Universidad Complutense de Madrid
ESCUELA DE VERANO 2007

Introducción al análisis de datos en microarrays

Gonzalo Gómez López
INB-CNIO
ggomez@cnio.es

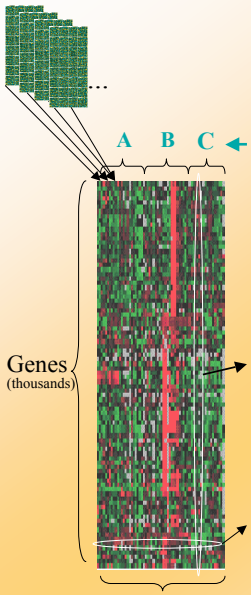
Madrid, 26 de Julio, 2007

Universidad Complutense de Madrid
ESCUELA DE VERANO
2007

1. Introducción
2. Microarrays (tipos, tratamiento de las muestras e hibridación, bases de datos, aplicaciones…)
3. Normalizacion y análisis exploratorio (JC Oliveros)
4. Análisis de datos (Análisis supervisado y no supervisado, algoritmos…)
5. Práctica

ggomez@cnio.es

The data



Source: J. Dopazo

Characteristics of the data:

- Low signal to noise ratio
- High redundancy and intra-gene correlations
- Most of the genes are not informative with respect to the trait we are studying (physiological conditions, etc.)
- Many genes have no annotations!!

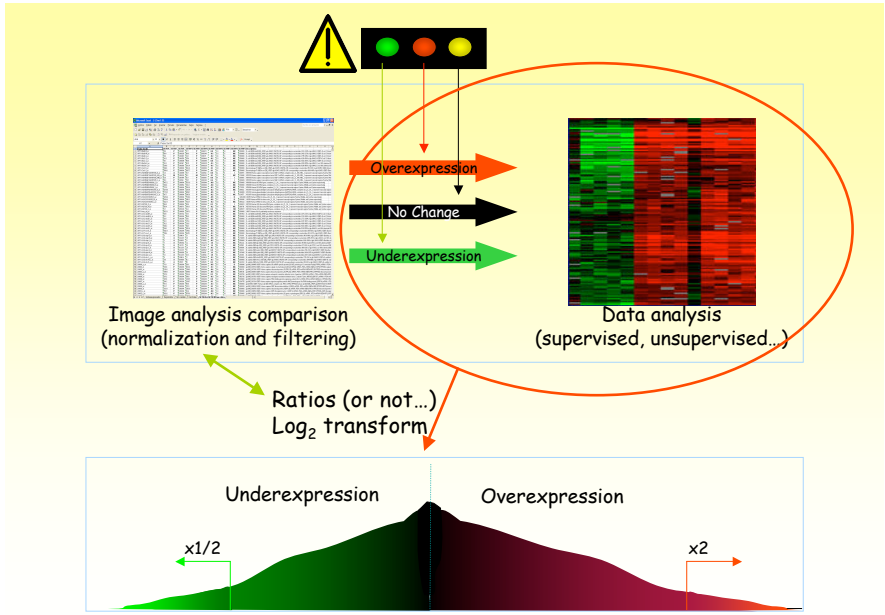
Brief history in microarray data analysis:

Unsupervised methods

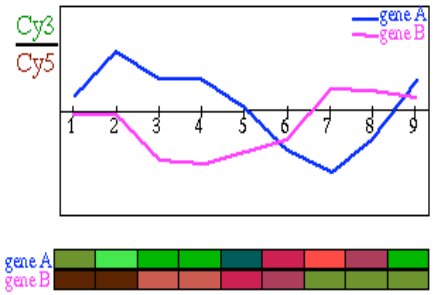
1. **Clustering methods**
   - UPGMA (Sneath and Sokal, 1973)
   - kmeans (Hartigan and Wong 1979)
   - kmedians (Hartigan and Wong 1979)
   - SOTA (Herrero et al. 2001)
   - SOM (Kohonen 1979, Tamayo et al 1999)
   - Gene Shaving (Hastie et al, 2000)
   - Fuzzy methods (**Dougherty ET AL. 2002**)
   - Probabilistic clustering (Bhattacharjee et al. 2001)
   - Metagenes (Pittman et al, 2004)
2. **Exploratory analysis**
   - parametric: ttest, SAM (Tusher et al, 2001), ANOVA…
   - non parametric: Welch ttest, Wilcoxon, Kruskal Wallis…
   - Sumarizing datasets: PCA
3. **Blocks of genes**
   - Kolmogorov-Smirnof: GSEA (Subramanian et al. 2005)
   - FatiScan (Al-Shahrour et al. 2005)
   - SAM 3.0 (Jan 07)
   - GeneTrail (Backes et al. 2007)

Supervised methods (predictors)

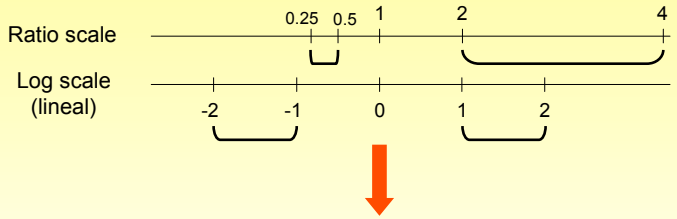
- Neural networks (Khan et al. 2001)
- SOMs (Kohonen et al. 1984)
- kNN (Ripley 1996; Hastie et al 2001)
- PAM (Tibshirani et al. 2002)
- DLDA (Dudoit et al. 2002)
- SVMs (Furey et al, 2000)

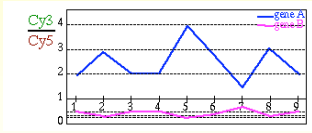


EJEMPLO

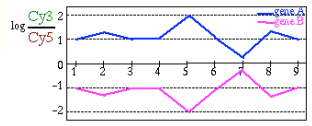


Why Log_2 transform??



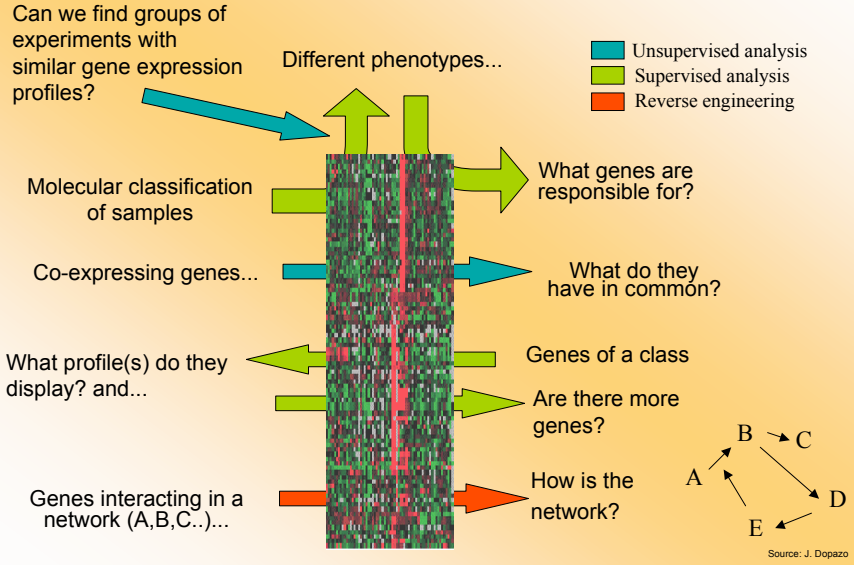


No Log transform

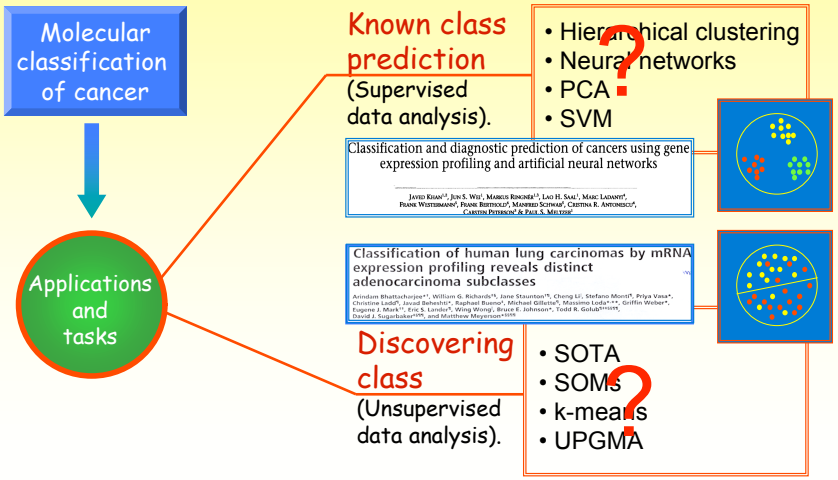


After Log transform

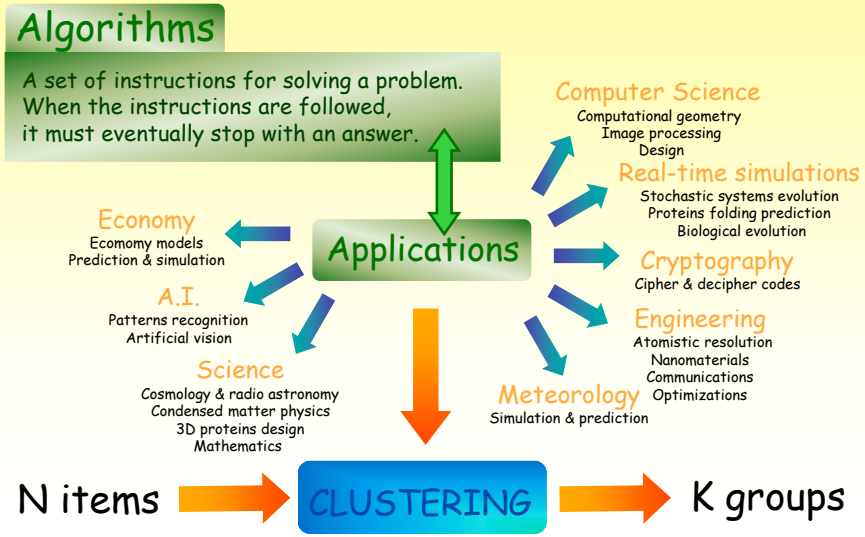
QUESTIONS & METHODOLOGICAL APPROACH



Reminding…Examples



Algorithms & clustering



Brief history in microarray data analysis:

