Universitat Autònoma de Barcelona

etse
Escola Tècnica Superior d'Enginyeria

INTERFAZ WEB PARA ESTUDIAR EL EFECTO DE DIFERENTES CONDICIONES SOBRE LA EXPRESIÓN DE LOS GENES

Realizado por : **José Fernández Márquez**
Director : **Jordi González Sabaté (CVC-UAB)**
Codirector 1 : **Mario Huerta (IBB-UAB)**
Codirector 2 : **Juan Antonio Cedano (IBB-UAB)**

PRESENTACIÓN

ESTADO DEL ARTE

OBJETIVOS

IMPLEMENTACIÓN

CONCLUSIONES

Instituto de Biotecnología y Biomedicina (IBB)

- En el IBB se desarrollan principalmente investigaciones de tipo biológico

- El trabajo se desarrolló en el IBB bajo la tutela de Mario Huerta y con la colaboración de Juan Antonio Cedano

- El trabajo realizado se enmarca en una linea de investigación dirigida por Mario Huerta y Juan Antonio Cedano que estudia el efecto del estrés en las células humanas, cómo el estrés puede generar células cancerígenas.

Tecnología de microarrays

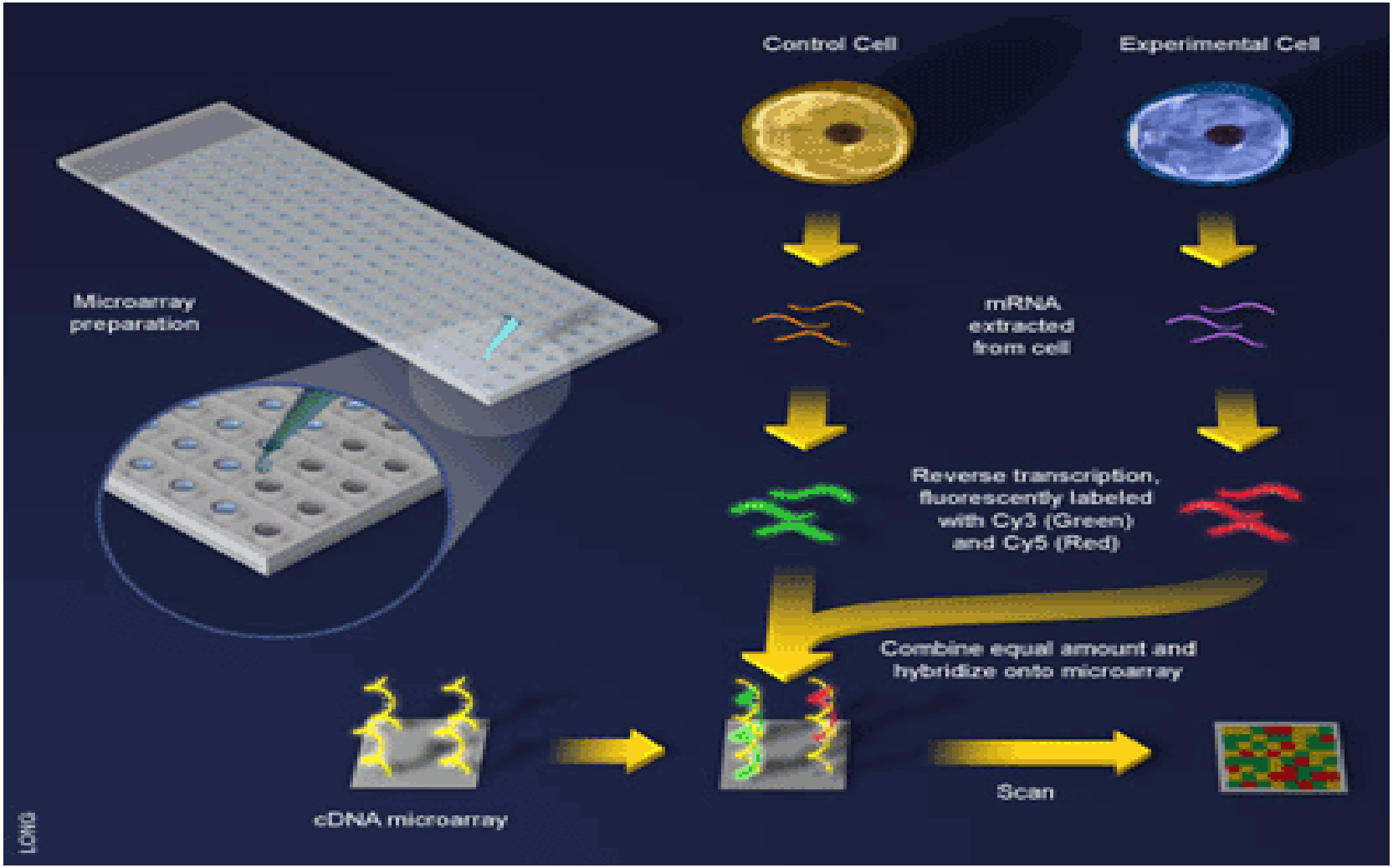
Métodos de agrupación

Índices de Integridad

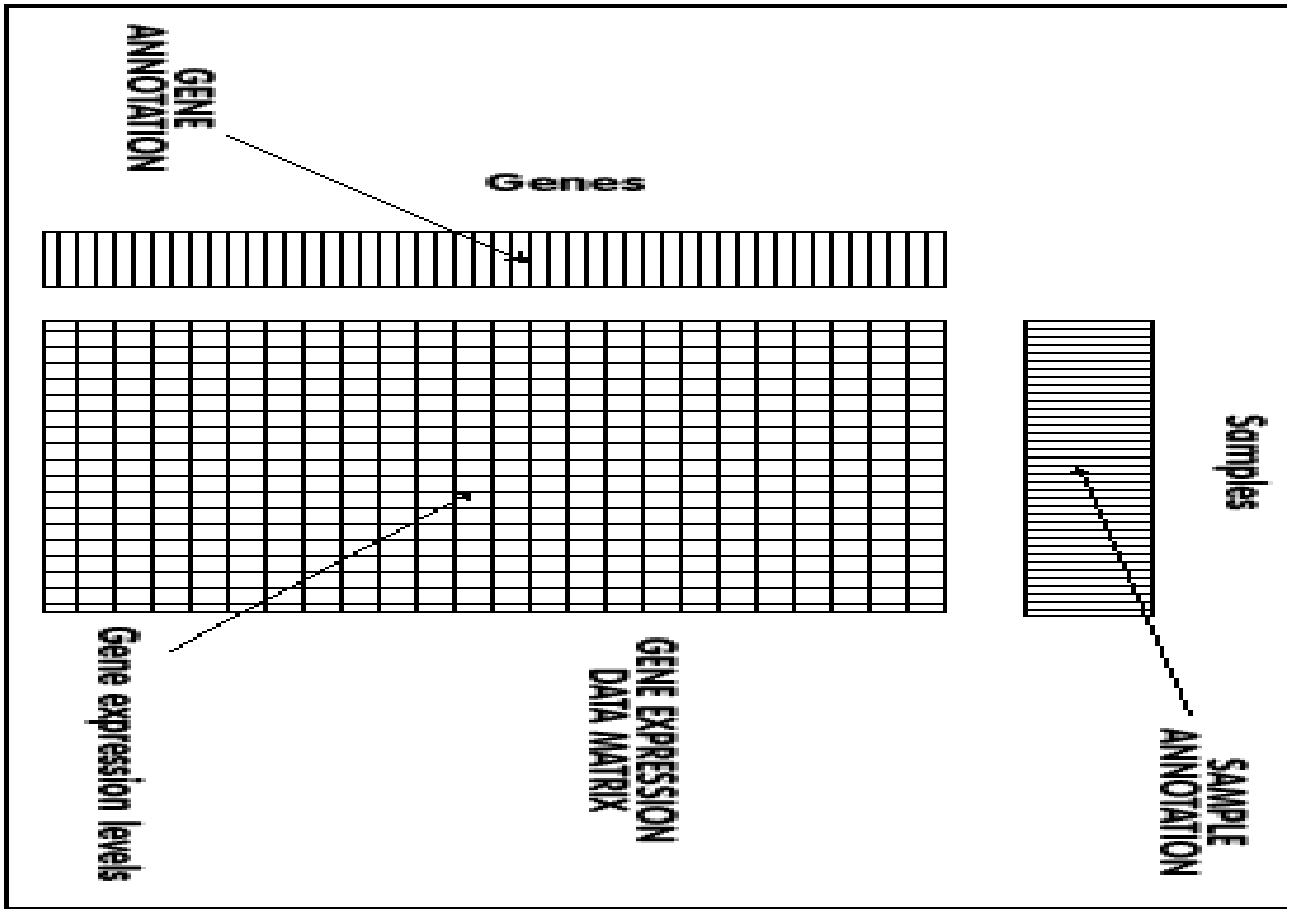
Intervalos de confianza

PCOPGene

Tecnología de microarrays



Tecnología de microarrays



Métodos de Agrupación más utilizados en el análisis de microarrays

Escalado matriz de datos:

- Multi Dimensional Scaling (MDS)
- Principal Components (PC)

Métodos agrupación :

Jerarquicos:

- Hierarchical Clustering (HC)

De particionamiento:

- K-Means
- Partitioning Around Medoids (PAM)
- Self-organizing Maps (SOM)
- Self-organizing Tree Algorithms (SOTA)

