Bioinformática : Base de datos de matrices de expresión génica para su análisis vía web

Universitat Autònoma de Barcelona

escola d'enginyeria

Daniel Sánchez Santolaya
Tutores: Mario Huerta(IBB)
Jordi Gonzàlez(CVC)

Índice

1. Introducción

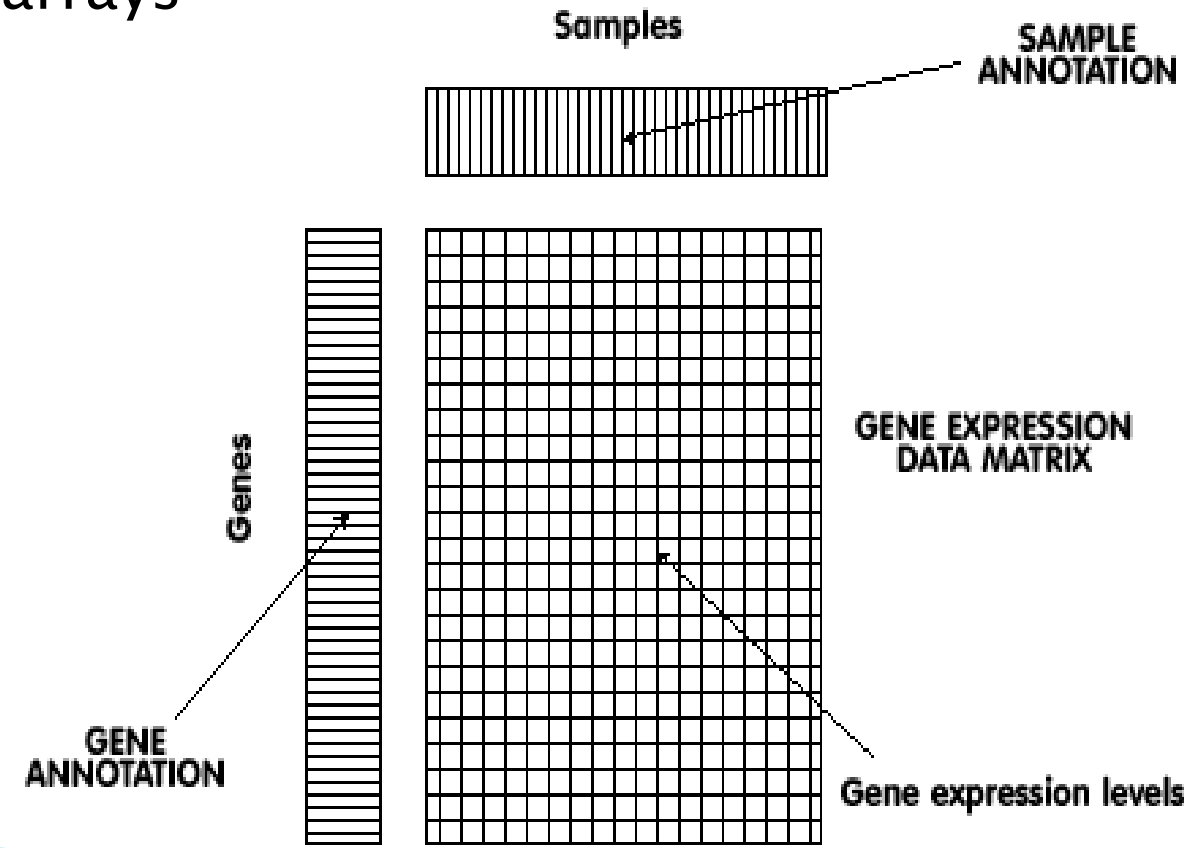
- Motivación
  - Aplicar mis conocimientos para ayudar en investigaciones científicas como la cura o tratamiento de enfermedades.
  - Oportunidad para realizar un proyecto de una aplicación real y en un centro de investigación real.

1. Introducción

- Estado del arte
  - Los genes al expresarse sintetizan las diferentes proteínas.
  - Las proteínas sintetizadas llevan a cabo diferentes funciones de la célula.
  - Los genes al expresarse determinan el estado celular.
  - Modificando la expresión génica se provoca un cambio celular que puede llevar de la salud a la enfermedad o viceversa.
  - El estudio de la expresión de los genes puede ayudarnos a salvar muchas vidas.