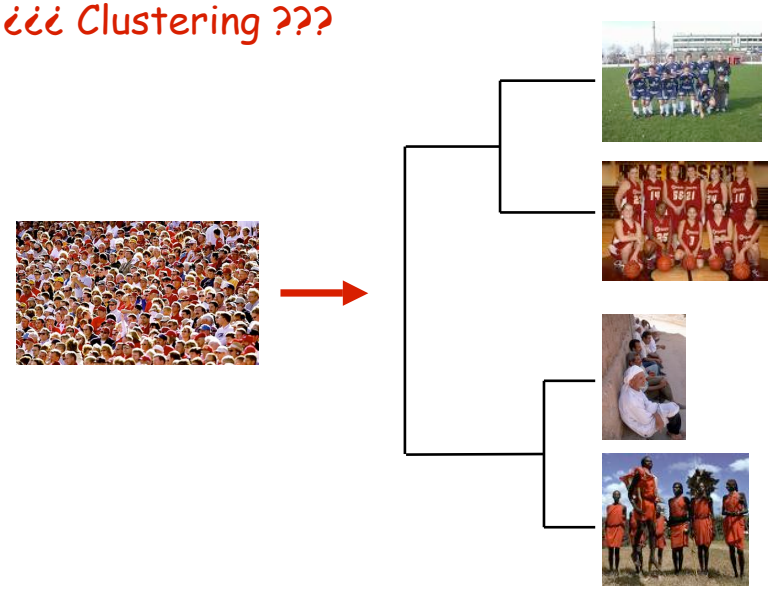
Unsupervised methods

1. **Clustering methods**
   - UPGMA (Sneath and Sokal, 1973)
   - kmeans (Hartigan and Wong 1979)
   - kmedians (Hartigan and Wong 1979)
   - SOTA (Herrero et al. 2001)
   - SOM (Kohonen 1979, Tamayo et al 1999)
   - Gene Shaving (Hastie et al, 2000)
   - Fuzzy methods (**Dougherty ET AL. 2002**)
   - Probabilistic clustering (Bhattacharjee et al. 2001)
   - Metagenes (Pittman et al, 2004)
2. **Exploratory analysis**
   - parametric: ttest, SAM (Tusher et al, 2001), ANOVA…
   - non parametric: Welch ttest, Wilcoxon, Kruskal Wallis…
   - Sumarizing datasets: PCA
3. **Blocks of genes**
   - Kolmogorov-Smirnof: GSEA (Subramanian et al. 2005)
   - FatiScan (Al-Shahrour et al. 2005)
   - SAM 3.0 (Jan 07)
   - GeneTrail (Backes et al. 2007)

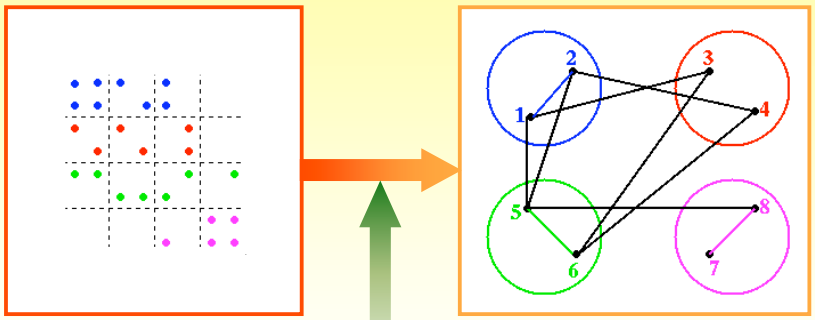
Supervised methods (predictors)

- Neural networks (Khan et al. 2001)
- SOMs (Kohonen et al. 1984)
- kNN (Ripley 1996; Hastie et al 2001)
- PAM (Tibshirani et al. 2002)
- DLDA (Dudoit et al. 2002)
- SVMs (Furey et al, 2000)

¿¿¿ Clustering ???



Algorithms & clustering

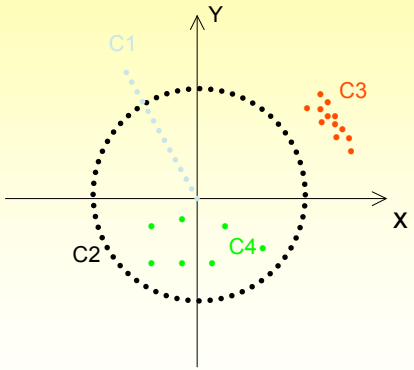


26 items → Clustering algorithm → 4 groups

Algorithms & clustering

- Different methods
  - Hierarchical clustering algorithms
    - Agglomerative
      - UPGMA
    - Divisive
      - SOTA
  - Non-hierarchical clustering algorithms
    - K-means
    - SOMs

Why different algorithms?



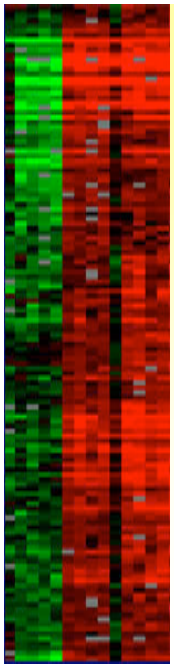
I) Cartesian algorithm (x, y):

a) (X>0,Y>0) ⇒ C3, 1/4 C2?
b) (X<0,Y>0) ⇒ C1, 1/4 C2?
c) (X<0,Y<0) ⇒ C4?, 1/4 C2?
d) (X>0,Y<0) ⇒ C4?, 1/4 C2?
…
aUbUcUd ⇒ C1, C2, C3, C4.

II) Polar algorithm (α, m):

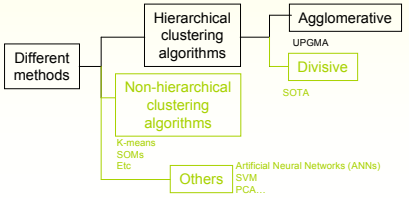
e) (α=cte, m) ⇒ C1…
f) (α, m=cte) ⇒ C2, C3?
…
g) (α, m) ⇒ C1, C2, C3, C4.

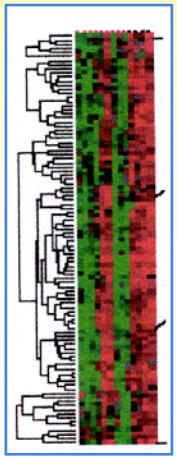
There´s no a perfect method…



Data analysis in microarrays

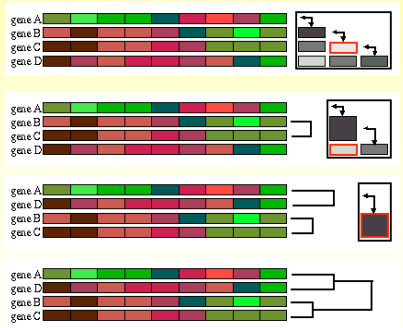
Hierarchical clustering & dendrograms





UPGMA

Unweighted Pair Group Method with Arithmetic mean



1) B y C son los dos genes más próximos (en sus medias)

2) B y C se unen y se recalcula la matriz de distancias empleando el nuevo cluster en vez de B y C. A y D son ahora los mas próximos.

3) A y D tambien se unen. Se reordenan los elementos para ajustar la topología del árbol. Se recalcula la matriz de distancias entre los grupos existentes

4) El proceso termina cuando el dendrograma esta construído.

Distances between genes??

Metrics

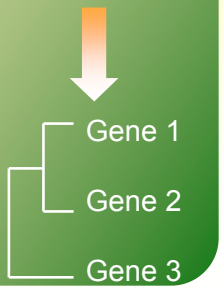
Euclidean distance
Pearson correlation coefficient
Spearman ρ correlation coefficient
...

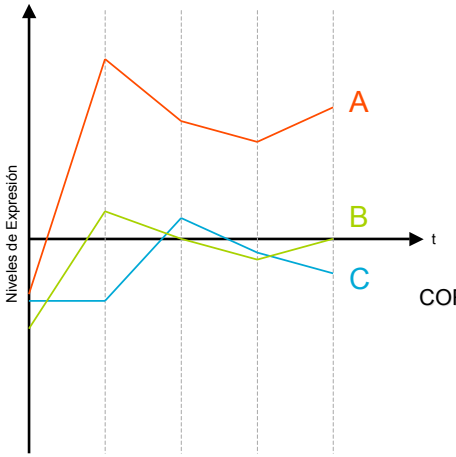
	RNA express Gene 1	RNA express Gene 2	RNA express Gene 3
Exp 1	0.7	0.3	7.3
Exp 2	1.2	1.9	6.5
Exp 3	1.1	0.9	8.9

Example

Pearson Correlation Coefficient

Gene 1			
Gene 2	0.88		
Gene 3	-0.19	-0.62	





DISTANCIA

EUCLIDEANA

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

CITY BLOCK (Manhattan or taxi cab)

$$|x_1-x_2|+|y_1-y_2|\,.$$

OTRAS (Canberra, Bray-Curtis…)

