Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria

Bioinformática:

Comparación de datos de expresión génica del servidor local con datos de una Base de Datos Remota

Por Marc Muñoz Escudero

Directors: Jordi Gonzàlez (CVC)
Mario Huerta (IBB)

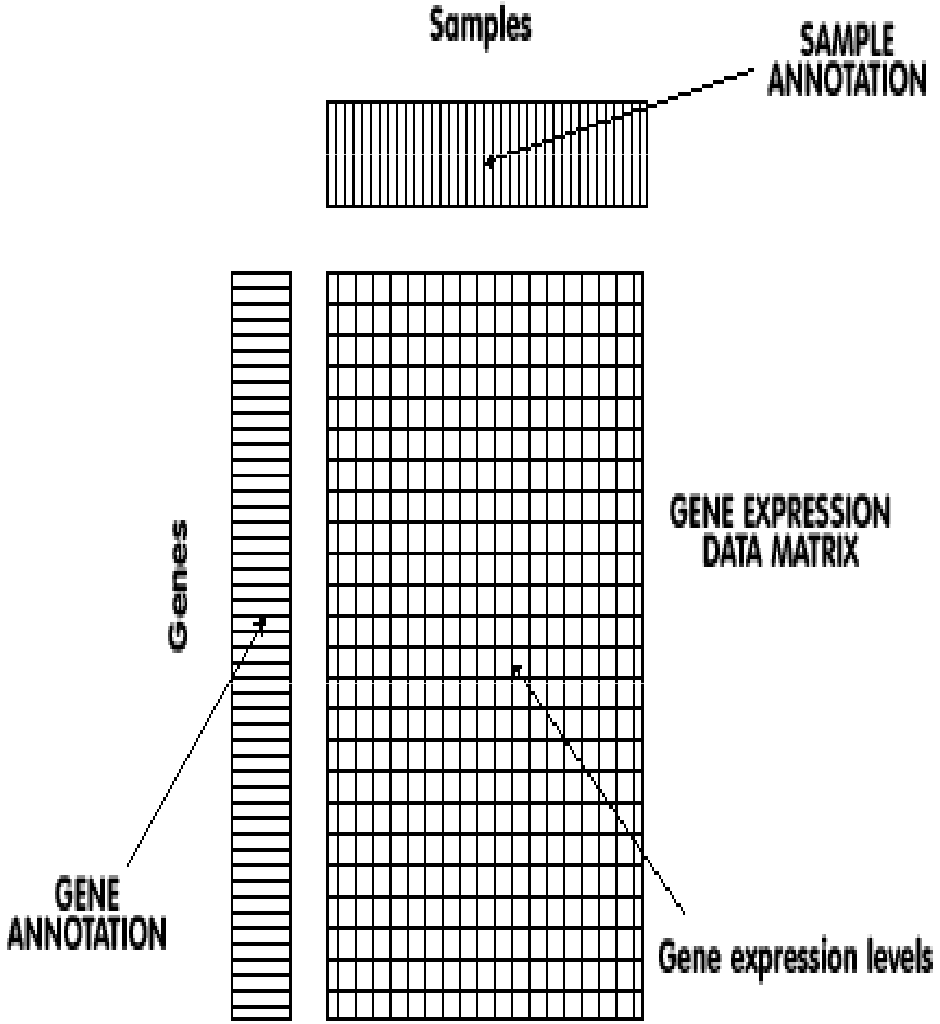
Índice

1. Introducción

El Instituto de Biotecnología y de Biomedicina (**IBB**) es un centro de investigación que forma parte de la Universidad Autónoma de Barcelona (**UAB**). En el IBB tienen una línea de investigación para el análisis de microarrays, que se desarrolla en el servidor de aplicaciones http://revolutionresearch.uab.es/.