1. Introducción

- Estado del arte
  - Microarrays

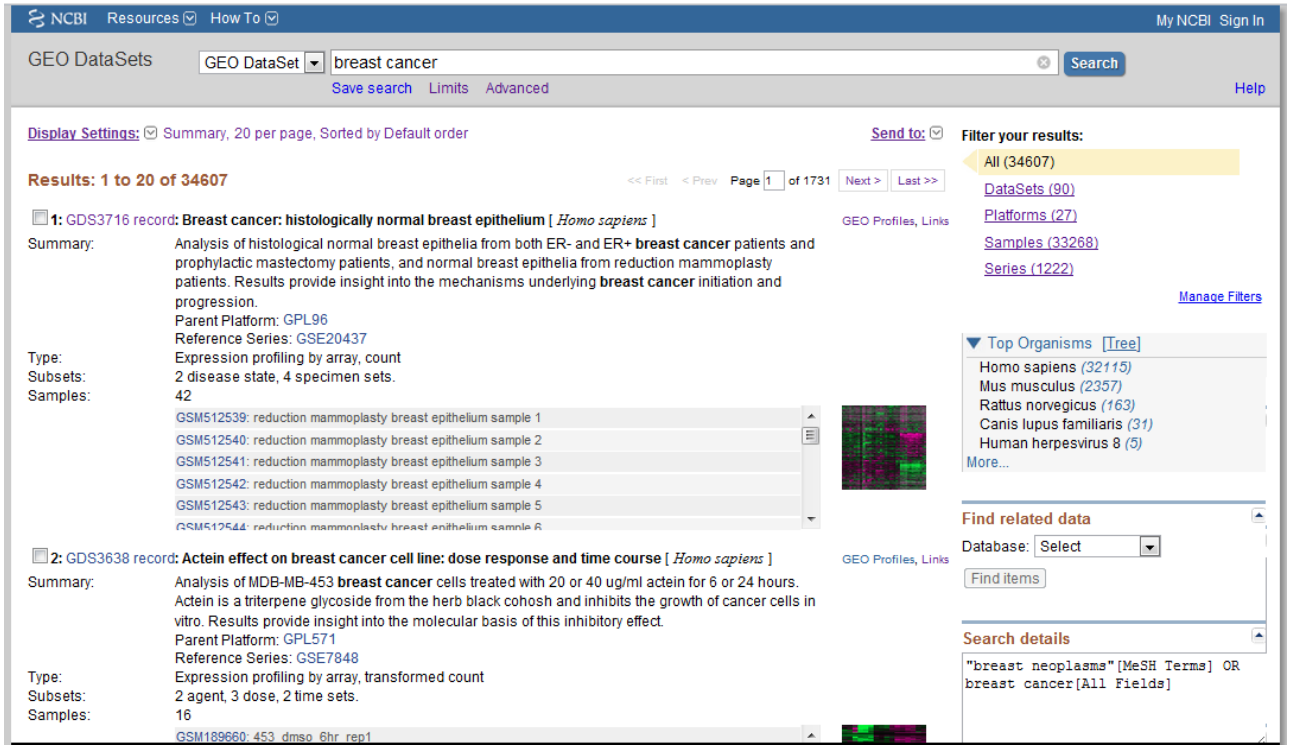


1. Introducción

- Estado del arte
  - Aplicación web en: http://revolutionresearch.uab.es/ para el análisis de microarrays
  - Problema: las microarrays se han de subir manualmente por los usuarios -> Pocas microarrays