Índices de integridad

Hartigan

Calinsky-Harabasz

Dunn

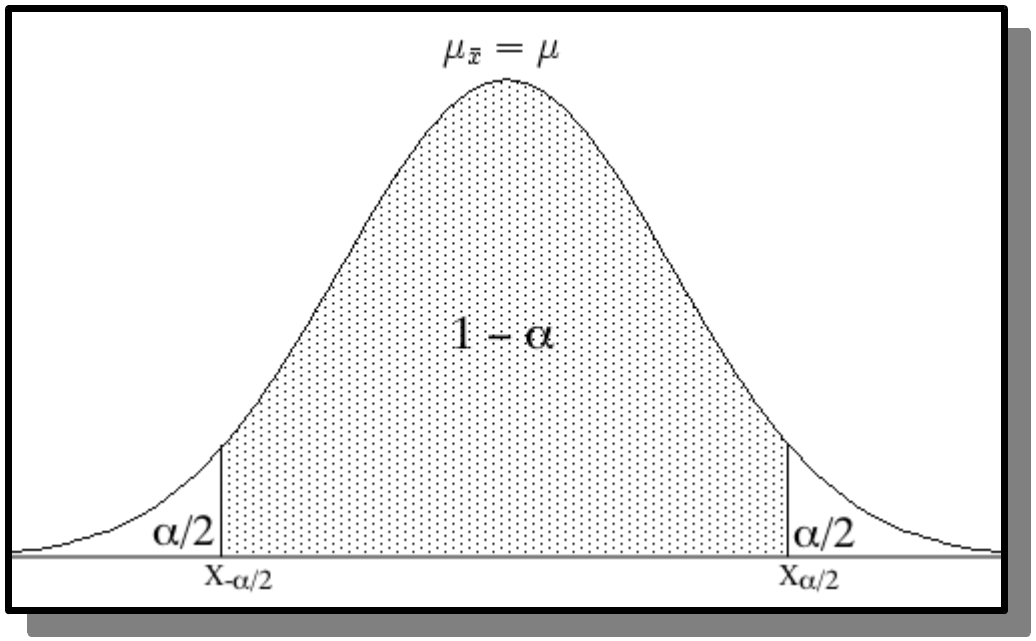
Silhouette Width

Connectivity

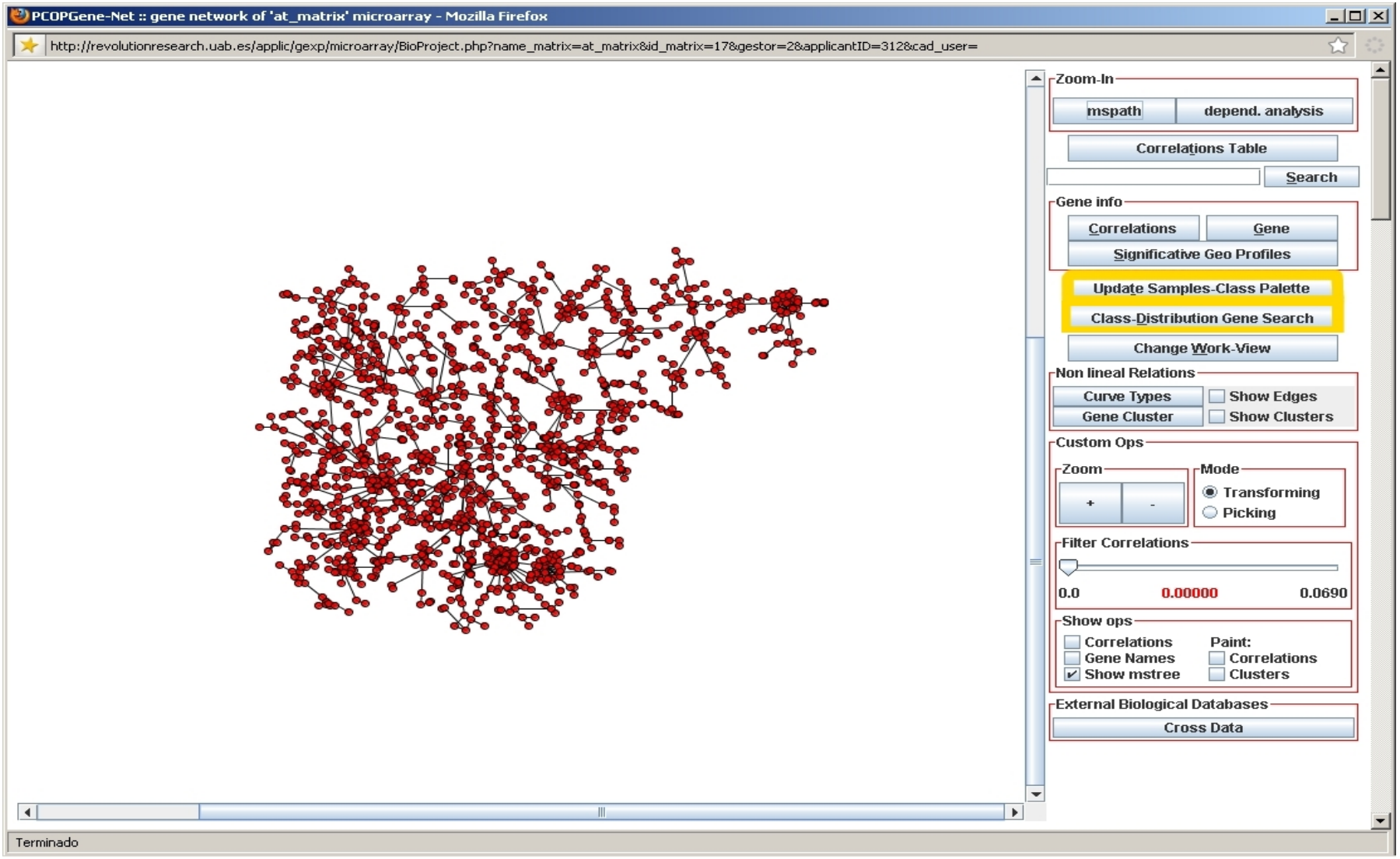
Búsqueda de Genes Marcadores para una Distribución de Clusters Concreta

Distribución normal (distribución T d Students)

Intervalo de confianza $intervalo = (\bar{X} - Kte * (\frac{desv}{\sqrt{n}}), \bar{X} + Kte * (\frac{desv}{\sqrt{n}}))$



PCOPGene:: Microarray analysis tool



OBJETIVOS

- Implementación algoritmos agrupación de las condiciones muestrales

- Integrar agrupación en el preproceso existente

- Integrar resultados agrupación en la interfaz web PCOPGene (http:://revolutionresearch.uab.es) y añadir nuevas funcionalidades

- Implementar algoritmo búsqueda de genes marcadores

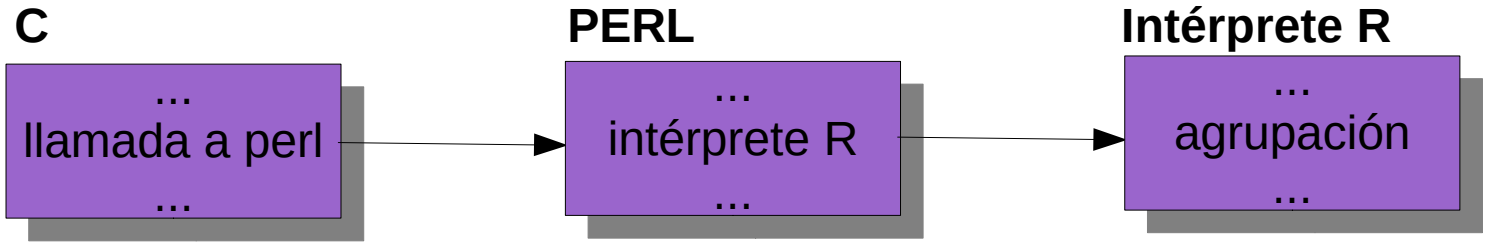
- Integrar implementación y resultados en el interfaz web PCOPGene (http:://revolutionresearch.uab.es)