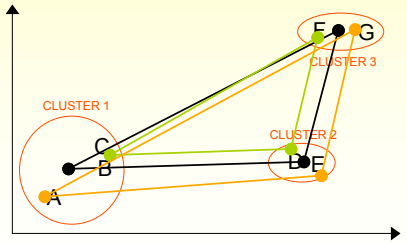
CORRELACIÓN

PEARSON $r=\frac{\sum z_x z_y}{N-1}$

SPEARMAN $\rho=1-\frac{6\sum D^2}{N(N^2-1)}$

Dendrograms

Linkage methods



Average-linkage method

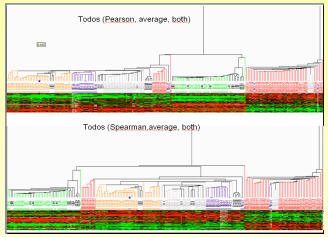
Single-linkage method

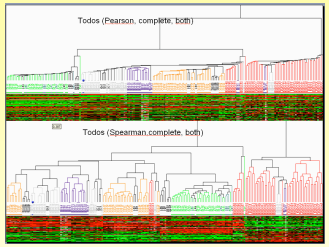
Complete-linkage method

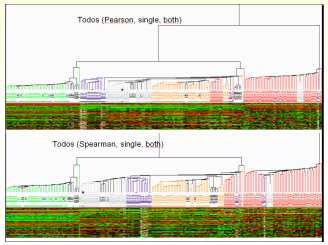
Others…

(Weighted pair-group average, Within-groups clustering Ward´s method…)

Ejemplos

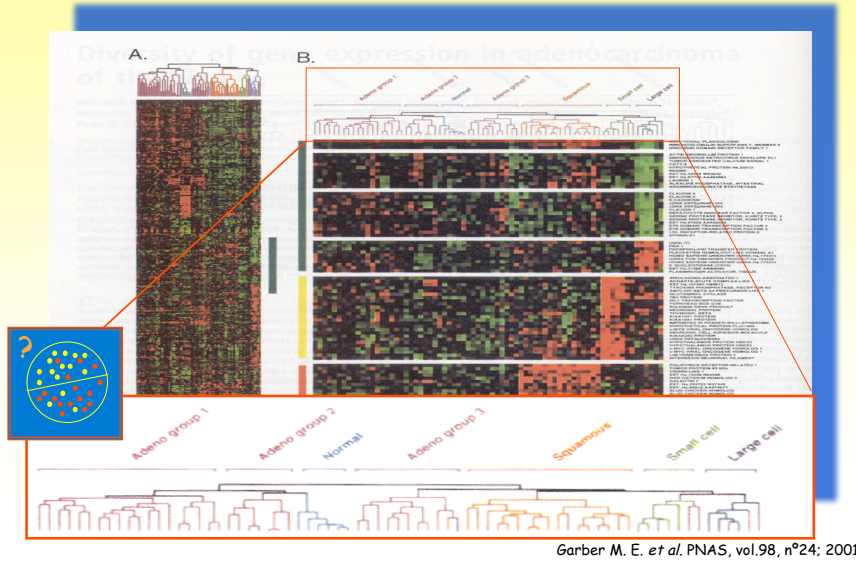




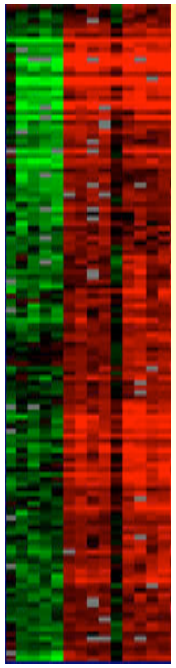


En nuestra práctica:
GEPAS

Hierarchical clustering - UPGMA

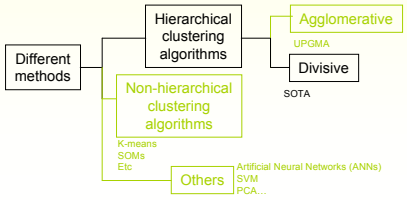


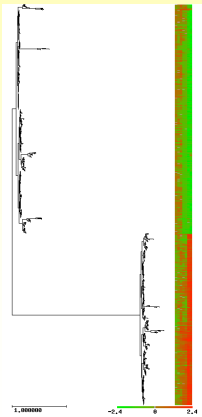
Garber M. E. *et al.* PNAS, vol.98, nº24; 2001.



Data analysis in microarrays

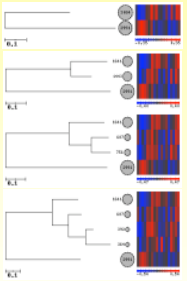
Hierarchical clustering & dendrograms





SOTA

Self-Organizing Tree Algorithm

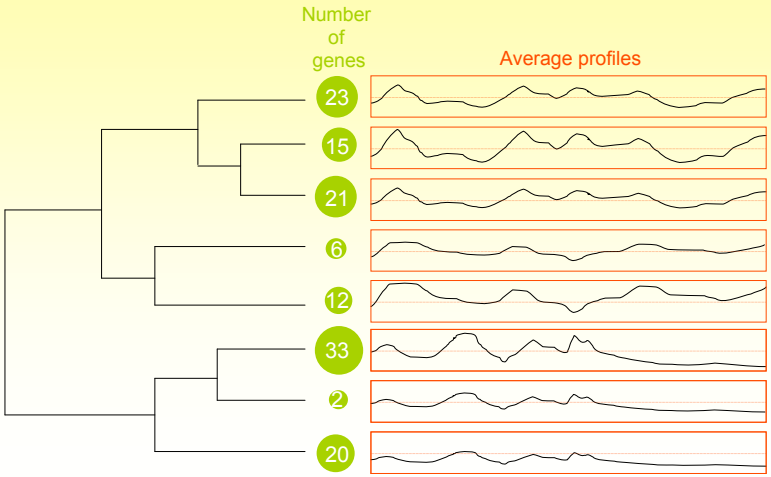


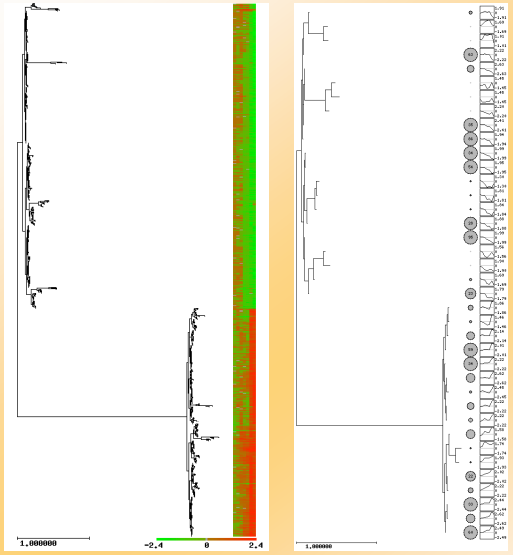
- Artificial Neural Network
- Growth from the root of the tree, toward the leaves
  (from lower to higher resolution)
- Threshold of resource value should be fixed
- Generation of a hierarchical cluster structure at the desired level of resolution

Herrero *et al.* Bioinformatics. 17: 126-136. 2001.

SOTA

Self-Organizing Tree Algorithm



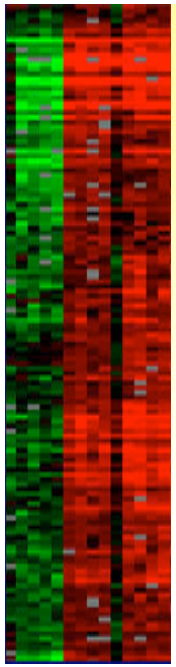


SOTA

Self-Organizing Tree Algorithm

Un ejemplo…

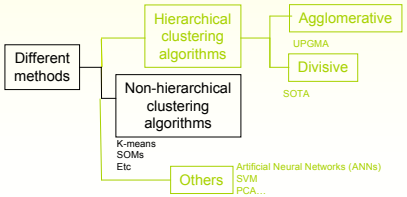
En nuestra práctica: GEPAS



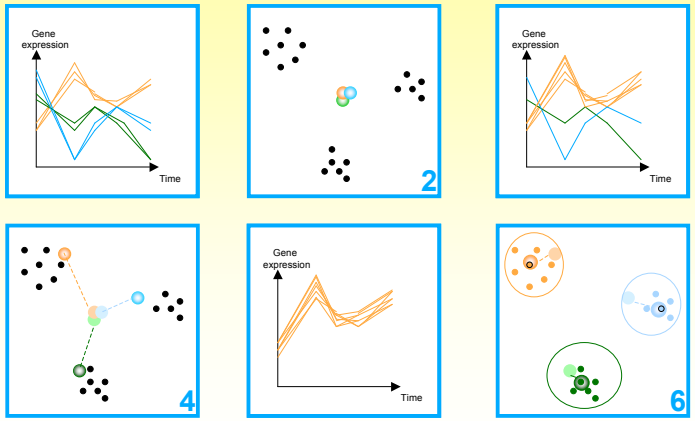
Data analysis in microarrays

K-means clustering
&
Self organizing-maps (SOMs)

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \|x_n - \mu_j\|^2,$$



K-means clustering



K-means clustering

1. Compute cluster centers:

$$x_{(k)j} = \frac{1}{n_k} \sum_{i=1}^{N} I[O_i \subset C_k] x_{ij}$$

2. Compute the **Euclidean Distance** of each object to the K centroids:

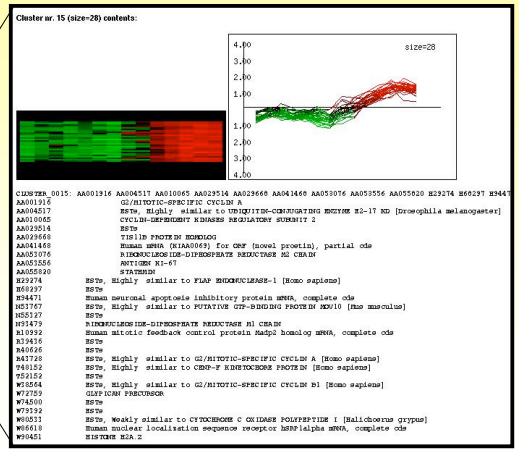
$$e^2_{i(k)} = \sum_{j=1}^{P} (x_{ij} - x_{(k)j})^2 = (\mathbf{x}_i - \mathbf{x}_{(k)})'(\mathbf{x}_i - \mathbf{x}_{(k)})$$

3. Reassign each objecte to its nearest *cluster*.

More information: http://biodiver.bio.ub.es

K-means clustering

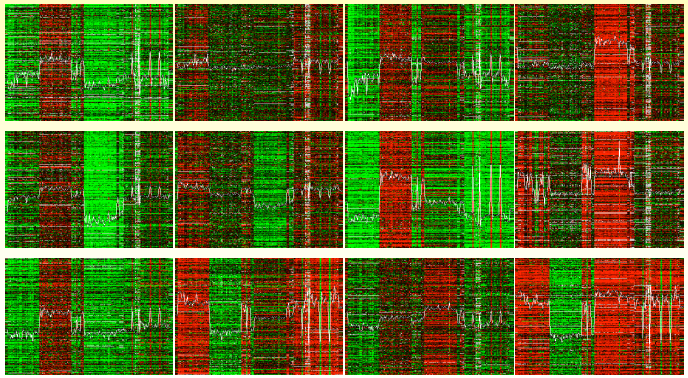
Example



www.isrec.isb-sib.ch.

