2661 Bioinformática:
Comparación de datos de expresión génica del servidor local con datos de una Base de Datos Remota

Memoria del Proyecto
de Ingeniería Informática
realizado por
Marc Muñoz Escudero
y dirigido por
Jordi Gonzàlez i Sabaté
y Mario Huerta
Bellaterra, 22 de Junio de 2011

El sotasignat, Jordi Gonzàlez i Sabaté

Professor/a de l'Escola Tècnica Superior d'Enginyeria de la UAB,

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat sota la seva direcció per en Marc Muñoz Escudero

I per tal que consti firma la present.

Signat:

Bellaterra, 22 de Junio de 2011

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria

El sotasignat, Mario Huerta

de l'empresa, Institut de Biotecnologia i de Biomedicina de la UAB,

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat l'empresa sota la seva supervisió mitjançant conveni amb la Universitat Autònoma de Barcelona.

Així mateix, l'empresa en té coneixement i dóna el vist-i-plau al contingut que es detalla en aquesta memòria.

Signat:

Bellaterra, 22 de Junio de 2011

Índice de contenido

1. Introducción 1
1.1 Motivación personal 1
1.2 Estado del arte 2
1.3 Objetivos 4
1.4 Organización de la memoria 6

2. Fundamentos teóricos 8
2.1 Bioinformática 8
2.2 Microarrays 9
2.2.1 Agrupamiento o clustering 11
2.2.2 Genes marcadores 11

3. Fases 12
3.1 Conocimientos previos en el ámbito de la bioinformática y del proyecto 13
3.1.1 Adquirir conocimientos sobre la bioninformática 13
3.1.2 Familiarización con el entorno del NCBI 13
3.1.2.1 Página web (GEO Profiles) 14
3.1.2.2 Método de descarga de información de la base de datos de GEO 16
3.1.3 Construcción de consultas E-Utils necesarias 17
3.1.4 Compresión del aplicativo PCOPGene, concretamente NCR-PCOPGene 20
3.2 Construcción de base de datos de genes marcadores de microarrays 22
3.2.1 Descargar la información del NCBI mediante la herramienta E-Utils 22
3.2.2 Tratar la información descargada de los genes marcadores 23
3.3 Cálculos online 26
3.4 Aplicación web 28
3.4.1 Generación de imágenes a partir de los perfiles de expresión de los genes 28
3.4.2 Diseño e implementación de la interfaz web 30
3.4.2.1 Conexión entre aplicaciones web 31
3.4.2.2 Vista general de la interfaz web 33
3.4.2.3 Vista detalle de la interfaz web 39
3.4.2.4 Borrado de archivos temporales 43
3.5 Automatización mensual de la base de datos de genes marcadores de microarrays 44

4. Informe técnico … 46
4.1 Estructura de directorios 46
4.2 Construcción de la base de datos de genes marcadores de microarrays 46
4.2.1 Descarga de genes marcadores a través de E-Utils 47

4.2.2 Agrupación de los genes marcadores por microarrays 47
4.3 Cruce online 49
4.4 Generación de imágenes a partir de los perfiles de expresión de los genes 50
4.5 Interfície web 51
4.6 Actualización 52
4.7 Limpieza del servidor 53

5. Conclusiones 55
5.1 Trabajos futuros 56
5.2 Presupuesto 57

6. Bibliografía 58

1. Introducción

1.1 Motivación personal

La elección de este proyecto ha supuesto un reto personal y una gran fuente de motivación. Como origen de tal motivación está la oportunidad de realizar un trabajo útil que ayudará en investigaciones científicas.

El desarrollo del trabajo me ha aportado una gran cantidad de conocimientos tanto en el campo de la informática como de la biotecnología. Algo que agradezco.

