Universitat Autònoma de Barcelona

escola d'enginyeria

ENGINYERIA INFORMÀTICA

MEMÒRIA PREVIA DEL PROJECTE:

4507 BIOINFORMÁTICA: BASE DE DATOS DE MATRICES DE EXPRESIÓN GÉNICA PARA SU ANÁLISIS VÍA WEB

Signatura de l'estudiant	Signatura del director/a o directors/es
Nom: Daniel Sánchez Santolaya Data:	Nom/s: Jordi Gonzalez Sabat Mario Huerta Dpt: CVC Data:

1. Objetivo del proyecto

En el Instituto de Biotecnología y de Biomedicina(IBB)[1] de la Universidad Autónoma de Barcelona(UAB) existe una base de datos de microarrays que suben los usuarios vía web, así como unas herramientas que las analizan en un proceso background una vez subidas y una interfaz web para poder gestionarlas y visualizar el resultado de los análisis. Estas herramientas se encuentran en un servidor llamado revolutionresearch[2].

La tecnología de microarrays nos permite observar el nivel de expresión de un número grande de genes bajo un número de circunstancias diferentes. De esta manera, permiten estudiar el comportamiento de los genes en diversas condiciones.

El objetivo principal del proyecto es ampliar la base de datos de microarrays a partir de microarrays con un gran número de muestras publicadas en el panorama internacional, de manera que esta se vaya actualizando periódicamente y automáticamente para mantener la base de datos siempre con las últimas publicadas. Como se ha comentado anteriormente, en el servidor revolutionresearch existen unas potentes herramientas para el análisis de microarrays, por lo que para sacar información realmente útil habrá que lanzar los preprocesos de análisis para las nuevas microarrays que se incorporen. Por último, para poder ver la información de estos análisis actualmente hay una interfaz web. Esta interfaz web deberá ser ampliada, de manera que permita a los usuarios acceder a las microarrays incorporadas de forma automática, realizar búsquedas por diversos campos y ofrecer la posibilidad de que el usuario incluya ciertas microarrays en su listado de favoritas, lo cual le permitirá un acceso más rápido a ellas. Esta interfaz web se ha de realizar de manera que el usuario pueda usar sus funciones de manera cómoda y sencilla.

De esta manera, al final del proyecto, los usuarios de la aplicación web verán que la información disponible se ha potenciado ampliamente, ya que podrán

ver los análisis de una gran cantidad de microarrays, además de las microarrays por ellos subidas.

2. Breve introducción al estado del arte del tema propuesto

Este proyecto está situado en el ámbito de la bioinformática. La bioinformática es una disciplina científica en el que se incluyen varios campos como la biología, la computación y las tecnologías de la información.
Concretamente, el proyecto está centrado en ofrecer herramientas a los investigadores para estudiar el comportamiento de la expresión de genes en diversas condiciones. Los genes al expresarse sintetizan las diferentes proteínas. Son estas proteínas sintetizadas las que realizan las diferentes funciones de la célula. Por esta razón, al expresarse los genes determinan el estado celular y modificando esta expresión génica, se puede provocar un cambio celular. Esto es muy importante, ya que estos cambios celulares pueden provocar el llevar a una persona de la salud a la enfermedad o al contrario. Por lo tanto, el estudio de la expresión de los genes nos puede ayudar a entender las causas de porqué se provocan ciertas enfermedades, o a encontrar que causas pueden producir la cura, de manera que puede ayudar a salvar muchas vidas.

Este estudio de los genes se puede realizar mediante la tecnología de microarrays. Los datos generados por las microarrays se pueden ver como una inmensa matriz, dónde las filas representan cada uno de los genes analizados, las columnas cada una de las condiciones a las que se han sometido los genes, y finalmente los valores de la matriz serán los niveles de expresión.

Estas matrices contienen una gran cantidad de información (miles de genes por cientos de condiciones muestrales), por lo que es necesario realizar análisis previos para extraer información útil para el investigador. En el servidor del IBB revolutionresearch se realizan algunos de estos análisis. Ejemplos de estos análisis es la aplicación a las microarrays de algoritmos de agrupamiento o clustering, los cuales agrupan los genes que se sobreexpresan en

determinadas condiciones experimentales pero no en otras, o el análisis de relaciones lineales o no lineales entre 2 genes. Estos análisis son los que proporcionan información realmente útil a los investigadores.