K-means clustering

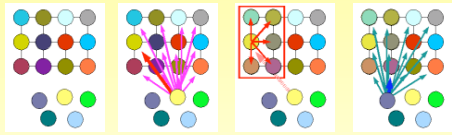
Example 2 ⇒ K= 12



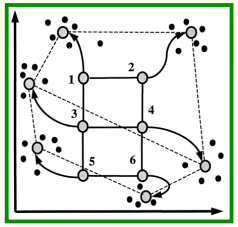
SOMs

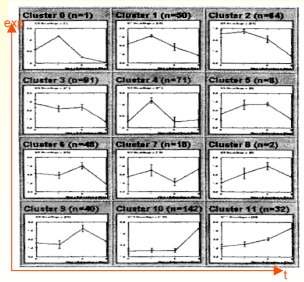
Related genes

Geometric dependence, i.e. 4 x 3



Self-organizing maps (SOMs) or Kohonen Maps

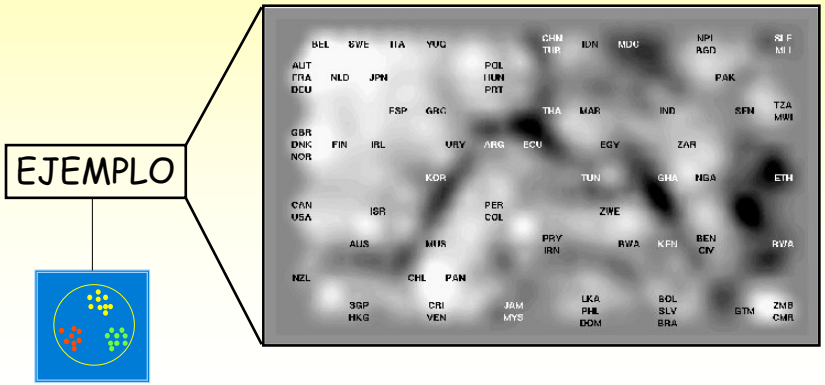




Kohonen T. Proc IEEE 78(9):1464-1480,1990
Tamayo P. *et al.* PNAS, 96: 2907-2912; 1999.

Self-organizing maps (SOMs)

Indices de bienestar mundiales



http://biodiver.bio.ub.es

Brief history in microarray data analysis:

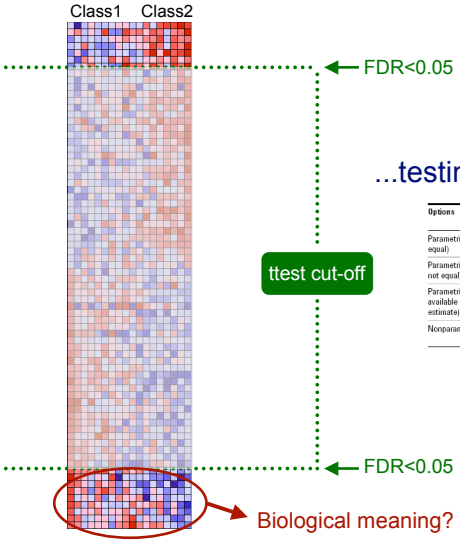
Unsupervised methods

1. **Clustering methods**
   - UPGMA (Sneath and Sokal, 1973)
   - kmeans (Hartigan and Wong 1979)
   - kmedians (Hartigan and Wong 1979)
   - SOTA (Herrero et al. 2001)
   - SOM (Kohonen 1979, Tamayo et al 1999)
   - Gene Shaving (Hastie et al, 2000)
   - Fuzzy methods (Dougherty ET AL. 2002)
   - Probabilistic clustering (Bhattacharjee et al. 2001)
   - Metagenes (Pittman et al, 2004)
2. **Exploratory analysis**
   - parametric: ttest, SAM (Tusher et al, 2001), ANOVA…
   - non parametric: Welch ttest, Wilcoxon, Kruskal Wallis…
   - Sumarizing datasets: PCA
3. **Blocks of genes**
   - Kolmogorov-Smirnof: GSEA (Subramanian et al. 2005)
   - FatiScan (Al-Shahrour et al. 2005)
   - SAM 3.0 (Jan 07)
   - GeneTrail (Backes et al. 2007)

Supervised methods (predictors)

- Neural networks (Khan et al. 2001)
- SOMs (Kohonen et al. 1984)
- kNN (Ripley 1996; Hastie et al 2001)
- PAM (Tibshirani et al. 2002)
- DLDA (Dudoit et al. 2002)
- SVMs (Furey et al, 2000)

Exploratory analysis



...testing genes independently...

Options	Specific test name: Analyzing 2 groups	Specific test name : Analyzing more than 2 groups
Parametric (variances equal)	Student's T-test	ANOVA
Parametric (variances not equal)	Welch t-test	Welch ANOVA
Parametric (use all available error estimate)	Welch t-test using error model variances	Welch ANOVA using error model variances
Nonparametric	Wilcoxon-Mann-Whitney test	Kruskal-Wallis test

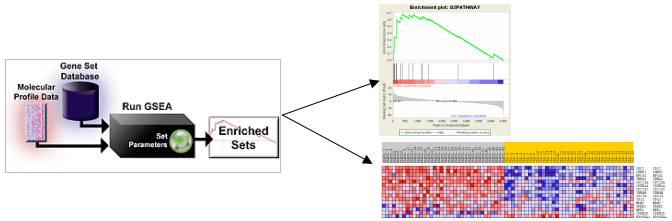
Currently…testing block of genes ~ biological pathways

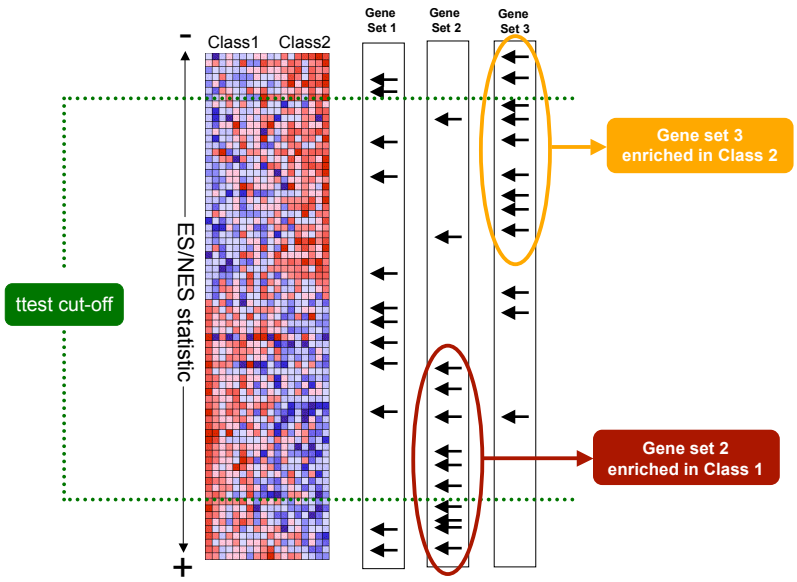
GSEA (Subramanian et al. PNAS. 2005.)

Gene Set Enrichment Analysis
Broad Institute
v 2.0 available since Jan 2007

New version includes Biocarta, Broad Institute, GeneMAPP, KEGG annotations and more...
Platforms: Affymetrix, Agilent, CodeLink, 2-color...

GSEA applies Kolmogorov-Smirnof test to find assymmetrical distributions for defined blocks of genes in datasets whole distribution.





Now…analizing block of genes ~ biological pathways

FatiScan

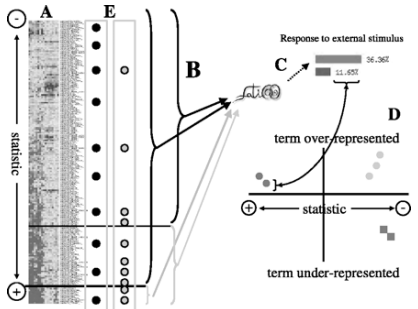
Al-Shahrour et al. Bioinformatics. 2005 Jul 1;21(13):2988-93.

Available in Babelomics suite: www.babelomics.org

Segmentation test for searching asymmetries

More statistically sensitive than GSEA

Few annotations implemented



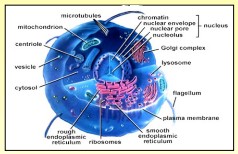
En nuestra practica:

www.babelomics.org

A discussion…

Mathematics methods for biology

?