En el campo de la informática, porque he obtenido los datos de bases de datos científicas internacionales que se actualizan constantemente con los últimos avances científicos, he diseñado las estructuras de datos del servidor local que se van actualizando con dicha información, he desarrollado el cruce de datos que se ejecuta en tiempo real a requerimiento del usuario y he diseñado e implementado el aplicativo web con el que se muestran los resultados y permiten al científico interactuar con ellos. El presente proyecto me ha dado la oportunidad de operar con los servidores del NCBI, fuente de datos en el campo de la biotecnología que es un referente a nivel mundial.

Agradezco también la formación biotecnológica recibida, porque me ha permitido descubrir como la rama de la genética se aplica dentro de la medicina. En mi caso concreto trabajando con datos de microarrays, que permiten analizar la expresión de miles de genes bajo cientos de condiciones experimentales diferentes (falta de agua, de oxígeno, introducción de agentes tóxicos, aplicación de medicamentos...).

Una vez realizado el proyecto y alcanzados los objetivos marcados, los investigadores tendrán a su disposición una aplicación web que facilitará la investigación génica para el estudio de todo tipo de enfermedades y descubrir en que diferentes patologías y procesos biológicos están implicados los diferentes grupos de genes.

La aplicación web realizada en el proyecto forma parte de una línea de investigación para el análisis de datos de microarrays llevada a cabo en el el Institut de Biotecnología y de Biomedicina (IBB) de la UAB.

Me enorgullece ver como he conseguido satisfacer las necesidades presentadas, un caso real con una aplicación real y con gran utilidad médica, un problema no solo mío, sino de toda la comunidad científica en el campo de la biotecnología. Y esta “necesidad científica” la he satisfecho aplicando los conocimientos que he adquirido en el transcurso de la carrera (más algunos nuevos adquiridos expresamente para el proyecto). De esta manera, puedo apreciar como lo aprendido en la universidad es útil una vez finalizada esta etapa de mi formación.

1.2 Estado del arte

La aplicación desarrollada pertenece al campo de la bioinformática. La bioinformática está formada por la unión de la biología molecular y la teoría de la información (una rama de las matemáticas) aplicando la potencia computacional de los ordenadores de hoy en día.

La bioinformática engloba todos los aspectos de la adquisición, procesamiento, distribución, análisis, interpretación e integración de la información biológica. Es decir, la bioinformática aplica las tecnologías de la información a la biología molecular así como a otras ciencias biomédicas.

Algunos experimentos que se realizan en biología molecular producen una gran cantidad de información. La tecnología de microarrays en concreto produce una gran cantidad de información sobre expresión génica. Estos datos sobre la expresión génica es necesario analizarlos posteriormente para extraer conocimiento útil. La tecnología de microarrays se encarga de obtener el nivel de expresión de un número elevado de genes (varias decenas de miles) para un gran número de condiciones experimentales. Como resultado se generan matrices de datos donde las filas representan genes, las columnas las condiciones experimentales y cada una de las celdas de la matriz, el nivel de expresión de cada gen para cada experimento.

La información generada por las microarrays dependerá de las condiciones experimentales aplicadas. Es decir, dependiendo de las condiciones experimentales, la microarray nos proporciona información de diferente tipo. Por ejemplo si las condiciones experimentales de la microarray son sobre el cáncer de colon, la microarray nos mostrará los niveles de expresión de los genes para el cáncer de colon, si en la microarray se estudia el efecto de fumar tabaco sobre los pulmones, la microarray nos proporcionará el nivel de expresión de los genes del tejido pulmonar bajo estas condiciones. De esta forma las condiciones experimentales pueden proporcionar la respuesta génica a distintos fármacos, a variaciones en las dosis, a diferentes fases en el progreso de una enfermedad, a diferentes tipos de células, a diferentes tipos de tejidos, al género u otras características de los pacientes, etc.

Un procedimiento habitual en el análisis de microarrays es agrupar las condiciones muestrales en clusters. Esto significa agrupar una serie de condiciones muestrales porque presentan un comportamiento similar, es decir, porque estadísticamente tienen el mismo efecto sobre la expresión de los genes (clustering estadístico). Los clusters pueden también tener un origen no estadístico (clustering por criterios biomédicos), por ejemplo agrupando los experimentos “manualmente” porque tengan algún atributo biológico o experimental común como los citados anteriormente[3].