2. Estado del arte

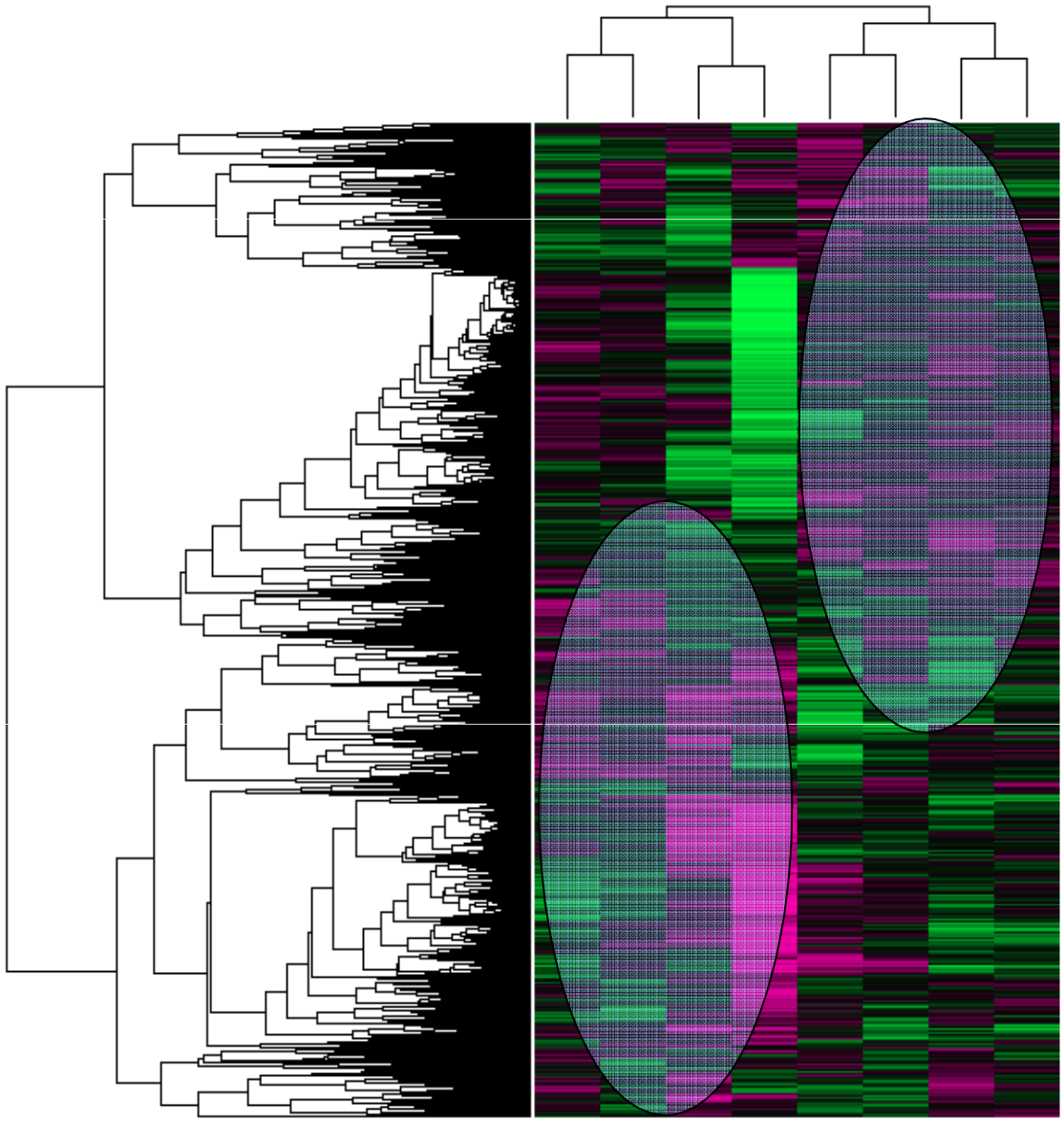
Microarray:



2. Estado del arte

Cluster:

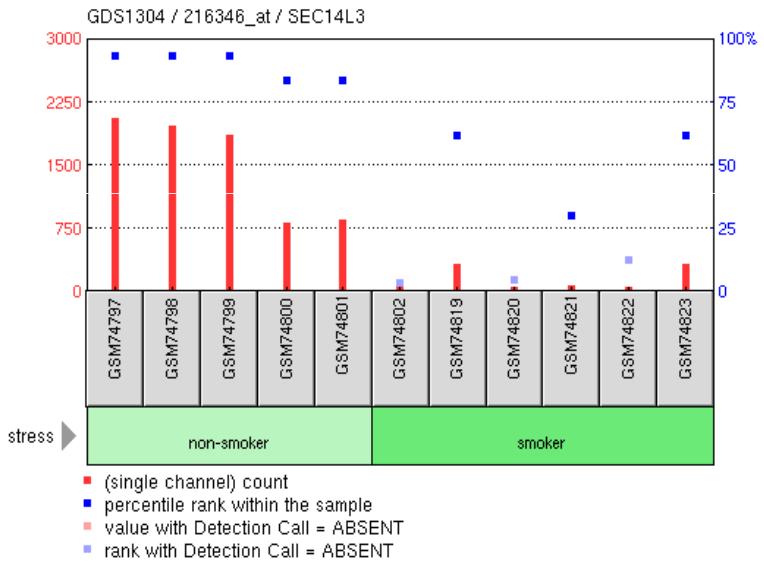
gds 3634:



2. Estado del arte

El “National Center for Biotechnology Information” (**NCBI**) es parte de la Biblioteca Nacional de Medicina de Estados Unidos. El NCBI es una importante fuente de información en biología molecular.