Current math´s methods are useful to explain real and meaningful processes from a biological point of view?

mmm…actually we don´t know…sometimes they are…sometimes not

New integrative methods should be developed…

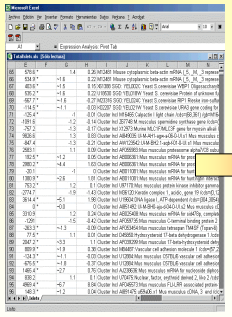
Datamining

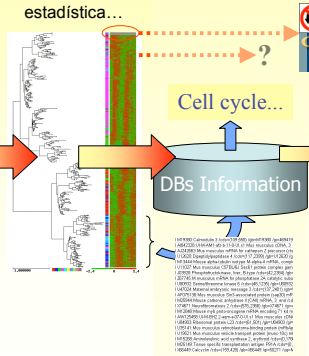
Qué es, que hace y que sabemos de este gen?

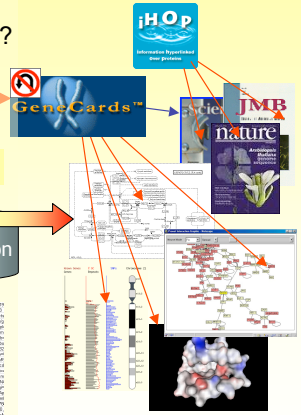
Tengo mis datos...

He analizado su estructura y "relevancia" estadística…

Que hay en este cluster?



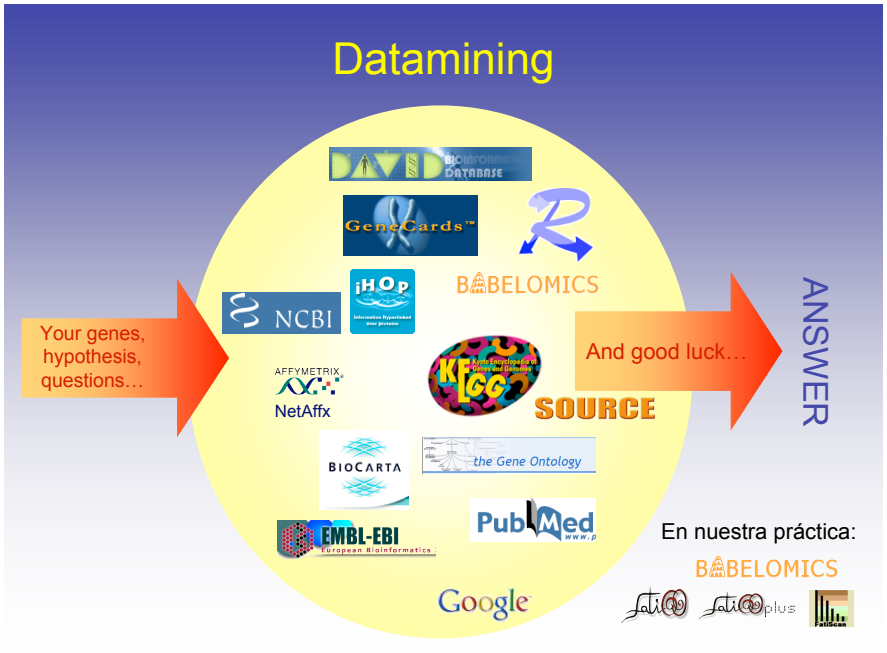




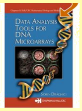
Clustering,estadística, etc — *Datamining* — Links

Source: J. Dopazo

Datamining



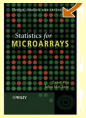
Para profundizar…

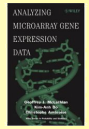


Data Analysis Tools for DNA Microarrays
by Sorin Draghic

DNA Microarrays and Gene Expression : From Experiments to Data Analysis and Modeling
by Pierre Baldi, G. Wesley Hatfield, Wesley G. Hatfield

Exploration and Analysis of DNA Microarray and Protein Array Data (Wiley Series in Probability and Statistics)
by Dhammika Amaratunga, Javier Cabrera



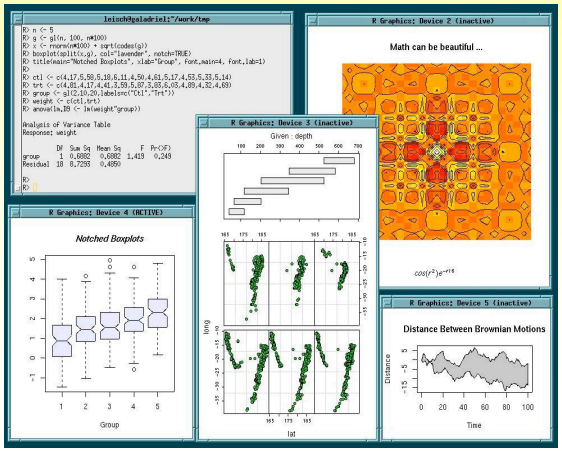


Statistics for Microarrays : Design, Analysis and Inference
by Ernst Wit, John McClure

Analyzing Microarray Gene Expression Data (Wiley Series in Probability and Statistics)
by Geoffrey J. McLachlan, Kim-Anh Do, Christophe Ambroise

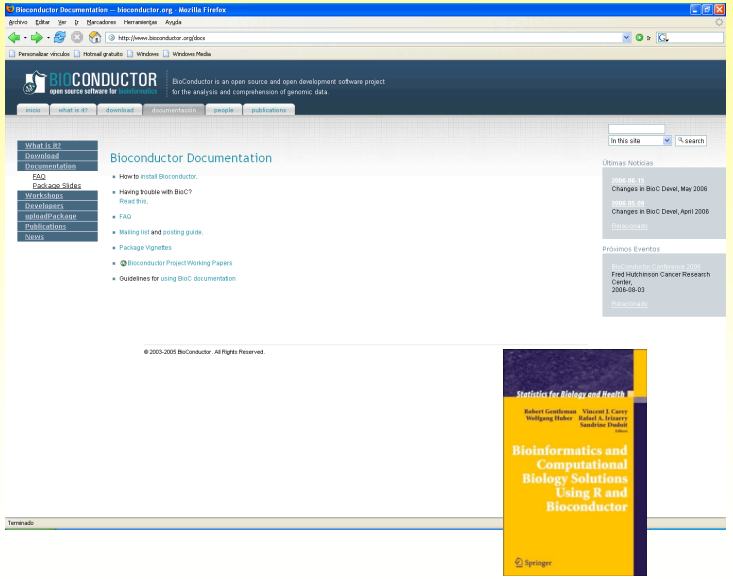
Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)
by Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit

Microarray Gene Expression Data Analysis: A Beginner's Guide
by Helen Causton

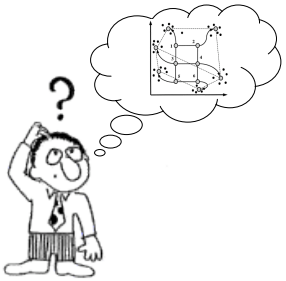


www.r-project.org

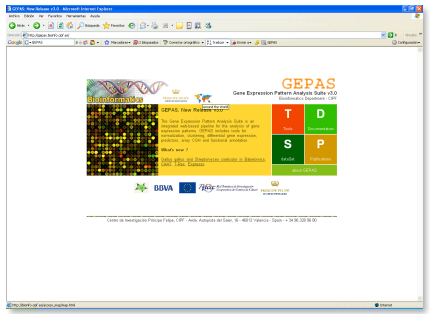
Windows
Mac OS X
LINUX/UNIX

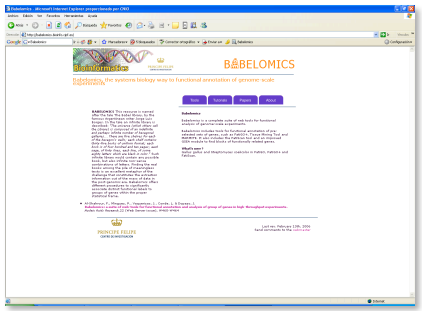


Gracias!!



Gonzalo Gómez López
ggomez@cnio.es

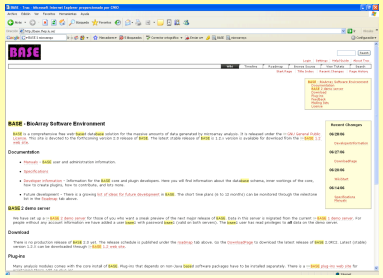


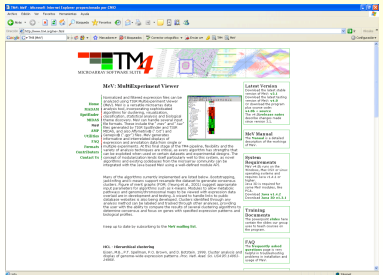


http://www.gepas.org
http://www.babelomics.org

Universidad Complutense de Madrid
ESCUELA DE VERANO
2007

1. Introducción
2. Microarrays (tipos, tratamiento de las muestras e hibridación, bases de datos, aplicaciones…)
3. Normalización y análisis exploratorio (JC Oliveros)
4. Análisis de datos (Análisis supervisado y no supervisado, algoritmos…)
5. Práctica





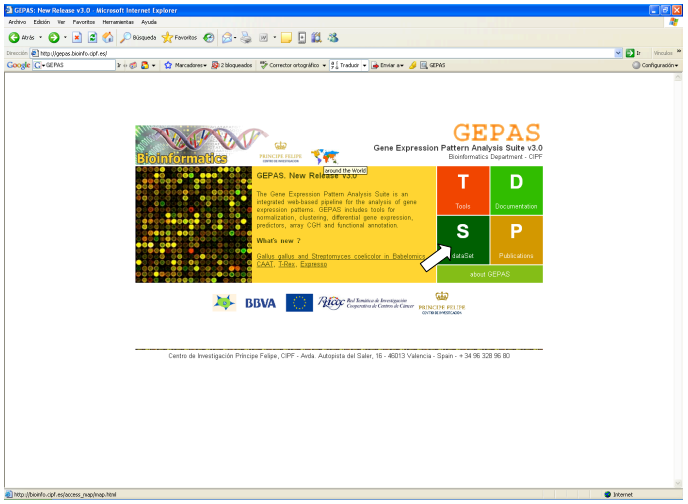
Otras herramientas gratuitas:

-BASE
http://base.thep.lu.se/

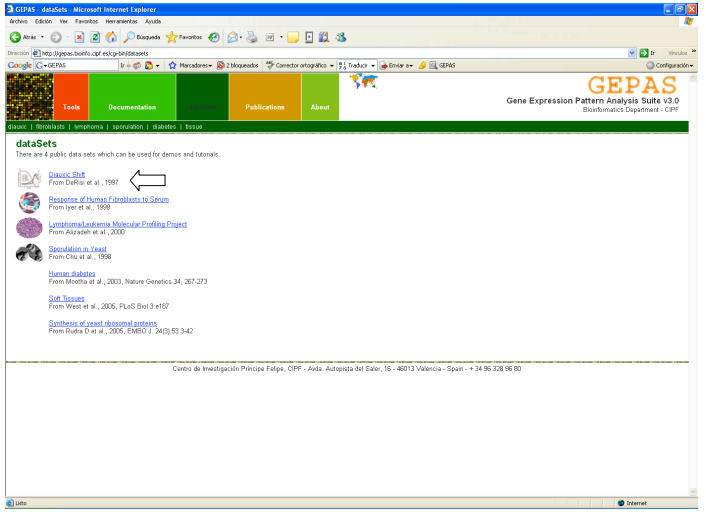
-BRBTools
http://linus.nci.nih.gov/BRB-ArrayTools.html

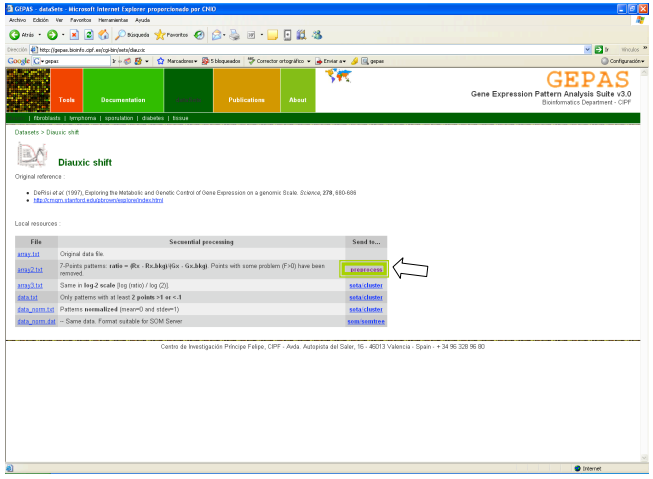
-TM4(MeV)
http://www.tm4.org/mev.html

-Otras
SNOMAD, DAVID, MIDAW…



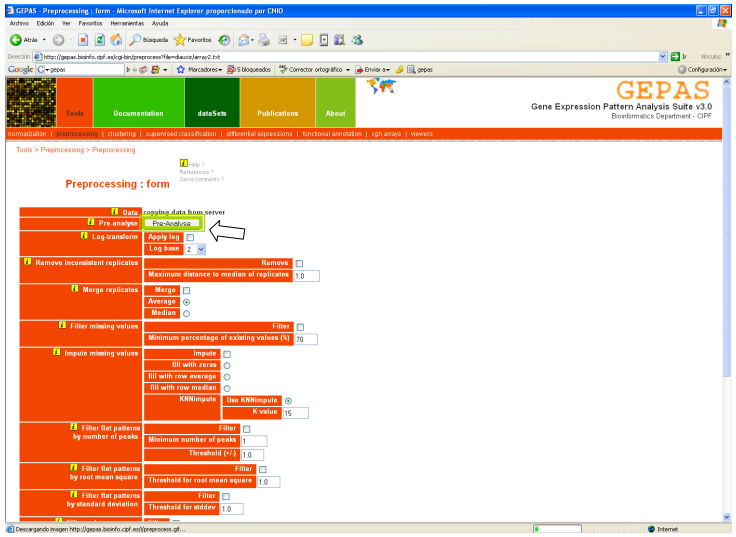
www.gepas.org

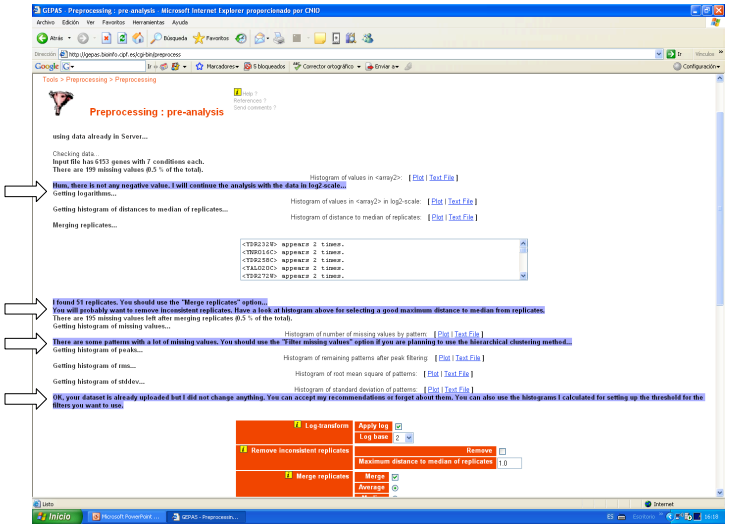


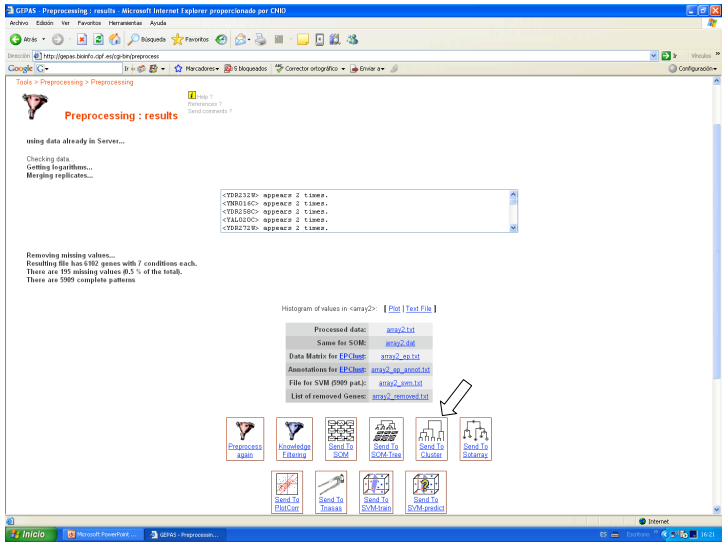


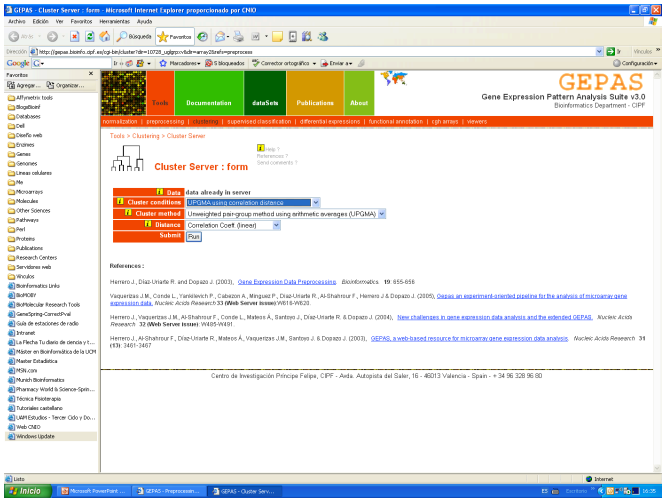
INPUT FILE FORMAT
(preprocessing)