Agrupación de condiciones muestrales

Herramientas de desarrollo:

- R-Statistics (R)
- PERL
- C

Modelo de implementación:



Tratamiento previo: Corrección de "celdas vacías" en la microarray de entrada

Implementación de los métodos de agrupación:

MDS + (K-MEANS, SOM, SOTA, PAM, HC)
PC + (K-MEANS, SOM, SOTA, PAM, HC)
SOM, SOTA, PAM, HC

Cálculo de la integridad de las distribuciones de clústers:

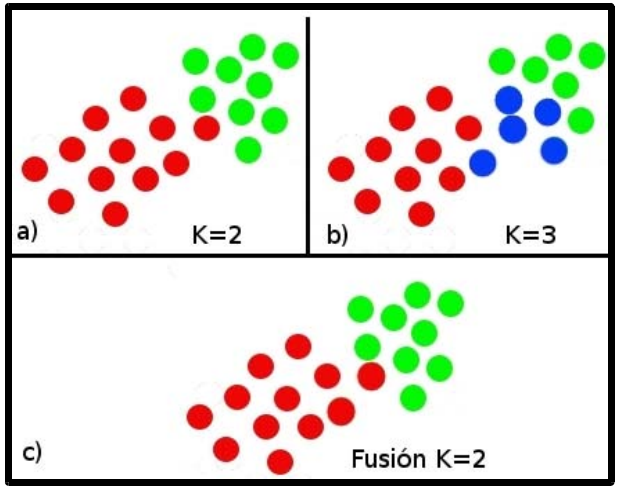
Dunn , Silhouette y Hartigan (*descartado*)

Agrupación de condiciones muestrales

Tratamiento de las condiciones muestrales *outlayers:**

Agrupaciones sin muestras *outlayers* , fusión de clústers

Agrupaciones con muestras *outlayers*



Para cada uno de los algoritmos se escogen las mejores agrupaciones según los índices Dunn y Silhouette

Tratamiento final para todas las agrupaciones:

-Normalizar identificadores de los clústers.

**Outlayer: muestras sin clúster asignado o muestra que pertenece a un grupo con pocas muestras (5% en este caso)*

Agrupación de condiciones muestrales

Tratamiento para las mejores agrupaciones:

- Eliminación *outlayers*

- Si la mejor agrupación tiene 9 clúster se elimina el clúster que contenga menos muestras.

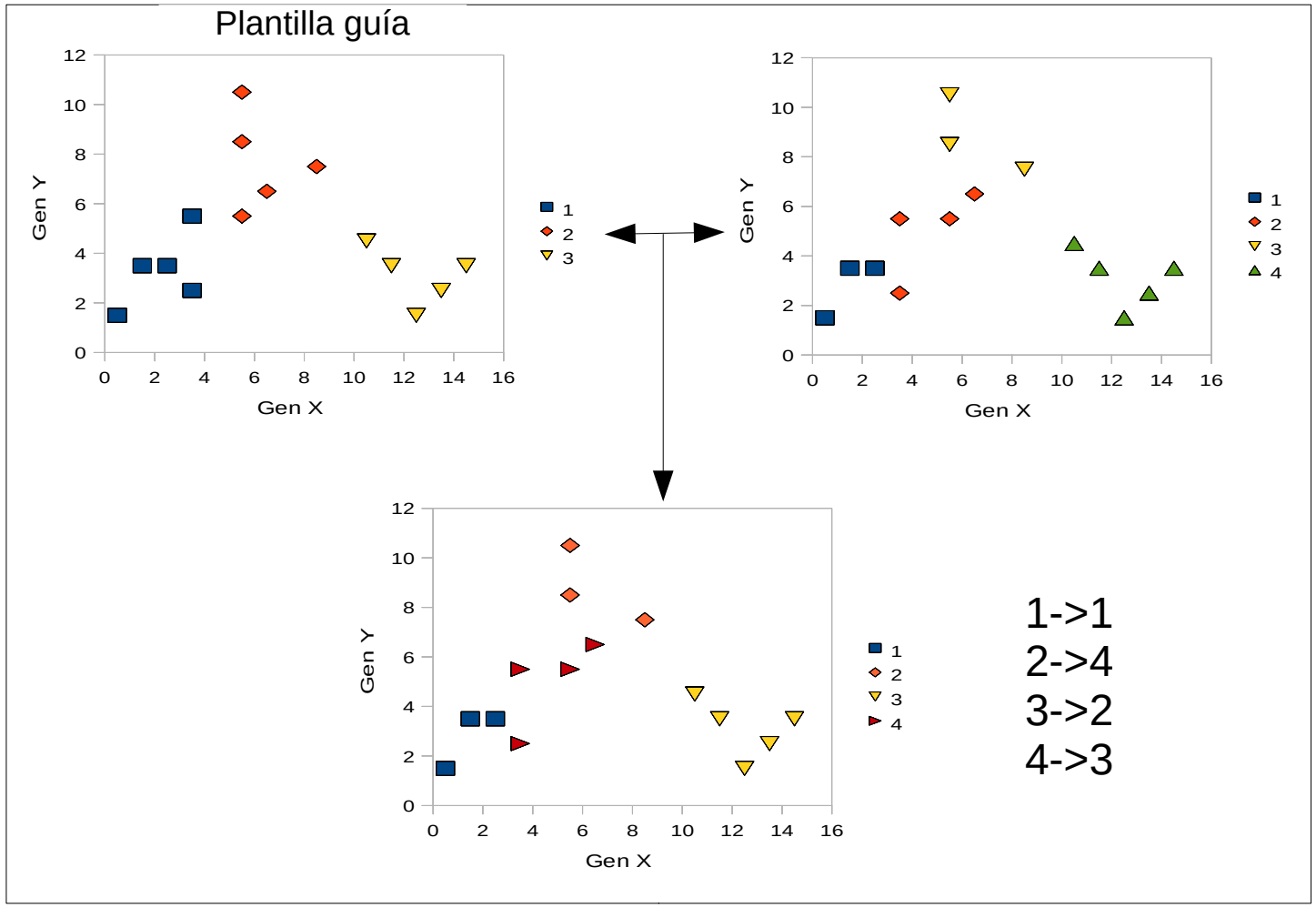
-Ordenación, agrupación y normalización de los ficheros de clústers:

· Proceso independiente de la agrupación de muestras

· Clustering de las mejores agrupaciones agrupándolas por similitud y ordenadas por disimilitud (HC)

· Normalización interna de cada grupo de ficheros de clústers a partir del fichero guía de cada grupo de ficheros.

Agrupación de condiciones muestrales



Agrupación de condiciones muestrales

Gestión de resultados de la agrupación:

Todos los resultados se guardan en ficheros en el servidor.

Los directorios más destacados son:

- Rclustering_Samples : se guardan todos los resultados de las agrupaciones

- Rclustering_Samples/Best: se guardan solo las mejores agrupaciones accesibles al usuario a través del aplicativo web

Integración de la agrupación en el preproceso

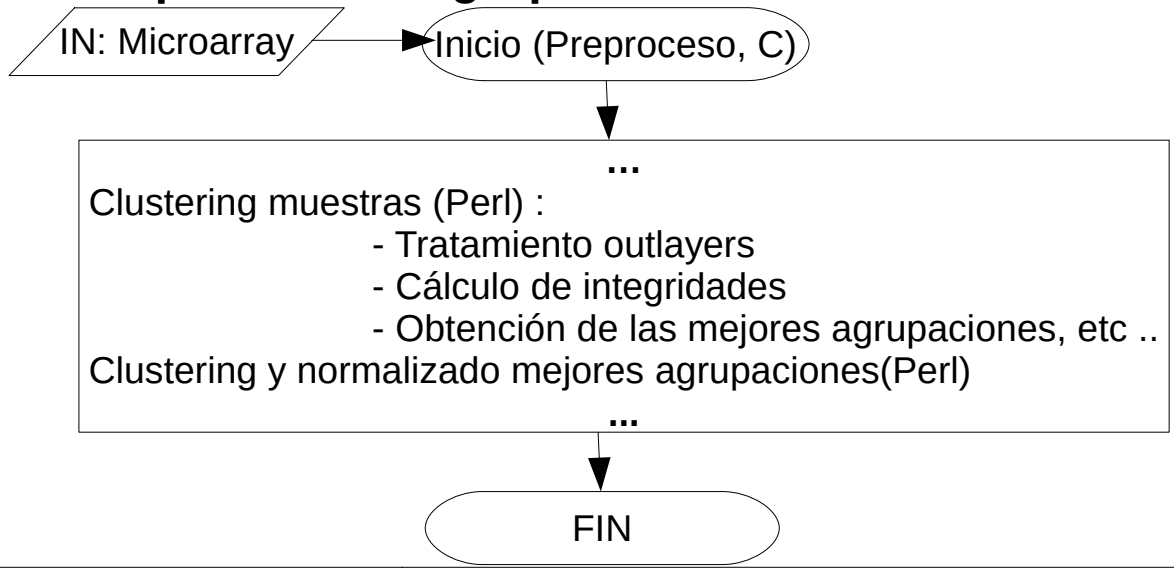
El preproceso es un conjunto de subprocesos que se ejecutan automáticamente al cargar una microarray en el sistema.

Solo se ejecuta una vez por microarray.

En este preproceso se añade el subproceso que realiza la agrupación de las condiciones muestrales de la microarray.

Debido a que el tiempo de ejecución es muy elevado se implementa una versión que solo realiza el proceso de agrupación

Diagrama de flujo:



Interfaz web: Integrar los resultados de las agrupaciones en el aplicativo web

Herramienta de desarrollo: PHP

Funcionalidades agregadas a la aplicación :