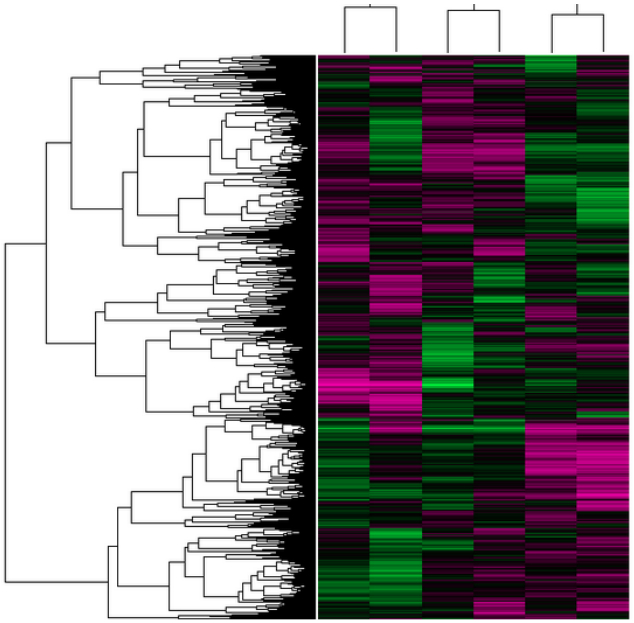
1. Introducción

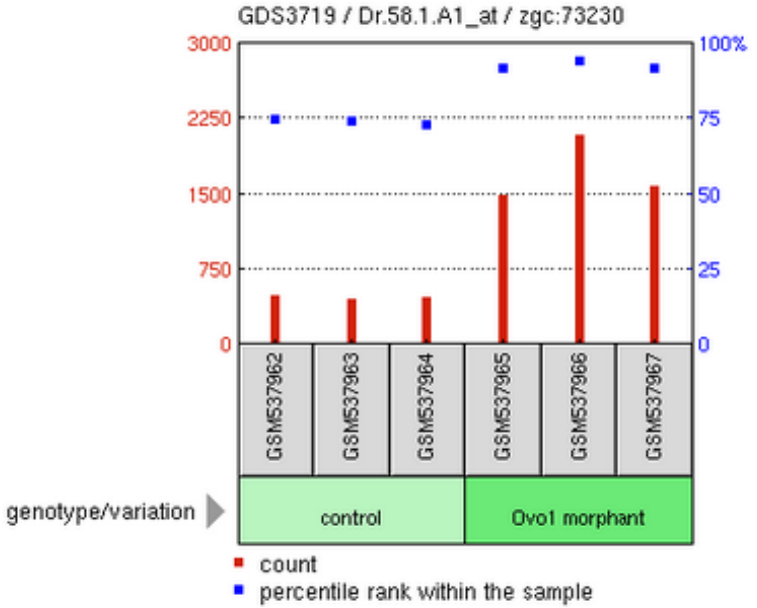
- Estado del arte



1. Introducción

- Estado del arte





2. Objetivos

- El objetivo principal es ampliar la base de datos de microarrays a partir de las microarrays del NCBI con un gran número de condiciones experimentales.
  - Actualización periódica y automática de la base de datos local de microarrays con las microarrays públicas de gran tamaño
  - Interfaz web para gestionar las nuevas microarrays

2. Objetivos

- Actualización periódica y automática de la base de datos local de microarrays con las microarrays públicas de gran tamaño
  - Identificar las nuevas microarrays de gran tamaño del NCBI.
  - Descargar y parsear los ficheros de las microarrays para que se adecuen al formato de las microarrays del servidor local. Tanto datos, como genes.
  - Subir a la base de datos local las microarrays descargadas y parseadas.
  - Realizar la actualización de manera que si los genes de la microarray cambian de nombre, estos puedan ser actualizados por el robot actualizador de nombres de genes que actualmente hay en el servidor.

2. Objetivos

- Actualización periódica y automática de la base de datos local de microarrays
  - La actualización ha de ser robusta a posibles errores o a la caída del servidor durante el proceso.
  - Eliminación de directorios y ficheros temporales que no son necesarios tras finalizar el proceso de actualización.
  - Realizar la actualización de manera periódica y sincronizada con el robot de actualización de genes marcadores de microarrays que existe actualmente en el servidor.

2. Objetivos

- Interfaz web para gestionar las nuevas microarrays
  - Adaptar la aplicación web para que las operaciones que se hacían anteriormente con las microarrays subidas por los usuarios se puedan realizar también con las microarrays subidas por la actualización automática.
  - Permitir a los usuarios realizar búsquedas de las microarrays públicas de gran tamaño insertando el tema de la microarray y/o especie sobre la que se realizaron los experimentos.
  - Mantener un listado de las microarrays públicas favoritas del usuario, de manera que le permitirá un acceso rápido a ellas.

3. Fases

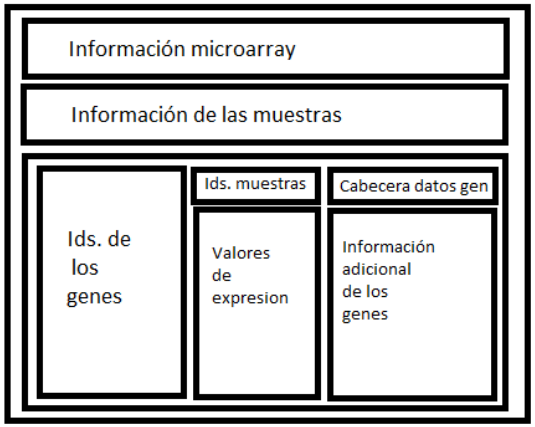
3.1 Adquirir conocimientos sobre la bioinformática y el ámbito del proyecto

- Adquirir conocimientos sobre la bioinformática
- Adquirir conocimientos necesarios sobre la aplicación web y las bases de datos del servidor local
  - Base de datos de microarrays
  - Microarray:
    - Fichero con los valores de expresión(samples)
    - Fichero con los nombres de las condiciones experimentales(snames)
    - Fichero con los nombres de los genes(genesorig)
    - Fichero con los nombres de los genes actualizados(genes)
  - Robot actualizador de nombres de gen
  - Robot de descarga de genes marcadores