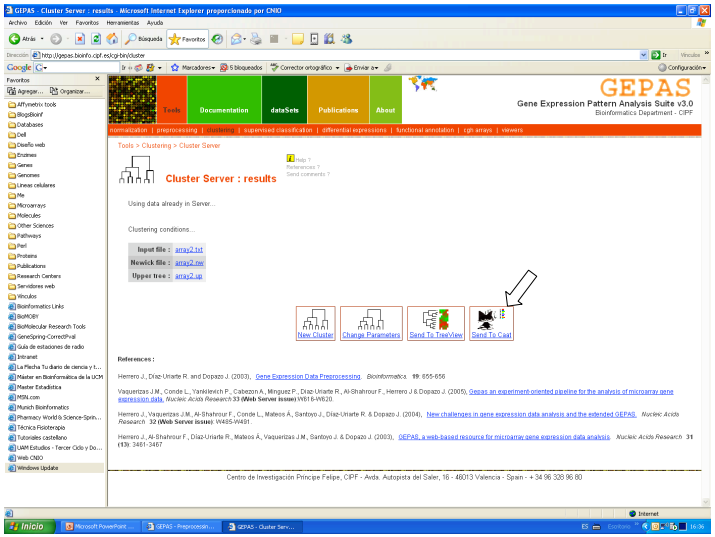
```
#NAMES    t0       t0.5     t2       t5       t7       t9       t11.5
#LABELS -00 "Metabolic" -01 "Early I" -02 "Early II" -03 "Early-Mid" -04 "Middle"
YHR007C   1.210    1.531    2.722    1.062    0.639    0.570    0.495
YOL109W   1.819    0.860    1.592    1.920    2.372    2.854    2.079
YAL056W   1.065    1.149    1.090    0.781             1.959    1.143
YAL058W   1.135    0.824    0.954    1.017    0.733    0.656    0.760
YAL059W   1.278    0.328    0.702    0.621    0.874    1.646    1.851
YAL060W   1.190    0.641    0.592    0.375    0.281    0.539    1.057
YAL061W   1.216    1.147    1.303    0.583    1.202    1.290    1.330
YAL062W@0 1.260    5.370    8.040    2.686    1.341    1.381    2.796
YAR002W   1.019    0.784    0.821    0.730    0.794    0.911    0.622
YAR003W@3 0.845    1.296    1.536    2.098    2.904    3.951    2.901
YAR007C@0 0.937    2.984    5.367    4.252    4.454    1.612    1.941
YAR008W   0.740    0.804    2.808    1.962    1.970    1.249    0.904
YAR009C   0.954    0.949    0.799    0.643    0.672    0.894    0.539
YAR010C   0.931    0.956    0.889    0.746    0.905    0.860    0.683
YAL001C   0.998    0.759    0.908    0.837             0.896    1.020
YAL002W   1.058    1.288    1.107    0.793    0.503    0.660    1.037
YAL003W   1.206    0.258    0.412    0.340    0.245    0.509    0.501
YAL004W   1.366    0.686    0.544    0.479    0.359    0.306    0.363
YAL005C   1.566    0.848    0.416    0.410    0.276    0.299    0.233
YAL007C   1.107    0.607    1.336    0.856    0.683    0.523    0.593
YAL008W   1.324    0.768    1.318    0.826    0.911    0.914    0.937
YAL009W   1.056    0.923    1.033    0.972    1.088    0.904    1.211
YAL010C   1.152    0.805    0.593    0.411    0.293    0.324    0.361
YAL011W   1.108    0.672    1.016    0.934    0.797    0.876    1.088
YAL012W   1.045    2.337    0.563    0.317    0.152    0.198    0.203
YAL013W   1.233    1.087    0.754    0.666    0.688    0.881    0.726
YAL014C   1.043    1.130    0.915    0.990    1.215    1.152    1.429
YAL015C   0.906    0.786    0.658    0.395    0.333    0.401    0.314
YAL016W   1.017    0.862    0.598    0.435    0.267    0.621    0.431
YAL017W   0.933    0.780    0.784    0.731    0.820    0.910    0.947
YAL018C@4 0.718    0.582    0.592    13.217   25.203   10.166   21.863
YAL019W   0.905    0.316    0.829    0.821    0.833    0.982    0.959
YAL020C   0.854    0.892    0.801    0.902    0.843    1.042    0.950
YAL021C   1.030    0.768    1.004    0.919    0.950    1.033    0.975
YAL022C   0.721    1.399    1.227    1.031    1.661    1.064    1.639
YAL023C   0.873    0.776    1.273    1.180    0.830    0.572    0.485
YAL024C   0.925    1.007    1.242    1.151    1.157    1.345    0.986
YAL025C   0.935    0.085    0.278    0.454    0.207    0.529    0.713
YAL026C   1.242             0.911    1.510    1.855    2.717    3.010
YAL027W   1.200    0.836    0.979    0.972    1.346    1.596    1.814
YAL028AW  1.265    1.451    0.767    1.035    0.897    1.120    1.109
YAL028W   1.150    1.028    0.742    0.603    0.578    0.588    0.732
YAL029C   1.066    0.664    0.876    1.147    1.186    1.177    0.830
YAL030W   1.146    1.028    1.354    1.433    1.505    1.461    1.386
YAL031C   1.272    1.321    1.067    0.896    0.982    0.908    0.783
YAL032C   1.172    0.863    0.687    0.658    0.596    0.633    0.762
YAL033W   1.242    0.960    1.125    0.843    0.987    1.380    0.970
YAL034AW  1.080    0.577    0.997    1.021    1.211    1.564    1.263
```

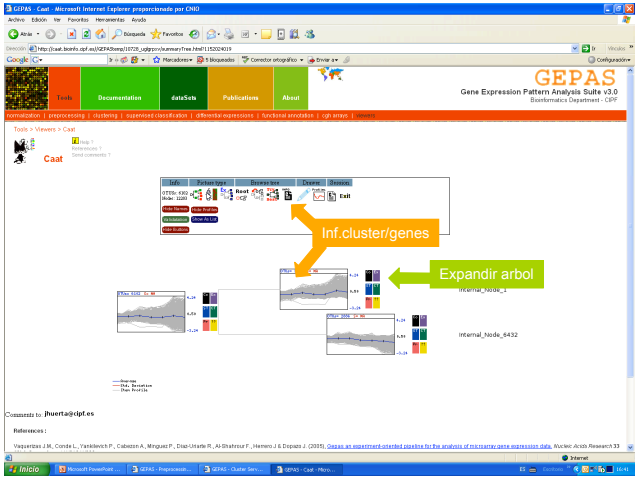

GEPAS - Preprocessing : form - Microsoft Internet Explorer proporcionado por CNIO
GEPAS
Gene Expression Pattern Analysis Suite v3.0
Bioinformatics Department - CIPF
Tools
Documentation
dataSets
Publications
About
Tools > Preprocessing > Preprocessing
Preprocessing : form
Data
copying data from server
Pre analyse
Pre-Analyse
Log-transform
Apply log
Log base
Remove inconsistent replicates
Remove
Maximum distance to median of replicates
Merge replicates
Merge
Average
Median
Filter missing values
Filter
Minimum percentage of existing values [%]
Impute missing values
Impute
fill with zeros
fill with row average
fill with row median
KNNimpute
Use KNNimpute
K value
Filter flat patterns by number of peaks
Filter
Minimum number of peaks
Threshold (+/-)
Filter flat patterns by root mean square
Filter
Threshold for root mean square
Filter flat patterns by standard deviation
Filter
Threshold for stddev

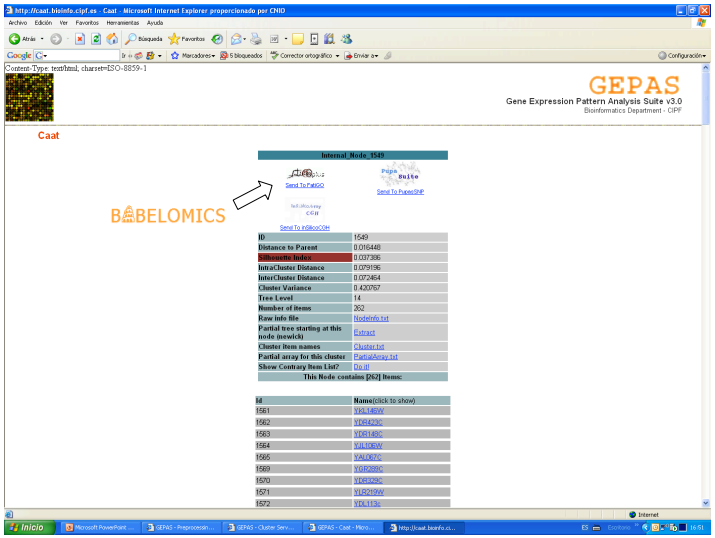

GEPAS - Preprocessing : pre_analysis - Microsoft Internet Explorer proporcionado por CNIO
Tools > Preprocessing > Preprocessing
Preprocessing : pre-analysis
using data already in Server...
Checking data...
Input file has 6153 genes with 7 conditions each.
There are 199 missing values (0.5 % of the total).
Hum, there is not any negative value. I will continue the analysis with the data in log2 scale...
Getting logarithms...
Getting histogram of distances to median of replicates...
Merging replicates...
<TDR232W> appears 2 times.
<TDR016C> appears 2 times.
<TDR258C> appears 2 times.
<TAL020C> appears 2 times.
<TDR272W> appears 2 times.
I found 51 replicates. You should use the "Merge replicates" option...
You will probably want to remove inconsistent replicates. Have a look at histogram above for selecting a good maximum distance to median from replicates.
There are 195 missing values left after merging replicates (0.5 % of the total).
Getting histogram of missing values...
There are some patterns with a lot of missing values. You should use the "Filter missing values" option if you are planning to use the hierarchical clustering method...
Getting histogram of peaks...
Getting histogram of rms...
Getting histogram of stddev...
OK, your dataset is already uploaded but I did not change anything. You can accept my recommendations or forget about them. You can also use the histograms I calculated for setting up the threshold for the filters you want to use.
Log-transform
Apply log
Log base
Remove inconsistent replicates
Remove
Maximum distance to median of replicates
Merge replicates
Merge
Average

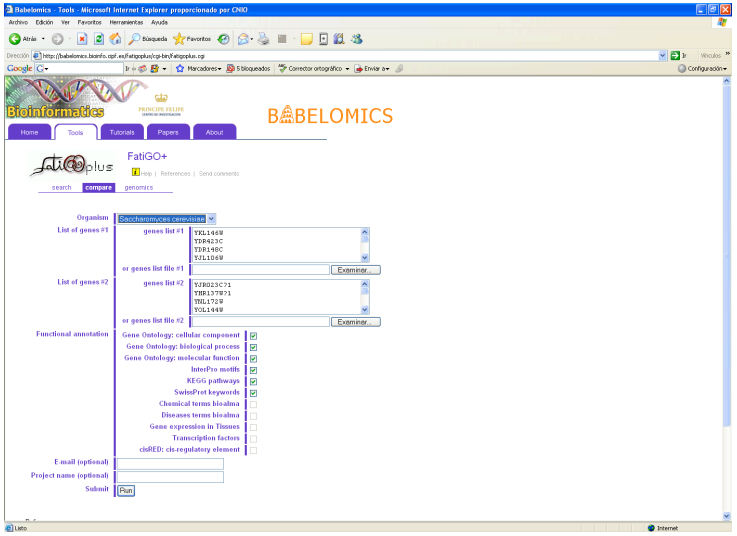

GEPAS - Preprocessing : results - Microsoft Internet Explorer proporcionado por CNIO
Tools > Preprocessing > Preprocessing
Preprocessing : results
using data already in Server...
Checking data...
Getting logarithms...
Merging replicates...
<TDR232W> appears 2 times.
<TDR016C> appears 2 times.
<TDR258C> appears 2 times.
<TAL020C> appears 2 times.
<TDR272W> appears 2 times.
Removing missing values...
Resulting file has 6102 genes with 7 conditions each.
There are 195 missing values (0.5 % of the total).
There are 5909 complete patterns
Histogram of values in <array2>: [Plot | Text File]
Processed data: array2.txt
Same for SOM: array2.dat
Data Matrix for EPChart: array2_ep.txt
Annotations for EPChart: array2_ep_annot.txt
File for SVM (5909 pat.): array2_svm.txt
List of removed Genes: array2_removed.txt
Preprocess again
Knowledge Filtering
Send To SOM
Send To SOM-Tree
Send To Cluster
Send To Sotaray
Send To PkdCorr
Send To Tnasas
Send To SVM-train
Send To SVM-predict