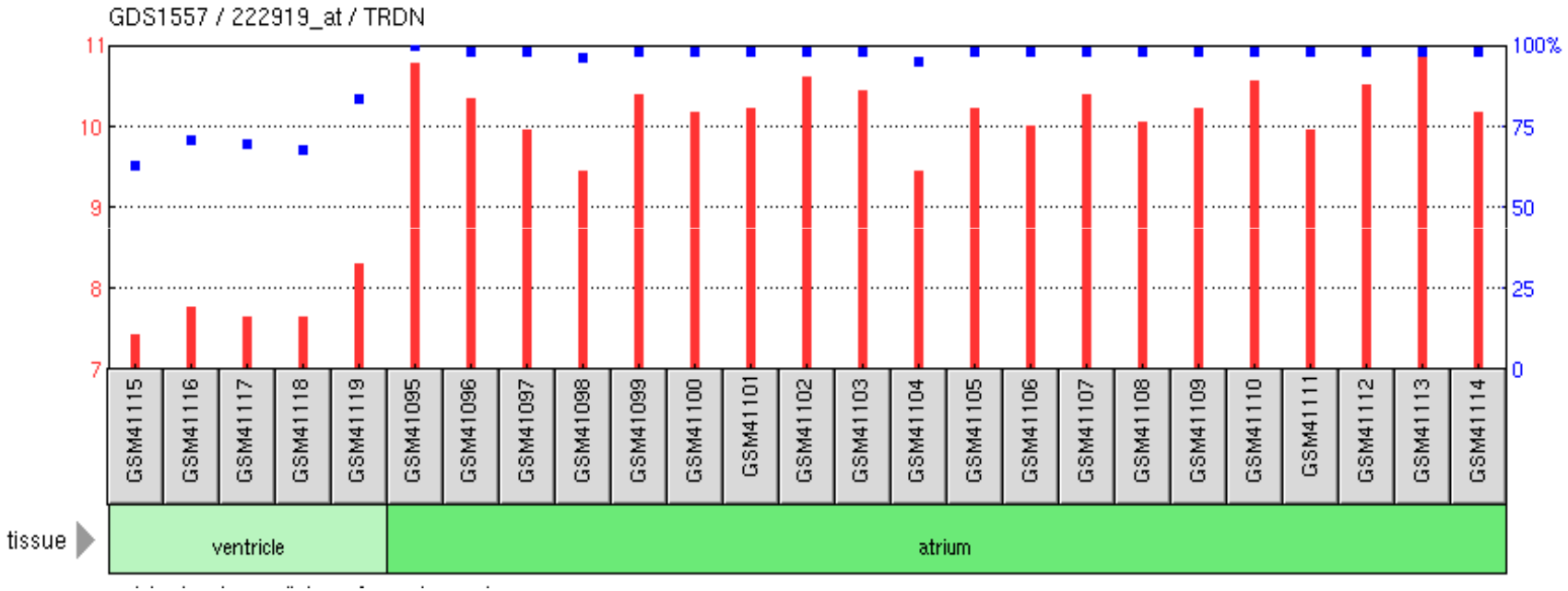
- **GEO Profiles**



2. Estado del arte

Genes marcadores:

Los genes marcadores son los genes de una microarray que se sobreexpresan en unas condiciones muestrales pero no en otras y por consecuencia, los genes que se sobreexpresan en unos clústers de condiciones muestrales pero no en otros.



3. Objetivos

- Enriquecer con información biomédica los clusters de condiciones muestrales de la microarray que el usuario este analizando en el servidor del IBB. Se enriquecerán los clusters de origen estadístico (o no) de la microarray del usuario a partir de cruzar los genes marcadores para estos clusters con la base de datos de genes marcadores para clusters basados en información biomédica de las microarrays del NCBI.

3. Objetivos

- Actualización periódica y automática de la base de datos local de genes marcadores.
- Cálculo en tiempo real de los genes marcadores comunes entre la microarray del usuario y la base de datos de genes marcadores de microarrays.
- **Interfaz web** para mostrar los resultados del cruce de los genes marcadores.

4. Desarrollo

Construcción de la BD de genes marcadores

1- Consulta a E-Utils (BD GEO Profiles), obtención de una lista de genes marcadores.
2- Consulta a E-Utils, obtener nombre, alias, especie, descripción,etc sobre cada gen marcador.
3- Tratar los genes marcadores descargados:
- Agrupar los genes marcadores por la microarray de la que son marcadores.
- Obtener el GeneID de cada gen marcador para posteriores consultas.