Otro tipo de análisis para los datos generados por la tecnología de microarrays es la búsqueda de genes marcadores. Los genes marcadores son los genes de una microarray que se sobreexpresan en unas condiciones experimentales pero no en otras y por extensión, los genes que se sobreexpresan en unos clústers de condiciones experimentales pero no en otros.

El Instituto de Biotecnología y de Biomedicina (IBB)[1] 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[4][5][6][7][8][9]. Dentro de esta linea de investigación se ha desarrollado el servidor de aplicaciones: http://revolutionresearch.uab.es/ [2] para el análisis de datos de microarray. En dicho servidor se alberga la aplicación web PCOPGene[4], que permite estudiar y analizar datos de microarrays con diversos métodos desarrollados en el IBB. Una de las herramientas de la aplicación PCOPGene permite buscar los genes de la microarray con uno u otro cluster sobreexpresado o infoexpresado respecto al resto, es decir obtener los genes marcadores para los clusters que el usuario ha requerido [5]. De esta forma el usuario obtiene los genes marcadores para pasar a analizarlos en detalle.

El "National Center for Biotechnology Information" (NCBI)[3] es parte de la Biblioteca Nacional de Medicina de Estados Unidos. El NCBI es una importante fuente de información en biología molecular. El NCBI contiene una enorme base de datos de microarrays llamada GEO Profiles[10][11][12][13][14] que almacena los perfiles de expresión de los genes de 2,720 microarrays y 580,049 condiciones muestrales.

GEO Profiles proporciona los perfiles de expresión de cada gen para cada microarray mediante un gráfico que muestra el nivel de expresión de dicho gen para todas las condiciones muestrales de la microarray.

Las microarrays contenidas en la base de datos de GEO Profiles son subidas por los investigadores que crearon la microarray. Los mismos investigadores que suben la microarray son quienes definen manualmente los clusters con base biomédica para las condiciones muestrales de la microarray. Usando estos clusters, GEO Profiles permite realizar consultas para obtener los genes marcadores de cada microarray. Estos genes marcadores serán los genes que se sobrexpresen o infoexpresen en los clusters de origen biomédico definidos por los investigadores.

Enriquecer los clusters de origen estadístico (o no), proporcionándoles nuevos atributos y significado sería de una utilidad médica y biológica enorme . Este enriquecimiento es posible si conseguimos extrapolar los atributos de los clusters biomédicos de una microarray, a los clusters de otra microarray distinta. ¿ Y Porque sería de una gran utilidad médica y biológica? Si los clusters de la microarray del usuario son de origen estadístico, al proporcionarles atributos biomédicos, los usuarios podrán conocer el motivo biológico por el que las condiciones muestrales del cluster provocan la misma reacción en los genes de la microarray. Si los clusters son de origen biomédico (no estadístico) se le añadirán nuevos atributos biomédicos de un origen totalmente distinto al de las condiciones experimentales de la microarray. Esto es posible al extrapolarse información de otras microarrays de origen muy diferente. El enriquecimiento de los clusters para la microarray del usuario en el servidor local, puede obtenerse a partir de los clusters de origen biomédico de las microarrays del NCBI.

Por ejemplo, en el caso de que los clusters de la microarray del usuario sean de origen estadístico y esta sea una microarray para el estudio del cáncer rectal. Al comparar con otra microarray de cáncer de cuello de útero y encontrar equivalencias entre sus clusters, puede establecerse una equivalencia entre las diferentes fases de estos tipos de cáncer, lo que puede ser útil para investigar si tratamientos para el cáncer

de cuello de útero podrían aplicarse al cáncer rectal.

La presente es una posibilidad que no ha sido abordada hasta el momento, con lo que si se satisface la citada necesidad en este proyecto dará lugar a una herramienta nueva y realmente útil para los investigadores.

