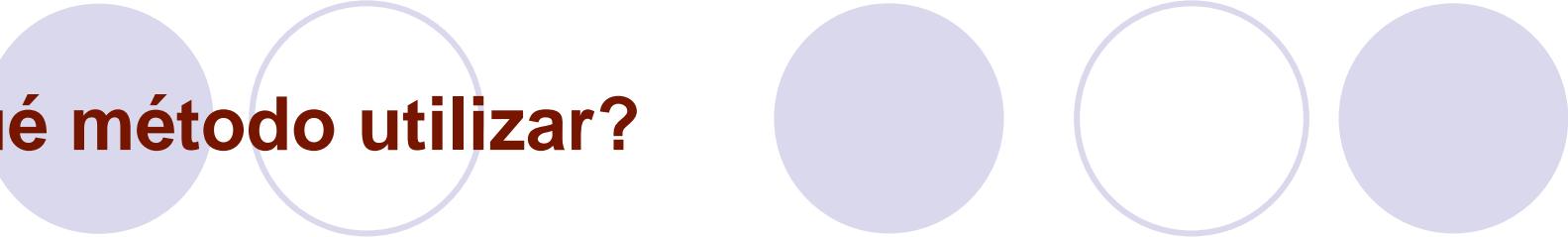
Neighborhood function: Bubble
Radius (begin x end): 2.5 x 0

Learning function: Linear
Initial rate: 0.05

Number of training steps: 12500

Output parameters:
MappingPrecision: 6.823e-3
TopologyPreservation: 0.1

Color by SOM clustering		
1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
All columns use the same scale.		



¿Qué método utilizar?

- Desgraciadamente **no existe un consenso** sobre cual es el mejor método a utilizar para hacer agrupamiento.
- No existe una manera simple de decidir cual es el mejor a partir de un conjunto de datos experimentales.
- Recomendación; Usar toda la información disponible y **utilizar varios métodos de agrupamiento y comparar!**

Paso final: Búsqueda de funciones de los genes

- Una vez creados los clusters, el paso final sería la búsqueda de las funciones de los genes que pertenecen a cada uno de los clusters.
- Generalmente los chips están compuestos por ESTs, lo que hace el proceso de búsqueda más largo.
- Bases de datos frecuentemente utilizadas para esto: GenBank, UniGen, OMIM, GeneCards, SwissProt, etc

Limitaciones:

Aunque los DNA microarrays son relativamente fácil de utilizar, existen ciertas limitaciones relacionadas con la información que estas técnicas brindan:

- El análisis del DNA **no puede predecir si las proteínas están en un estado activo.**
- A pesar de la correlación existente entre la cantidad de mRNA producido en la célula y la cantidad de proteína sintetizada, su cuantificación no es directa, por lo que **la cuantificación del RNA no siempre refleja los niveles correspondientes de proteínas.**
- **Múltiples proteínas pueden ser obtenidas de un mismo gen** cuando se tienen en cuenta la postraducción y el mRNA splicing.

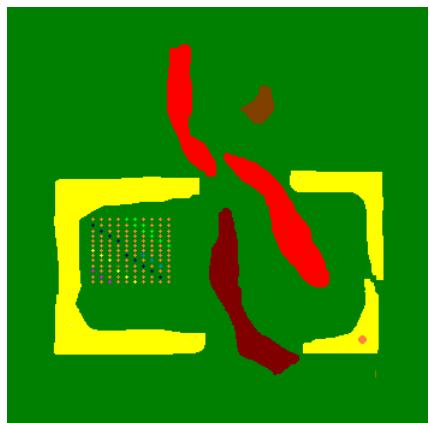
Por lo tanto la técnicas de microarray solo permiten una estimación cualitativa del proteoma. Técnicas mas avanzadas se necesitan para el estudio del proteoma: **La proteómica.**

Software:



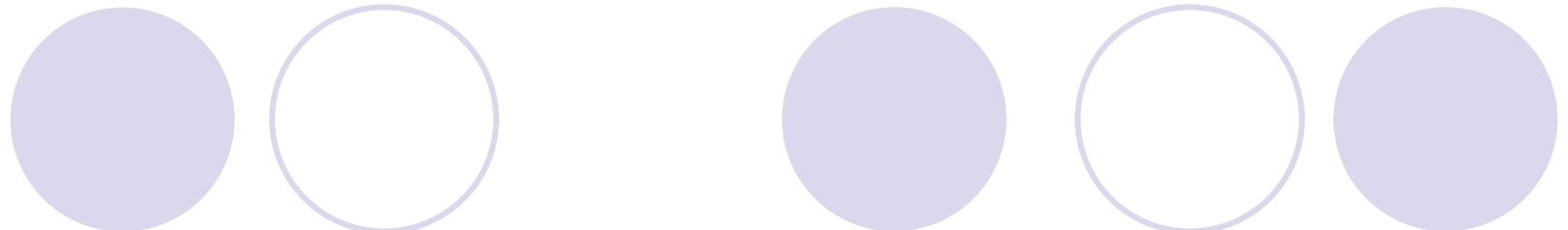
SpotFire DecisionSite for
Functional Genomics

www.spotfire.com



Engene: Gene Expression Data
Processing and Exploratory Data
Analysis

www.biocomp.cnb.uam.es



Gracias por vuestra atención



<http://www.integromics.com>