3.1 Adquirir conocimientos sobre la bioinformática y el ámbito del proyecto

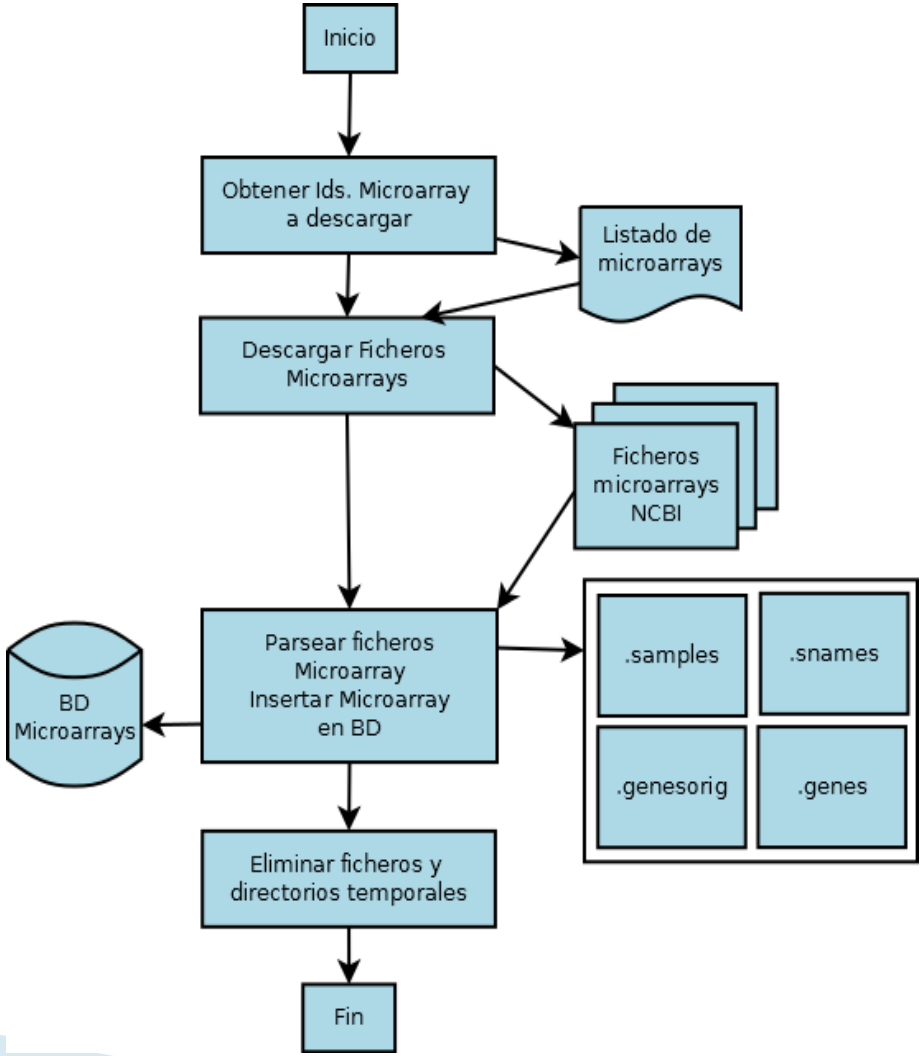
- Adquirir conocimientos necesarios sobre el entorno NCBI
  - Analizar los formatos y los métodos de descarga de las microarrays en GEO Datasets.
    - Hay que encontrar la mejor manera para descargar las microarrays