1.3 Objetivos

El objetivo principal del trabajo es enriquecer con información biomédica los clusters de condiciones muestrales de la microarray que el usuario este analizando en el servidor del IBB[1]. 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. Esta base de datos de genes marcadores se construirá a partir de la base de datos GEO[14] del NCBI[3]. La base de datos GEO permite consultas remotas para obtener los genes marcadores de cada una de sus microarrays, para los clusters de condiciones muestrales de cada microarray. Como se explicó en la sección anterior estos son clusters basados en criterios biomédicos o experimentales. Si los genes marcadores de un cluster de una microarray de GEO, es decir los genes que se sobrexpresan en dicho cluster pero que se infoexpresan en el resto de clusters, son también genes marcadores de un cluster que está analizando el usuario para su microarray en el servidor local, se puede establecer una equivalencia entre el cluster con base biomédica de la microarray del NCBI y el cluster de origen estadístico de la microarray del usuario. De esta forma los atributos biomédicos que dan origen al cluster de la microarray del NCBI pueden extrapolarse al cluster de la microarray del usuario, con lo que enriquecemos dichos clusters del usuario atribuyéndoles una información biológica extra. Si esto se realiza de forma masiva para 2,720 microarrays (el total de microarrays de GEO), el usuario puede obtener muchísima información relevante para sus clusters de muestras, con información que va mucho más allá de la información contenida en la microarray original que está analizando.

Para poder alcanzar el objetivo principal del proyecto será necesario cumplir y desarrollar los objetivos descritos a continuación:

- Actualización periódica y automática de la base de datos local de genes marcadores:
  - Extracción y construcción de la base de datos local a partir de la información del NCBI.
  - Actualizar de manera automática cada cierto tiempo la información almacenada en la base de datos de forma que se vayan incorporando los genes marcadores de las nuevas microarrays subidas al servidor del NCBI.
  - La actualización será robusta a posibles errores o caída del servidor durante el proceso de actualización.
- Cálculo en tiempo real de los genes marcadores comunes entre la microarray del usuario y la base

de datos de genes marcadores de microarrays:

- Búsqueda online de los genes marcadores comunes entre la base de datos de genes marcadores de las microarrays del NCBI y los genes marcadores de la microarray del usuario. Los genes marcadores de la microarray del usuario habrán sido calculados previamente (mediante la herramienta NCR-PCOPGene[5]) y también de forma online, para los clusters que el usuario haya especificado en su consulta.
- Permitir realizar el cruce restringiéndolo a microarrays con un atributo concreto (por ejemplo el cáncer de colon). Permitir realizar la búsqueda de genes comunes no sólo por el nombre oficial del gen sinó por todos los nombres posibles del gen (alias del gen), así como todos los nombres de la proteína sintetizada por dicho gen.
- Desarrollar un algoritmo de cruce óptimo que permita realizar la consulta online en el menor tiempo posible.

- Interfaz web para mostrar los resultados del cruce de los genes marcadores:
  - Vista general con el listado de microarrays que comparten genes marcadores con la búsqueda del usuario para su microarray:
    - Proveerá una lista de microarrays que comparten genes marcadores con la microarray de interés del usuario y se mostrará información relevante para cada microarray listada.
    - La información relevante mostrada será: el nombre de la microarray, el porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario (la lista de microarrays se ordenará inicialmente por este campo), el porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray listada, el título y descripción de la microarray listada. Una imagen con la distribución de los clusters de condiciones muestrales de la microarray listada.
  - Vista detalle con el listado de genes marcadores comunes y no comunes de cada microarray listada:
    - Mostrará información relevante sobre cada uno de los genes marcadores comunes para las dos microarrays, como el nombre del gen y su descripción, la separación en términos de expresión de los clusters para el gen marcador y la microarray del usuario (a mayor diferencia a la hora de expresa uno u otro clúster más marcador será el gen) y links referentes al gen (el grafo de correlación del gen marcador y el resto de genes de la microarray del usuario e información del gen de la base de datos Gene).
    - Se mostrarán también imágenes previamente generadas para el gen marcador donde se mostrarán los niveles de expresión del gen para cada cluster de la microarray. Por cada gen se mostrará una imagen para la microarray del NCBI y sus clusters, y otra imagen para la microarray del usuario y sus clusters. Las imágenes de los genes marcadores para las microarrays de GEO Profiles se habrán generado durante el proceso de actualización de la base de datos de genes marcadores y las imágenes para los genes marcadores de la microarray del usuario se generarán en el momento de realizar la consulta, puesto que