Una cuestión importante es la obtención de las nuevas microarrays que generen científicos de todo el mundo. En el proyecto esto se realizará a partir de la base de datos GEO (Gene Expression Omnibus) Datasets del NCBI(National Center for Biotechnology Information)[3]. El NCBI es parte de la Biblioteca Nacional de Medicina de Estados Unidos, lo cual nos servirá como una importante fuente de información, ya que se publican las últimas microarrays en el panorama internacional.

El proyecto deberá integrarse en el servidor revolutionresearch del IBB, concretamente con la aplicación PCOPGene[4][5][6][7][8][9] la cual permite realizar análisis i gestionar microarrays subidas por los usuarios.

3. Estudio de viabilidad del proyecto

Tanto la base de datos de microarrays, como la integración con los preprocesos de análisis y la interfaz web correrán sobre el servidor del IBB revolutionresearch. En este servidor se puede programar con algunos lenguajes, como Perl, XML, PHP, SQL y C, que son los que se utilizaran principalmente para implementar el proyecto.

En este servidor ya nos encontramos con que existe una base de datos de microarrays. Las aplicaciones de este servidor hacen uso de esta base de datos, por lo que nuestra base de datos de microarrays obtenidas del NCBI se deberá integrar con esta. Esta base de datos es relacional. Una vez realizado el primer análisis, se ha observado que en la base de datos existe una tabla de genes, la qual por cada microarray incluye la información de todos los genes que se encuentran en la microarray. Esta información incluye la posición del gen dentro de la microarray, por lo cual la tabla es vital para el funcionamiento de las aplicaciones. Actualmente, el número de microarrays en la base de

datos es muy inferior al que habrá al realizar las actualizaciones con la descarga de datos del NCBI, lo que puede producir que al descargar una gran cantidad de microarrays la tabla de genes acabe llegando al límite de registros posibles. Una posible solución de este problema es la siguiente:
Para los genes que sean comunes en varias microarrays, guardar un solo registro en la tabla de genes. Entonces, en el campo de posición, incluir una cadena donde tengamos por cada microarray en que esté el gen, el identificador de la microarray y la posición.

Por otra parte, como se ha comentado previamente, las microarrays se obtendrán de la base de datos GEO Datasets del NCBI. Estas microarrays pueden encontrarse en varios formatos. Tras analizar las opciones, se descargaran las microarrays en el formato llamado GDS Full, la cual incluye los valores de expresión normalizados, así como información extra de los genes como pueden ser identificadores únicos que nos servirá para llevar un control del nombre de los genes, ya que estos pueden cambiar desde el momento de su descubrimiento. Otro tema a tratar es como se realizará la descarga de las microarrays que necesitamos, ya que el NCBI proporciona diversas opciones. Tras analizarlo, la opción viable es realizar una consulta mediante la herramienta del NCBI eutils, la cual dada una consulta permite obtener identificadores que cumplen los términos introducidos en dicha consulta. Con esto, obtendremos los identificadores de las microarrays que deberemos descargar. Con estos identificadores y realizando una conexión FTP al NCBI, podremos descargar los archivos de las microarrays que necesitamos.

4. Planificación temporal del trabajo

El proyecto se ha planificado en las siguientes fases:

Fase 1: Conocimientos previos en el ámbito de la bioinformática y del proyecto

- Adquirir conocimientos sobre la bioinformática en general i sobre la tecnología de microarrays.
- Comprender las aplicaciones actuales en las que ha de integrarse el

proyecto. Comprender como está estructurado el servidor y entender las bases de datos con las que se integrará el proyecto.

- Entender el entorno del NCBI, sus bases de datos, su herramienta E-Utils, las opciones de descarga de microarrays y la estructura del FTP.

Fase 2: Implementación del Robot de Actualización de microarrays.

- Consulta a realizar para obtener los identificadores de las microarrays a descargar.
- Descarga de las microarrays por FTP.
- Parsear los ficheros de las microarrays descargadas.
- Subir a la base de datos las microarrays parseadas correctamente.
- Programar al robot para que se ejecute periódicamente.
- Pruebas de funcionamiento correcto del robot de actualización.