GDS Clustering

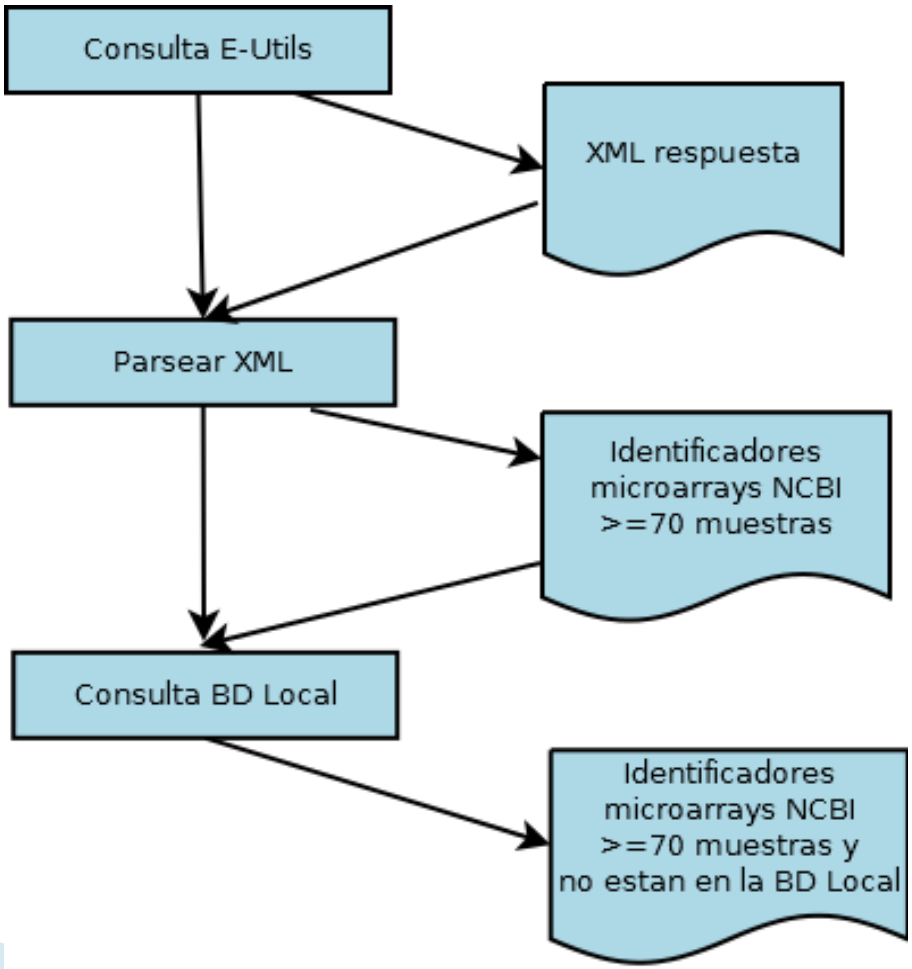


GDS Full

3.2 Crear robot de descarga de microarrays públicas de gran tamaño del NCBI.



3.2.1 Identificar las nuevas microarrays públicas de gran tamaño a descargar del NCBI

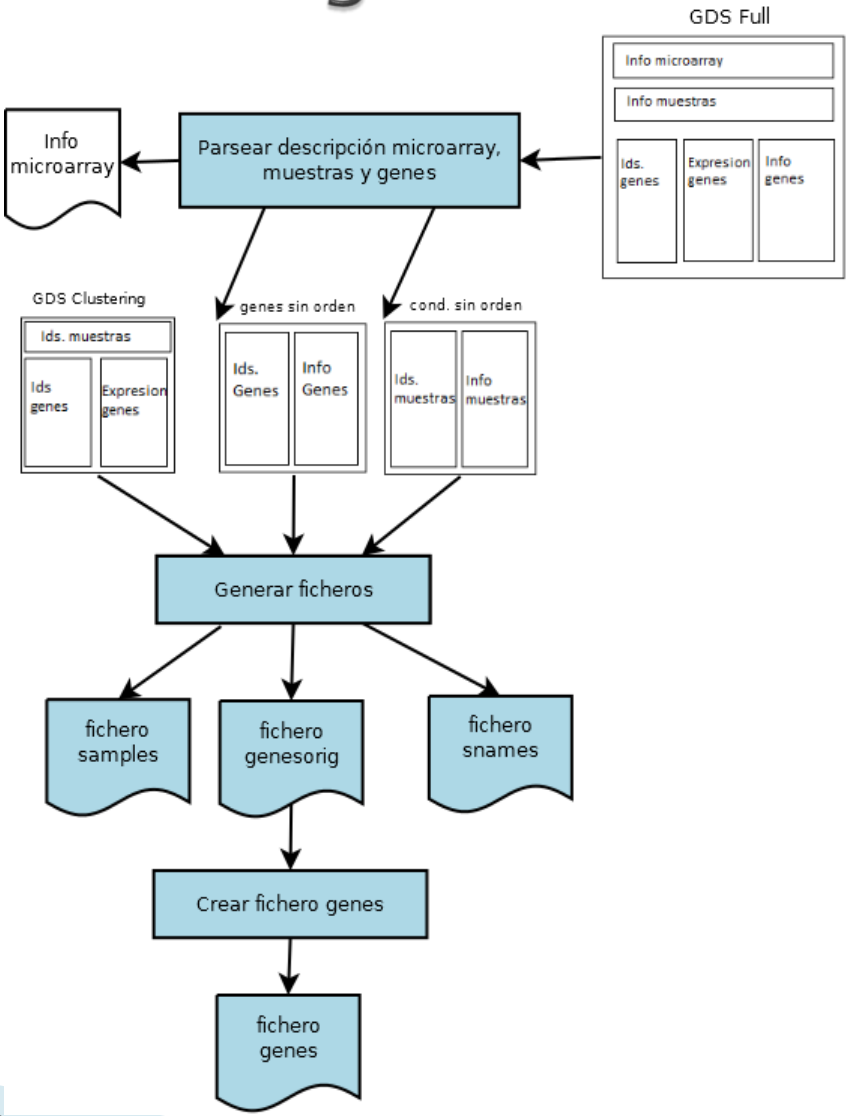


3.2.2 Descargar los ficheros de la microarrays nuevas del NCBI al servidor

- Ficheros GDS Full

- Ficheros GDS Clustering

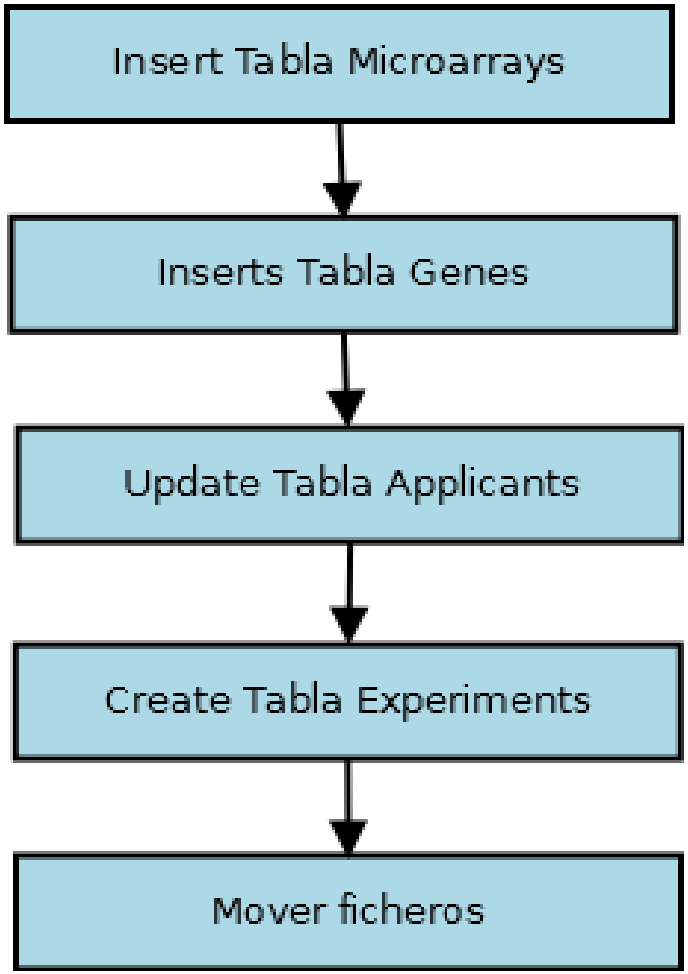
3.2.3 Parsear los ficheros de las microarrays descargadas



3.2.3 Parsear los ficheros de las microarrays descargadas

- Puntos clave:
  - Parsear la información que describe la microarray.
    - Se necesita obtener su descripción para posteriormente poder realizar búsquedas por palabra clave y especie
    - Es necesario crear una tabla con la relación entre nombres y códigos de especies
  - Parsear los genes para que puedan ser actualizados por el robot actualizador de nombres
    - Coger códigos de secuencia y códigos Unigene

3.2.4 Subir las microarrays a la base de datos local



3.2.5 Control de errores, recuperación tras caída del servidor en el proceso de actualización y eliminación de directorios y ficheros temporales

- Durante la actualización podría producirse una caída del servidor:
  - Problema:
    - Podría quedarse una microarray subida incompletamente
  - Solución:
    - Comprobar si la última microarray no se ha subido completamente
    - Si tenemos el registro en la tabla microarrays pero no están los ficheros movidos:
      - Eliminamos los registros de esa microarray y se subirán en la próxima actualización.
  - Ejecutado al iniciarse el servidor