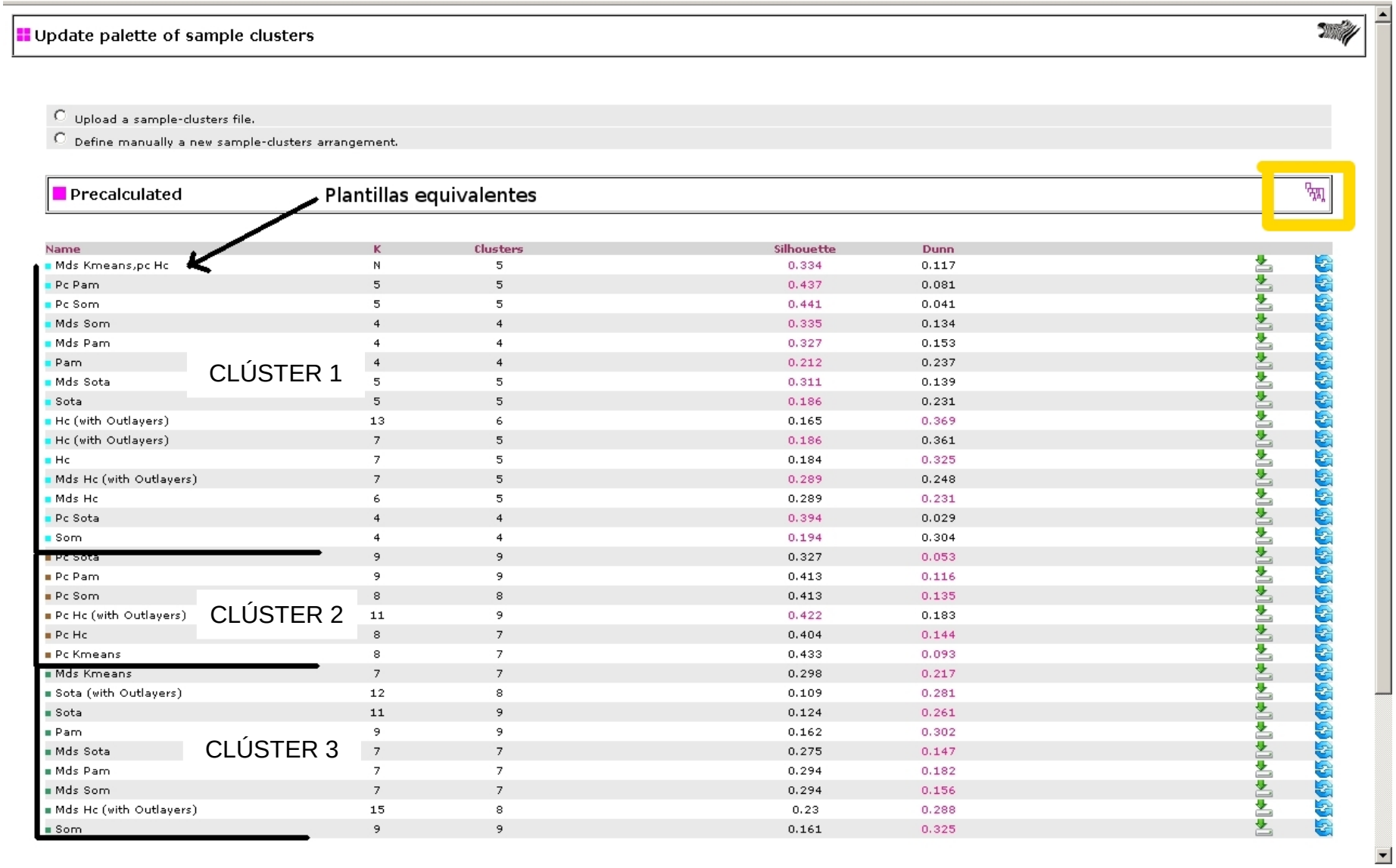
- Listado de las mejores agrupaciones ordenadas por similitud

- Actualización de la agrupación actual por la agrupación seleccionada por el usuario

- Descarga de la agrupación precalculada seleccionada por el usuario

- Gestor del histórico del usuario:
  - Guardar la agrupación actual con el nombre fijado por el usuario
  - Descargar la agrupación del histórico seleccionada
  - Actualizar la agrupación actual con la agrupación del histórico seleccionada
  - Eliminar uno o todas las agrupaciones del histórico
  - Normalizado histórico (clustering HC del histórico)

Interfaz web: Integrar los resultados de la agrupación en el aplicativo



Name	K	Clusters	Silhouette	Dunn
Mds Kmeans,pc Hc	N	5	0.334	0.117
Pc Pam	5	5	0.437	0.081
Pc Som	5	5	0.441	0.041
Mds Som	4	4	0.335	0.134
Mds Pam	4	4	0.327	0.153
Pam	4	4	0.212	0.237
Mds Sota	5	5	0.311	0.139
Sota	5	5	0.186	0.231
Hc (with Outlayers)	13	6	0.165	0.369
Hc (with Outlayers)	7	5	0.186	0.361
Hc	7	5	0.184	0.325
Mds Hc (with Outlayers)	7	5	0.289	0.248
Mds Hc	6	5	0.289	0.231
Pc Sota	4	4	0.394	0.029
Som	4	4	0.194	0.304
Pc Sota	9	9	0.327	0.053
Pc Pam	9	9	0.413	0.116
Pc Som	8	8	0.413	0.135
Pc Hc (with Outlayers)	11	9	0.422	0.183
Pc Hc	8	7	0.404	0.144
Pc Kmeans	8	7	0.433	0.093
Mds Kmeans	7	7	0.298	0.217
Sota (with Outlayers)	12	8	0.109	0.281
Sota	11	9	0.124	0.261
Pam	9	9	0.162	0.302
Mds Sota	7	7	0.275	0.147
Mds Pam	7	7	0.294	0.182
Mds Som	7	7	0.294	0.156
Mds Hc (with Outlayers)	15	8	0.23	0.288
Som	9	9	0.161	0.325

Búsqueda de los genes marcadores

Herramientas de desarrollo:

- C

Fundamentos teóricos:

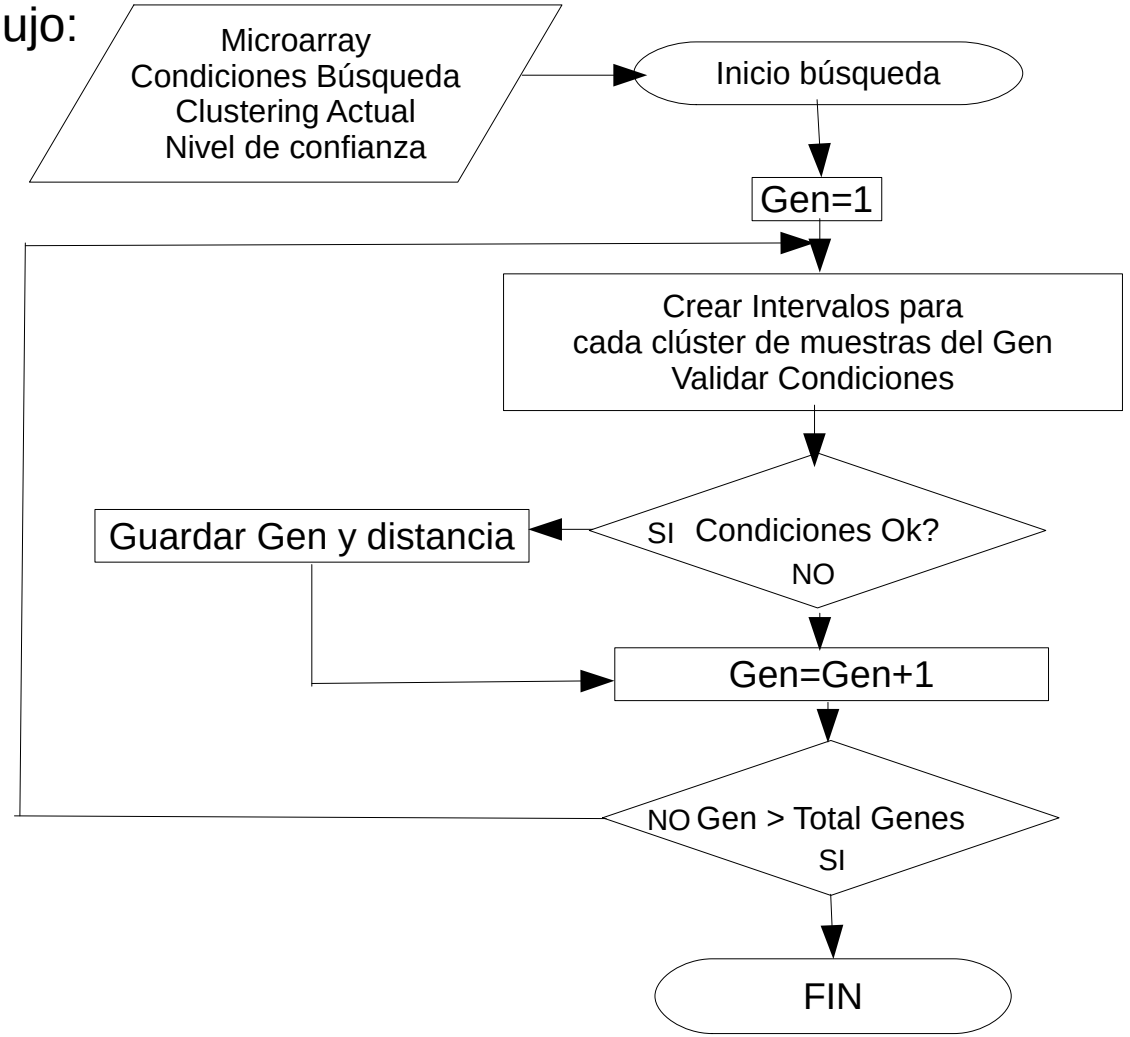
- Distribución T d Student
- Intervalos de confianza

Resultados:

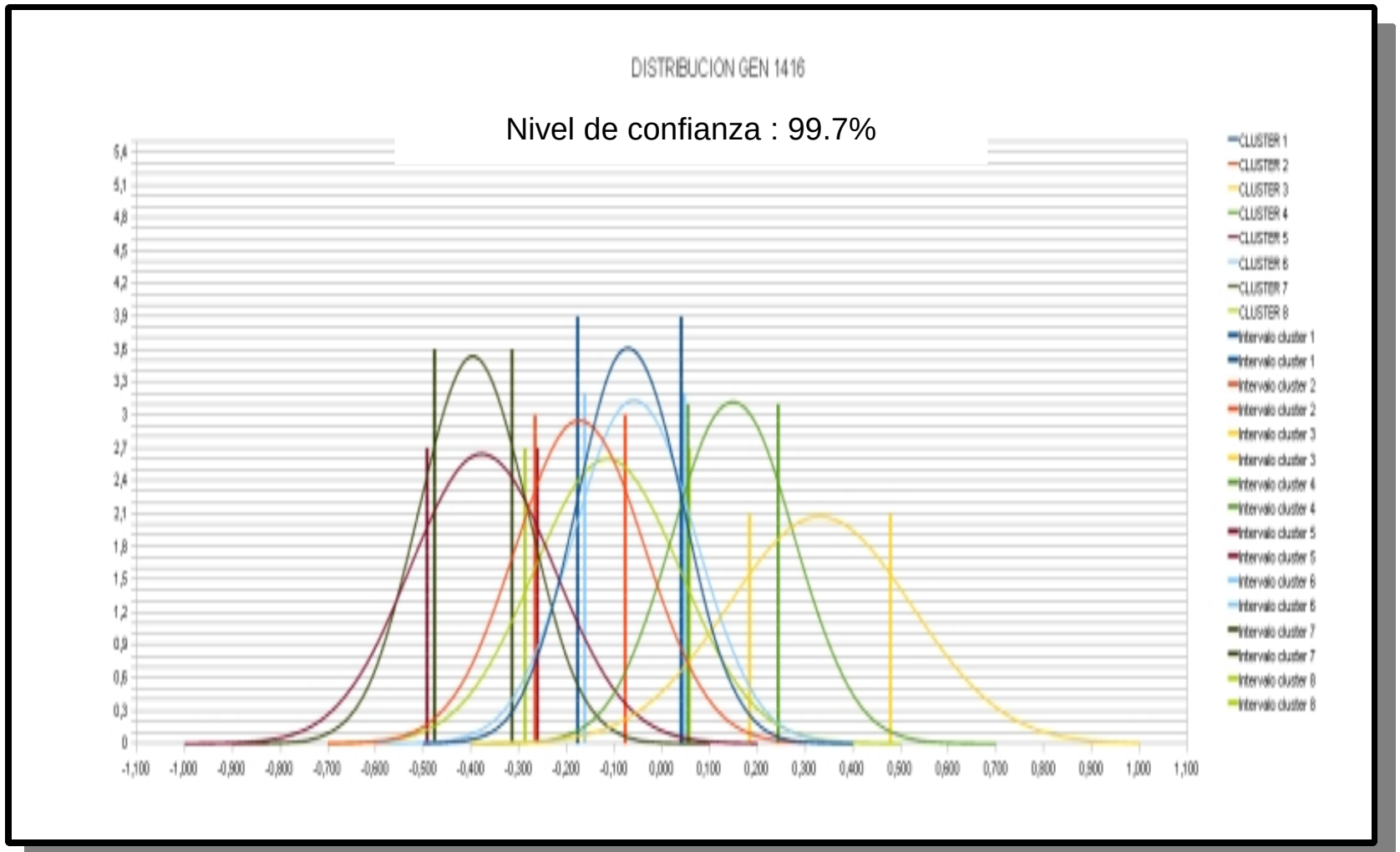
- Fichero con el identificador de los genes marcadores que cumplan las condiciones exigidas por el usuario para la agrupación actual y con la distancia total de los clúster validados de cada gen marcador.