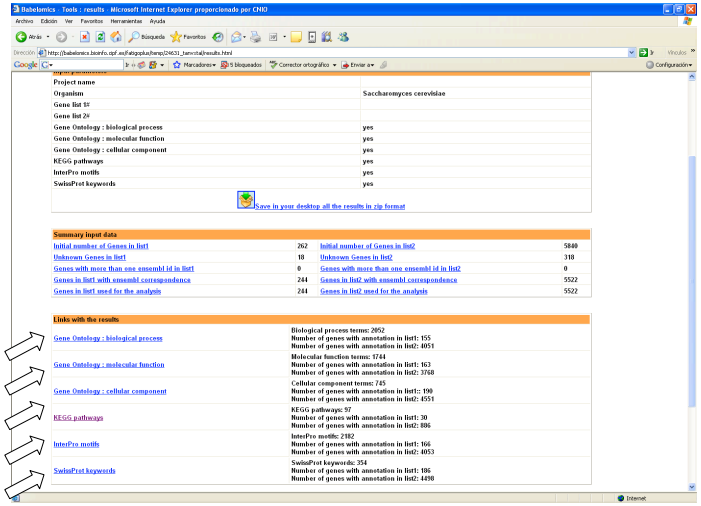

GEPAS - Cluster Server : form - Microsoft Internet Explorer proporcionado por CNIO
GEPAS
Gene Expression Pattern Analysis Suite v3.0
Bioinformatics Department - CIPF
Tools
Documentation
dataSets
Publications
About
Tools > Clustering > Cluster Server
Cluster Server : form
Data
data already in server
Cluster conditions
Cluster method
Unweighted pair-group method using arithmetic averages (UPGMA)
Distance
Correlation Coeff. (linear)
Submit
Run
References :
Centro de Investigación Príncipe Felipe, CIPF - Aeda. Autopista del Saler, 16 - 46013 Valencia - Spain - + 34 96 328 96 80


GEPAS
Gene Expression Pattern Analysis Suite v3.0
Bioinformatics Department - CIPF
Cluster Server : results
New Cluster
Change Parameters
Send To TreeView
Send To Caat

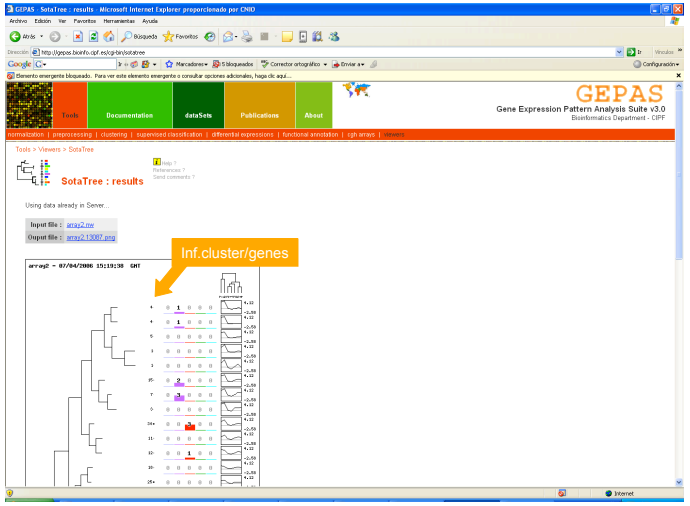

GEPAS
Gene Expression Pattern Analysis Suite v3.0
Caat
Inf.cluster/genes
Expandir arbol


GEPAS
Gene Expression Pattern Analysis Suite v3.0
Caat
BABELOMICS

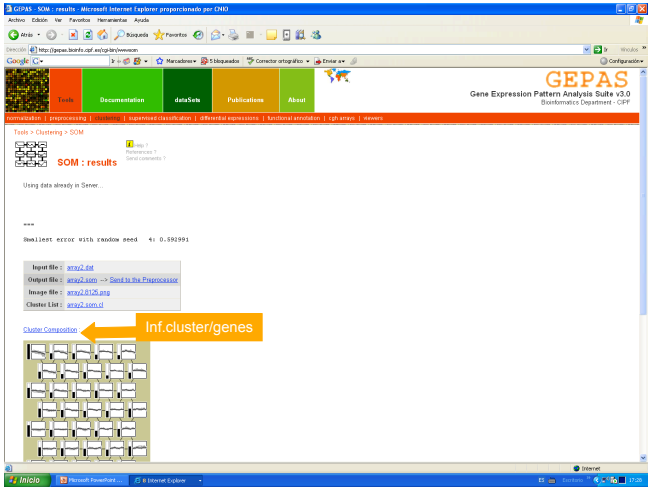

BABELOMICS
FatiGO+



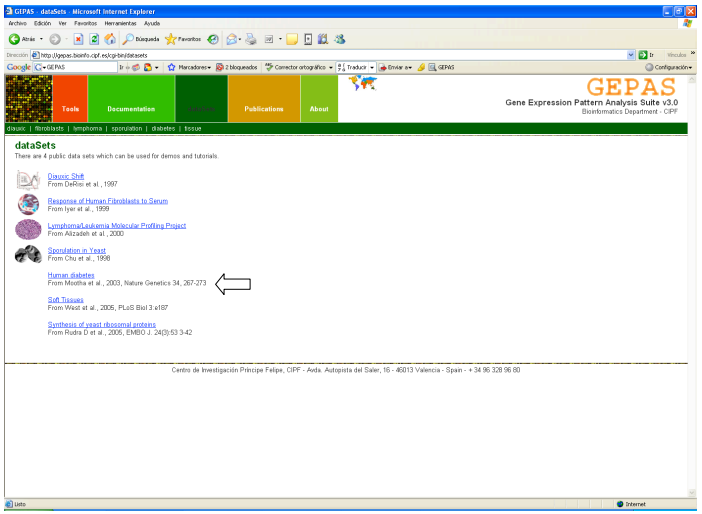
SOTA…

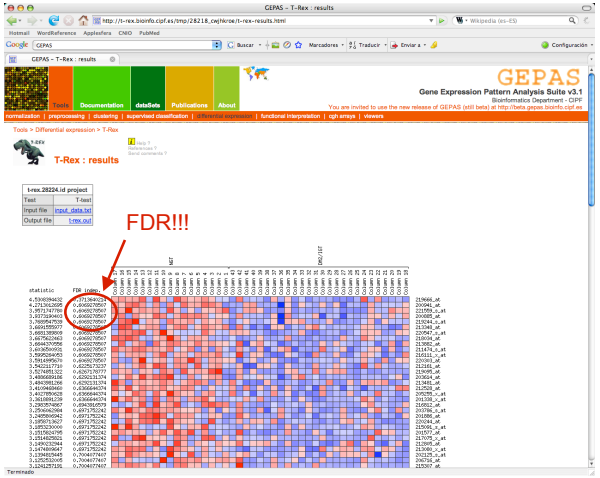


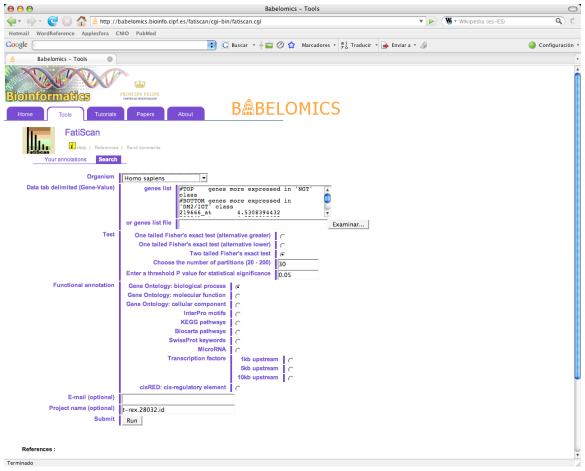
SOMs…



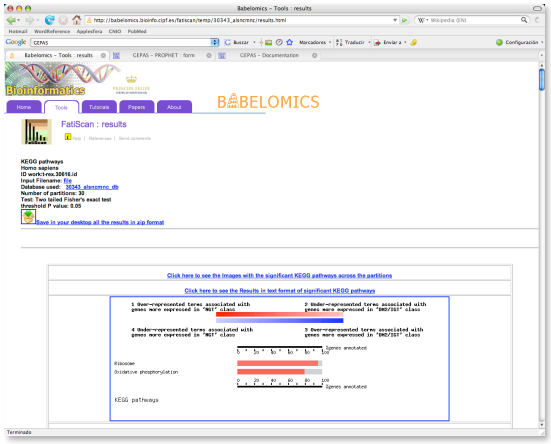
FatiScan



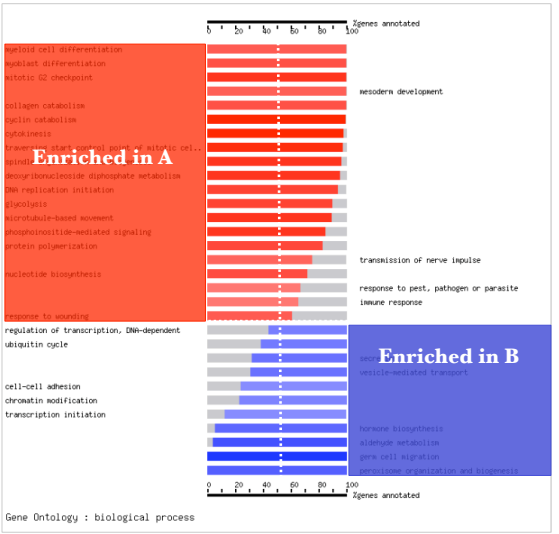




FatiScan Output



FatiScan output



% Genes with the specific GO annotation for each partition