3.3 Sincronización con el robot de genes marcadores y programación periódica de ambos robots

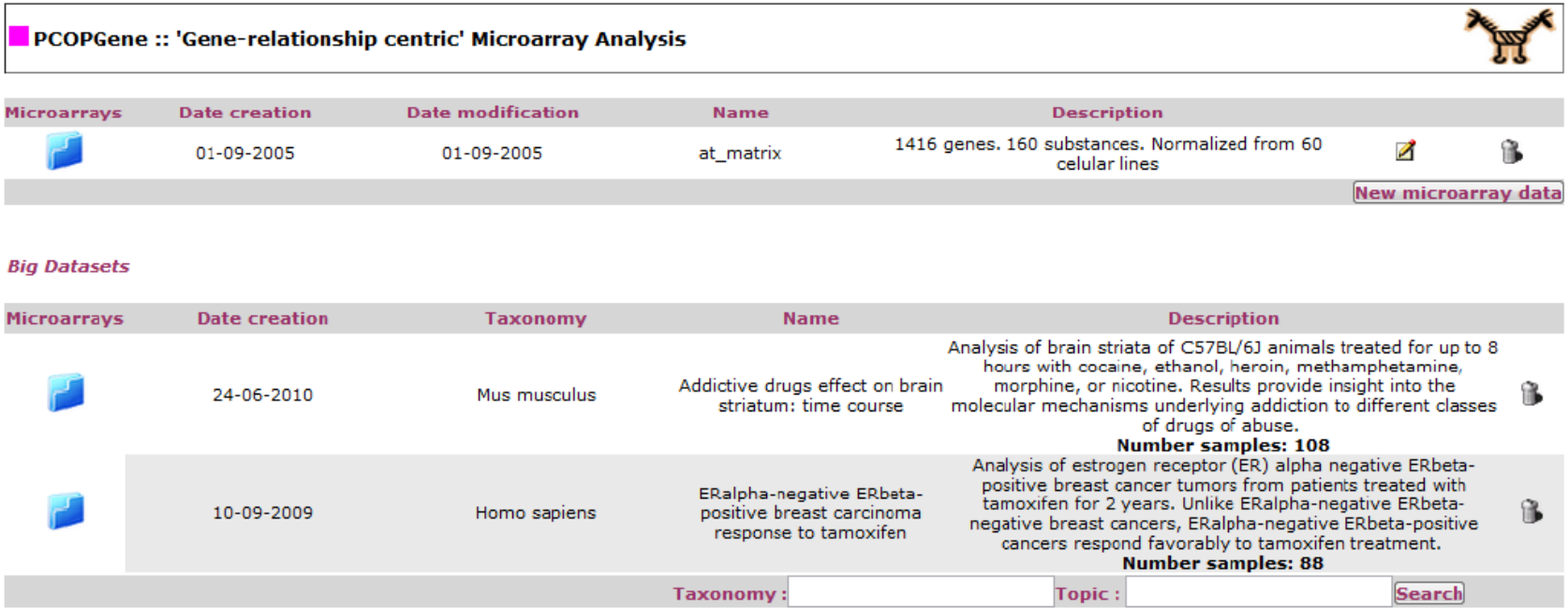
- Se ejecutan secuencialmente sin tener ningún conflicto en los directorios.
- Se ha utilizado el Cron de Linux para programar la ejecución

3.4 Aplicación web

- Crear nueva interfaz para la búsqueda y gestión de microarrays públicas de gran tamaño
- Realizar todos los cambios necesarios para que la aplicación web actual funcione con las microarrays públicas de gran tamaño de la misma manera que funciona actualmente con las microarrays subidas por los usuarios.

3.4.1 Crear nueva interfaz para la búsqueda y acceso de microarrays de gran tamaño

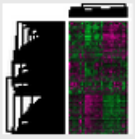
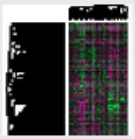
- Búsqueda por palabra clave y especie y listado de microarrays públicas favoritas



3.4.1 Crear nueva interfaz para la búsqueda y acceso de microarrays de gran tamaño

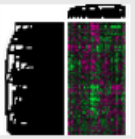
- Resultado búsqueda

GDS found

GDS Id	Date	Taxonomy	Title & Summary	GDS Analysis	Add
GDS3116	12-12-2008	Homo sapiens	Letrozole effect on breast cancer tumors Analysis of breast cancer tumors following treatment with letrozole for 14 days. The aromatase inhibitor letrozole is an anti-estrogen drug used to treat postmenopausal women with breast cancer. Results provide insight into the molecular mechanism of action of letrozole in breast cancer. **Number samples: 116**		
GDS1627	24-08-2006	Homo sapiens	Breast cancer cell lines response to chemotherapeutic drugs: time course Analysis of 4 breast cell lines treated for up to 36 hrs with the chemotherapeutic agents 5-fluorouracil (5FU), doxorubicin (DOX), or etoposide (ETOP), a drug mechanistically similar to DOX. Expression profiles for DOX- and 5FU-treated cells were used to successfully predict the response to ETOP. **Number samples: 83**		

1

GDS found in user list

GDS Id	Date	Taxonomy	Title & Summary	GDS Analysis
GDS2827	10-09-2009	Homo sapiens	ERalpha-negative ERbeta-positive breast carcinoma response to tamoxifen Analysis of estrogen receptor (ER) alpha negative ERbeta-positive breast cancer tumors from patients treated with tamoxifen for 2 years. Unlike ERalpha-negative ERbeta-negative breast cancers, ERalpha-negative ERbeta-positive cancers respond favorably to tamoxifen treatment. **Number samples: 88**	

<< Back

3.4.2 Realizar todos los cambios para que la aplicación web funcione con las microarrays públicas de gran tamaño de la misma manera que funcionan con las microarrays subidas por los usuarios.