Fase 3: Tratamiento de las microarrays.

- Utilizar las herramientas disponibles para el análisis de microarrays para generar los diversos datos que son utilizados por la interfaz web. Este preproceso se realiza una vez por microarray.
- Pruebas de funcionamiento correcto del preproceso de microarrays.

Fase 4: Gestión web de la base de datos de microarrays.

- Incluir en la interfaz web una opción de búsqueda de microarrays. Esta búsqueda se podrá realizar mediante campos como la descripción de la microarray o la especie de la microarray.
- Incluir opción para que los usuarios puedan agregar microarrays a su lista de microarrays.
- Pruebas de la interfaz web de gestión de microarrays.

Fase 5: Pruebas, corrección de errores y optimización.

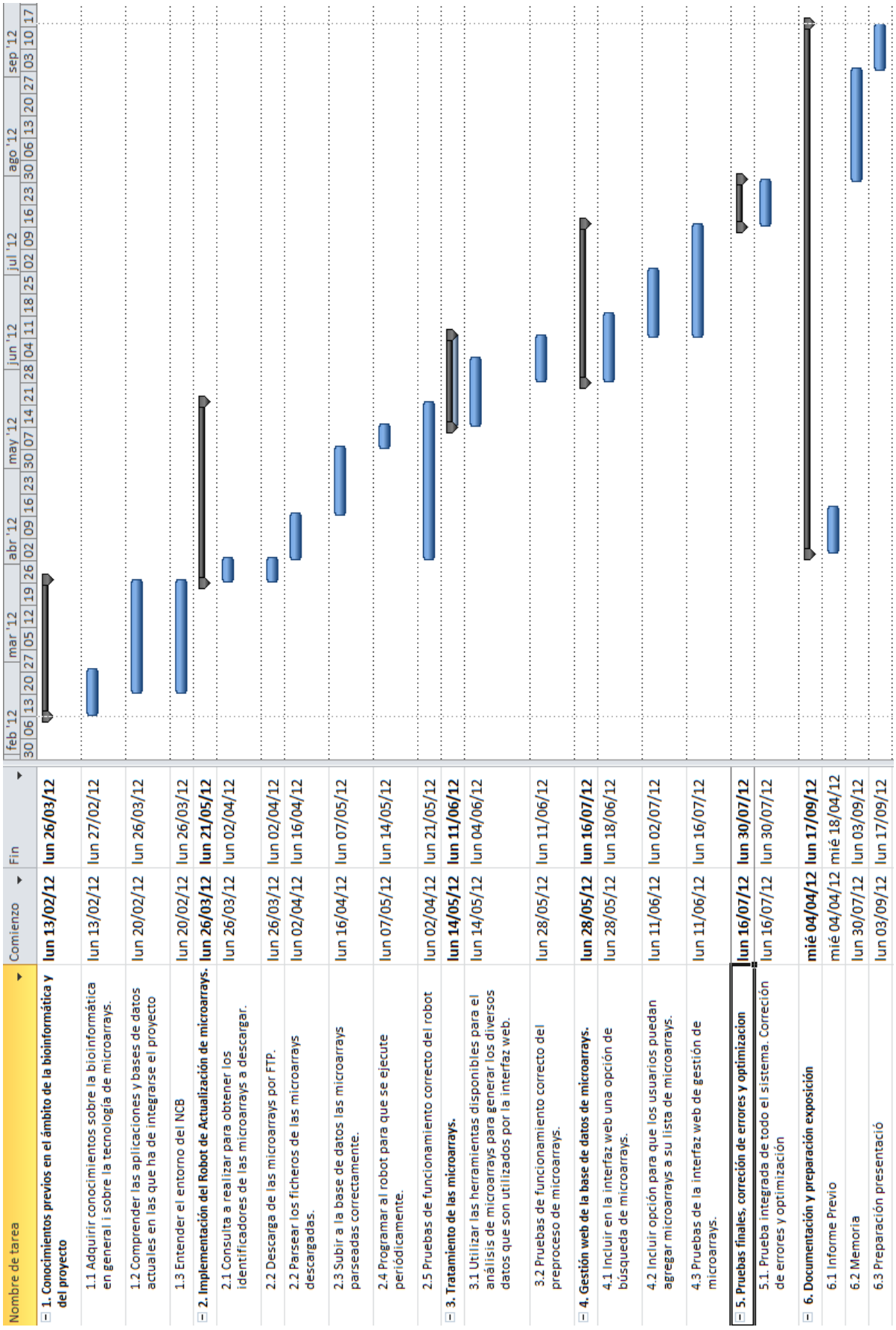
- Prueba integrada de todo el sistema. Incluye la corrección de errores y optimización del sistema.