Resultado:

- 600mil genes marcadores
- Expresándose de forma diferenciada en 2510 microarrays (XML)

4. Desarrollo

Cruce online de genes marcadores

Consiste en buscar los genes marcadores que retorna la aplicación web para los clusters actuales de la microarray que está analizando el usuario, en la base de datos de genes marcadores.

Métodos de búsqueda:

- Búsqueda Simple
- Búsqueda por Alias
- Búsqueda por Filtro

Resultado:

- Listado de compatibilidad de microarrays
- Por cada microarray, los genes marcadores comunes

Tiempo de respuesta: 5 segundos

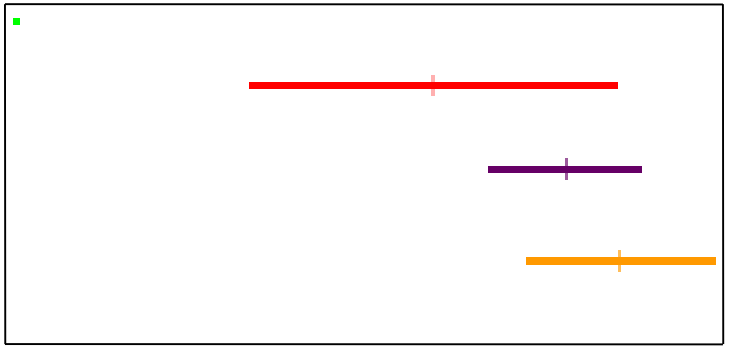
4. Desarrollo

Generación de imágenes

Consiste en generar una imagen por cada gen marcador donde se muestra los niveles de expresión del gen para cada cluster de la microarray. Se mostrarán en la vista detalle de la interfaz web.

Imagen:

- Microarray del NCBI
- Microarray del usuario



4. Desarrollo

UAB Universitat Autònoma de Barcelona

etse Escola Tècnica Superior d'Enginyeria

Genes found

Rank	Dist	Id	Name
1	0.195283	962	EIF3EIP: eukaryotic translation initiation factor 3, subunit E interacting protein
2	0.182004	961	ESTs Chr.22 [486514, (IW), 5º:AA043037, 3º:AA042937]
3	0.173826	968	ETV4: ets variant gene 4 (E1A enhancer binding protein, E1AF)
4	0.149907	966	MARCKSL1: MARCKS-like 1
5	0.144405	963	KIAA0430: KIAA0430
6	0.131102	953	HISPPD2A: histidine acid phosphatase domain containing 2A
7	0.130379	908	HBE1: hemoglobin, epsilon 1
8	0.124758	905	PSAT1: phosphoserine aminotransferase 1
9	0.104588	915	HBA2: hemoglobin, alpha 2
10	0.096631	919	SID W 296310, ESTs [5º:W03157, 3º:N74445]
11	0.081990	883	YARS: tyrosyl-tRNA synthetase
12	0.073853	912	PIBF1: progesterone immunomodulatory binding factor 1
13	0.071261	957	PIM3: pim-3 oncogene
14	0.065067	877	HIST1H1C: histone cluster 1, H1c
15	0.061815	952	NET1: neuroepithelial cell transforming gene 1
16	0.055803	955	CECR5: cat eye syndrome chromosome region, candidate 5

16 genes

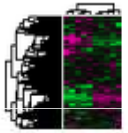
Use Gene Alias ☐ Filter by : Search Gene Markers

4. Desarrollo

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria

Matching GDS found

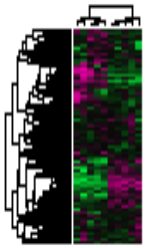
GDS	% user gds	% matching gds	Title & Summary	GDS Analysis	% gene markers
GDS1312	43.750000	0.664137	Squamous lung cancer *[Homo sapiens]* Expression profiling of squamous lung cancer biopsy specimens and paired normal specimens from 5 patients. Differentially expressed genes integrated with protein interaction maps. Results suggest that differentially expressed genes are highly connected through protein interactions. **Subsets: 2 disease state ,5 individual sets.**		4.730063
			Cigarette smoking effect on lung adenocarcinoma *[Homo sapiens]*		