- Se han debido realizar diferentes cambios, principalmente debido a que las microarrays nuevas tienen un tamaño muy superior

4. Conclusiones

- Objetivos relacionados con la actualización periódica y automática de la base de datos local de microarrays
  - Se identifcan las nuevas microarrays de gran tamaño publicadas en la base de datos GEO Datasets del NCBI.
  - Se descargan y se parsean los ficheros de las microarrays del NCBI de manera que se adaptan al formato de las microarrays del servidor local

4. Conclusiones

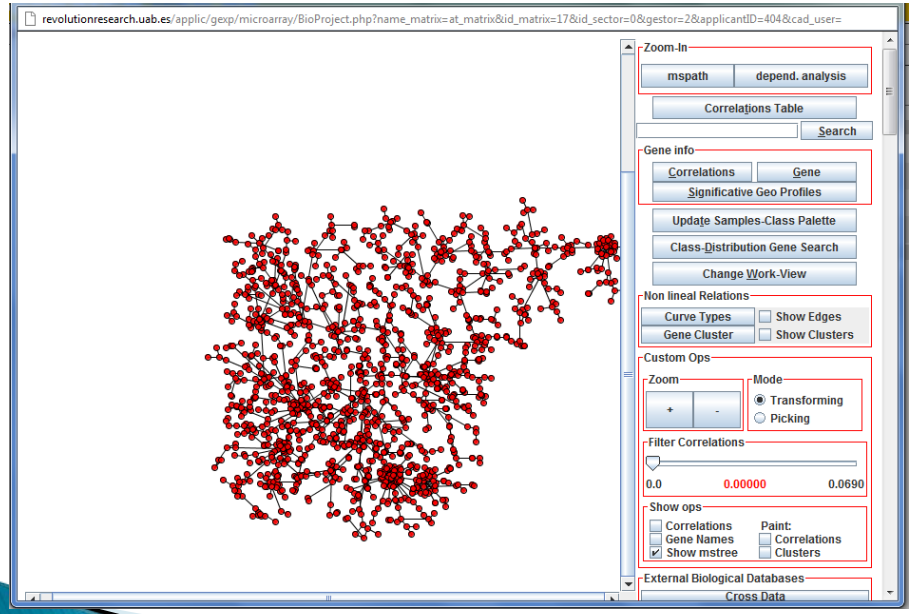
- Objetivos relacionados con la actualización periódica y automática de la base de datos local de microarrays
  - La actualización se realiza de manera que los genes puedan ser actualizados por el robot actualizador de nombres de gen
  - La actualización es robusta a posibles errores o la caída del servidor.
  - La actualización se realiza de manera periódica y sincronizada con el robot de genes marcadores.

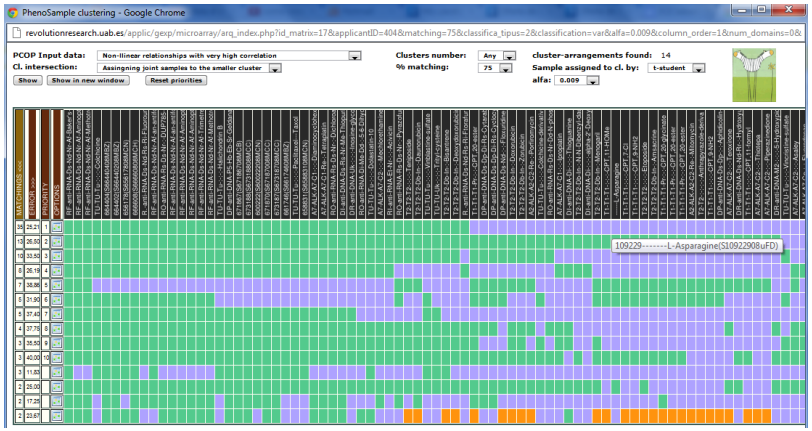
4. Conclusiones

- Objetivos relacionados con la interfaz web para gestionar las nuevas microarrays
  - Se ha creado la Interfaz web para realizar búsquedas por palabra clave y/o especie
  - Se ha creado la Interfaz web con el listado de microarrays públicas que el usuario considera de interés.
  - Se ha adaptado la aplicación web para poder realizar los mismos análisis y operaciones de gestión con las nuevas microarrays, una vez estas pasan a la lista de favoritas.

5. Trabajos futuros

- Realizar el preproceso de análisi de http://revolutionresearch.uab.es/ para realizarlo con las nuevas microarrays.





6. Bibliografía

http://revolutionresearch.uab.es : A web server for on-line microarray analysis supported by t Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UA

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 automati analysis and extraction of relevant biological knowledge. Comput Biol Med. 37:1672-1675.

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 Eff on Gene-Expression Relationships Advances in Bioinformatics, vol. 2008.

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.