Fase 6: Documentación y preparación exposición

- Informe previo

- Memoria
- Preparación presentación

En la figura 1 se puede observar la planificación de estas tareas según los periodos en los que están previsto realizarse. Se puede apreciar su Diagrama de Gant correspondiente.



Nombre de tarea	Comienzo	Fin
1. Conocimientos previos en el ámbito de la bioinformática y del proyecto	**lun 13/02/12**	**lun 26/03/12**
1.1 Adquirir conocimientos sobre la bioinformática en general i sobre la tecnología de microarrays.	lun 13/02/12	lun 27/02/12
1.2 Comprender las aplicaciones y bases de datos actuales en las que ha de integrarse el proyecto	lun 20/02/12	lun 26/03/12
1.3 Entender el entorno del NCB	lun 20/02/12	lun 26/03/12
2. Implementación del Robot de Actualización de microarrays.	**lun 26/03/12**	**lun 21/05/12**
2.1 Consulta a realizar para obtener los identificadores de las microarrays a descargar.	lun 26/03/12	lun 02/04/12
2.2 Descarga de las microarrays por FTP.	lun 26/03/12	lun 02/04/12
2.2 Parsear los ficheros de las microarrays descargadas.	lun 02/04/12	lun 16/04/12
2.3 Subir a la base de datos las microarrays parseadas correctamente.	lun 16/04/12	lun 07/05/12
2.4 Programar al robot para que se ejecute periódicamente.	lun 07/05/12	lun 14/05/12
2.5 Pruebas de funcionamiento correcto del robot	lun 02/04/12	lun 21/05/12
3. Tratamiento de las microarrays.	**lun 14/05/12**	**lun 11/06/12**
3.1 Utilizar las herramientas disponibles para el análisis de microarrays para generar los diversos datos que son utilizados por la interfaz web.	lun 14/05/12	lun 04/06/12
3.2 Pruebas de funcionamiento correcto del preproceso de microarrays.	lun 28/05/12	lun 11/06/12
4. Gestión web de la base de datos de microarrays.	**lun 28/05/12**	**lun 16/07/12**
4.1 Incluir en la interfaz web una opción de búsqueda de microarrays.	lun 28/05/12	lun 18/06/12
4.2 Incluir opción para que los usuarios puedan agregar microarrays a su lista de microarrays.	lun 11/06/12	lun 02/07/12
4.3 Pruebas de la interfaz web de gestión de microarrays.	lun 11/06/12	lun 16/07/12
5. Pruebas finales, corrección de errores y optimizacion	**lun 16/07/12**	**lun 30/07/12**
5.1. Prueba integrada de todo el sistema. Corrección de errores y optimización	lun 16/07/12	lun 30/07/12
6. Documentación y preparación exposición	**mié 04/04/12**	**lun 17/09/12**
6.1 Informe Previo	mié 04/04/12	mié 18/04/12
6.2 Memoria	lun 30/07/12	lun 03/09/12
6.3 Preparación presentació	lun 03/09/12	lun 17/09/12

Figura 1: Diagrama de Gant con la planificación de las tareas

5. Referencias

[1] Instituto de Biotecnología y de Biomedicina (IBB) de la Universidad Autónoma de Barcelona. http://ibb.uab.es/ibb/

[2] 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).

[3] Página web oficial de la base de datos Geo Datasets del NCBI. http://www.ncbi.nlm.nih.gov/gds

[4] D**elicado, P.(2001)** *Another look at principal curves and surfaces.* Journal of Multivariate Analysis, 77, 84-116 .

[5] **Delicado, P. and Huerta, M. (2003**): *'Principal Curves of Oriented Points: Theoretical and computational improvements'.* Computational Statistics 18, 293-315.

[6] **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.

[7] **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.

[8] **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

[9] **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.