Squamous lung cancer *[Homo sapiens]*

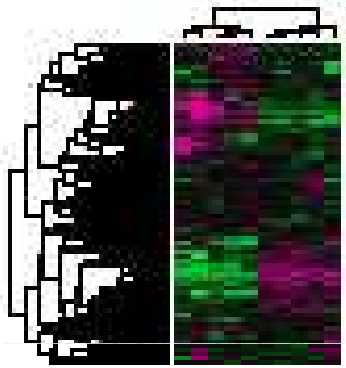
GDS1312 43.750000 0.664137

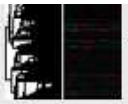
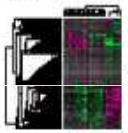
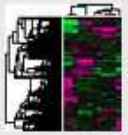
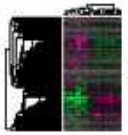
Expression profiling of squamo … aired normal specimens from 5 patients. Differentially expressed genes integrated wit … gest that differentially expressed genes are highly connected through protein inter …

Subsets: 2 disease state ,5 indi …



4.730063



GDS2214	37.500000	2.857143	Analysis of septic neutrophils treated … acute lung injury (ALI). HMGB1 is a l… (HMGB1) protein. Neutrophils isolated from patients with sepsis-induced … ay a role in endotoxemia-associated ALI. **Subsets: 3 agent ,8 individual se …**		0.942422
GDS1673	37.500000	2.083333	Non-diseased lung tissue *[Homo sapi …* Analysis of non-diseased lungs from … … ral sections of the lung were examined. Donors varied in age, sex, … many diseases. **Subsets: 2 tissue ,2 gender ,3 st …**		0.526749
GDS2499	12.500000	0.550964	Anti-cancer agent sapphyrin PCI-2050 effect on lung cancer cell line: dose response *[Homo sapiens]* Analysis of A549 lung cancer cells following treatment with anti-cancer agent sapphyrin PCI-2050 or transcription inhibitor actinomycin D. Hydrophilic sapphyrins localize to tumors, generate oxidative stress, and inhibit gene expression. **Subsets: 3 agent ,4 dose sets.**		0.663923
GDS1650	12.500000	1.886792	Pulmonary adenocarcinoma *[Homo sapiens]* Analysis of pulmonary adenocarcinomas (AC). Carcinogen exposure is responsible for the majority of ACs. Results compared with those obtained from a urethane-induced lung tumor model in the mouse (GDS1649), and provide insight into the conserved pathways underlying the development of AC. **Subsets: 2 tissue sets.**		0.839604

7 matching gds with gds-user % > 10% & 13 matching gds with gds-user % < 10%

Show/Hide % user-gds < 10%

4. Desarrollo

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria

Matching gene markers GDS881: Breast cancer and selective estrogen receptor modulators

Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
RAB31: RAB31, member RAS oncogene family		0.103162	
RAB31: RAB31, member RAS oncogene family		0.103162	
ID2: inhibitor of DNA binding 2, dominant negative helix-loop-helix protein		0.089515	
PRDX1: peroxiredoxin 1		0.073552	
HLA-B: major histocompatibility complex, class I, B		0.055599	
GNAS: GNAS complex locus		0.046373	
UBB: ubiquitin B		0.035885	
UBB: ubiquitin B		0.035885	

8 genes matched & 442 genes mismatched

Subsets: time agent

Show/Hide missmatched genes

4. Desarrollo

Actualización de la BD de genes marcadores

Consiste en:

- Cada 2 meses (programado en el cron)
- Nuevos genes marcadores
- Generación de imágenes

5. Conclusiones

- La creación de la base de datos de genes marcadores de microarrays en el servidor local y su actualización periódica se ha conseguido con éxito.
- El cruce online que se realiza para buscar genes marcadores comunes entre la microarray del usuario y la base de datos de microarrays proporciona su respuesta en un tiempo récord.
- La interfaz web realizada proporciona un aplicativo altamente usable, entendible y con una alta operativilidad.
- La aplicación web implementada logra cumplir con el objetivo marcado de permitir el cruce de los genes marcadores de la microarray de estudio con la base de datos de genes marcadores de microarrays.

6. Bibliografía

- Delicado, P.(2001) Another look at principal curves and surfaces. Journal of Multivariate Analysis, 77, 84-116 PCOP theoretical definition
- Delicado, P. and Huerta, M. (2003): 'Principal Curves of Oriented Points: Theoretical and computational improvements'. Computational Statistics 18, 293-315 PCOP theoretical and computacional Improvements
- 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.Pattern analysis, clustering and new-sample classification based on PCOP
- 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 Navigation trhough lineal and non-lineal gene-expression relationships
- Cedano J, Huerta M, Querol E. (2008) NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships. Advances in Bioinformatics, vol. 2008.Navigation trhough non-continuous gene-expression relationships
- 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 on gene-expression relationships. BMC Bioinformatics., 9;10:138 Microarray interactive gene networks