Búsqueda de los genes marcadores

Diagrama de flujo:



Búsqueda de los genes marcadores para una distribución de clusters concreta



Interfaz web: Integrar la búsqueda de genes marcadores en el aplicativo web

Genes found

Rank	Dist	Id	Name
1	0.158306	927	LYZ: lysozyme (renal amyloidosis)
2	0.156075	962	EIF3EIP: eukaryotic translation initiation factor 3, subunit E interacting protein
3	0.154085	938	RREB1: ras responsive element binding protein 1
4	0.152776	953	HISPPD2A: histidine acid phosphatase domain containing 2A
5	0.152140	940	PCK2: phosphoenolpyruvate carboxykinase 2 (mitochondrial)
6	0.148792	968	ETV4: ets variant gene 4 (E1A enhancer binding protein, E1AF)
7	0.136902	961	ESTs Chr.22 [486514, (IW), 5º:AA043037, 3º:AA042937]
8	0.135707	949	CABC1: chaperone, ABC1 activity of bc1 complex homolog (S. pombe)
9	0.135451	924	ASNS: asparagine synthetase
10	0.133060	963	KIAA0430: KIAA0430
11	0.128942	950	AKAP1: A kinase (PRKA) anchor protein 1
12	0.128450	839	THEM4: thioesterase superfamily member 4
13	0.126250	943	IFI30: interferon, gamma-inducible protein 30
14	0.124413	934	ADD3: adducin 3 (gamma)
15	0.122691	926	SID 360210, ESTs, Weakly similar to !!!! ALU SUBFAMILY J WARNING ENTRY !!!! [H.sapiens] [5º:AA013089, 3º:AA013090]
16	0.120184	930	RPS16: ribosomal protein S16
17	0.118686	951	CKMT1B: creatine kinase, mitochondrial 1B
18	0.117606	902	USP8: ubiquitin specific peptidase 8
19	0.115052	964	UQCRH: ubiquinol-cytochrome c reductase hinge protein
20	0.114574	965	PDXK: pyridoxal (pyridoxine, vitamin B6) kinase
21	0.113416	937	RDH13: retinol dehydrogenase 13 (all-trans/9-cis)
22	0.111934	954	USP37: ubiquitin specific peptidase 37
23	0.110990	914	SID 75340, [5º:T57560, 3º:T57514]
24	0.110626	966	MARCKSL1: MARCKS-like 1

CONCLUSIONES

Los objetivos se marcados se han cumplido con creces incluso se han desarrollado nuevas funcionalidades

La consecución de los objetivos resulta una herramienta especialmente útil y práctica para los investigadores :

-<u>Útil:</u>

- **Análisis de los distintos estados celulares.**
- **Encontrar genes marcadores responsables de estos estados celulares.**

-<u>Práctica:</u>

- **Agrupaciones de condiciones muestrales pre calculadas**
- **Manipulación y almacenaje de estas agrupaciones en un histórico personal.**
- **Búsqueda automática de genes marcadores**

CONCLUSIONES

A nivel teórico una de las principales conclusiones que pueden extraerse es sobre los actuales índices de integridad:

- No son nada precisos para encontrar una única agrupación como la óptima.

- Ayudan a discriminar agrupaciones de entre todas las calculadas.

Aspectos positivos del desarrollo del proyecto :

- Aplicar conceptos teóricos, matemáticos y estadísticos al mundo .real

- Participar en un proyecto conjunto dedicado a la investigación de los genes como responsables de enfermedades como el cáncer

- http://revolutionresearch.uab.es : Web server for on line microarray analysis supported by the Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UAB).
- Huerta M, Cedano J, Querol E. (2008) Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach , J Bioinform Comput Biol. 6:367-386.
- Cedano J, Huerta M, Querol E. (2008) NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships, Adv Bioinformatics. 2008:789026. Epub 2008 Dec 10.
- Huerta M, Cedano J, Peña D, Rodriguez A, Querol E. (2009) PCOPGene-Net: holistic characterisation of cellular states from microarray data base on continuous and non-continuos analysis og gene-expression relationships BMC Bioinformatics. 2009 May 9;10:138.
- Delicado, P.(2001) Another look at principal curves and surfaces. Journal of Multivariate Analysis, 77, 84-116 .
- Delicado, P. and Huerta, M. (2003): 'Principal Curves of Oriented Points: Theoretical and computational improvements'. Computational Statistics 18, 293-315.
- Cedano J, Huerta M, Estrada I, Ballllosera F, Conchillo O, Delicado P, Querol E. (2007) A web server for automatic analysis and extraction of relevant biological knowledge. Comput Biol Med. 37:1672-1675.

GRACIAS POR SU ATENCIÓN

AGRADECIMIENTOS

A mi padre JOSÉ

A mi madre FILO

A mis HERMANOS

Al resto de mi familia

A Mario Huerta y Juan Antonio Cedano

Etc ...

Gracias a todos por vuestra paciencia y atención