dependerá de la distribución de clusters que el usuario haya requerido [5].

1.4 Organización de la memoria

Para llevar a cabo los objetivos propuestos anteriormente he seguido la planificación que se expone a continuación.

En una primera fase me dedique a adquirir los conocimientos necesarios sobre la bioinformática y familiarizarme con el entorno del NCBI. Tanto su interfície web como los diversos métodos para acceder a la información almacenada.

Una vez comprendida la mejor manera de obtener los genes marcadores para poder construir la base de datos de genes marcadores de microarrays. Descargué y traté la información. Dí forma a la base de datos local que más tarde se actualizará de forma automática.

La tercera fase consiste en implementar el cruce entre genes marcadores de la microarray del usuario y la base de datos de genes marcadores de microarrays. Los genes marcadores comunes y las microarrays a las que pertenecen se mostrarán al investigador mediante la interfície web que se diseña en la siguiente fase.

En la cuarta fase del trabajo desarrollé la aplicación web e incorporé en el aplicativo la generación de imágenes que muestran la distribución de clusters de condiciones muestrales de la microarray a lo largo de la expresión de cada gen marcador.

Para poder generar las imágenes de la distribución de clusters, previamente se descargaron las microarrays del FTP que contienen los valores de expresión de los genes para cada condición muestral.

Diseñe e implemente la interfície web de manera que fuera clara e intuitiva para el usuario, mostrando información extra sobre microarrays y genes marcadores para así poder ayudar al investigador en su estudio.

La fase final del trabajo ha sido dedicada ha realizar test de pruebas a la aplicación web y a la automatización mensual de la base de datos de genes marcadores de microarrays.

Durante cada fase del trabajo se han realizado diferentes modificaciones para mejorar la funcionalidad, operatividad y usuabilidad del conjunto de la aplicación con la supervisión y asesoramiento del codirector Mario Huerta.

En los capítulos venideros se describe con minuciosidad el trabajo realizado. La estructura seguida es:

- Fundamentos teóricos: Expongo los conocimientos biológicos y técnicos necesarios para entender el proyecto.
- Fases: Describo el trabajo desarrollado para resolver el proyecto, los problemas encontrados y la solución adoptada.
- Informe técnico: Describo los programas implementados y la estructura de directorios detalladamente para facilitar su reusabilidad y adaptabilidad.
- Conclusiones
- Bibliografía
- Resumen

2. Fundamentos teóricos

2.1 Bioinformática

Bioinformática es una disciplina científica emergente que utiliza tecnología de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biología. Bioinformática es un área de investigación multidisciplinaria, la cual puede ser ampliamente definida como la interfase entre dos ciencias: Biología y Computación y esta impulsada por la incógnita del genoma humano y la promesa de una nueva era en la cual la investigación genómica puede ayudar dramáticamente a mejorar la condición y calidad de vida humana.

Avances en la detección y tratamiento de enfermedades y la producción de alimentos genéticamente modificados son entre otros ejemplos de los beneficios mencionados más frecuentemente. Involucra la solución de problemas complejos usando herramientas de sistemas y computación. También incluye la colección, organización, almacenamiento y recuperación de la información biológica que se encuentra en base de datos.

Según la definición del Centro Nacional para la Información Biotecnológica "National Center for Biotechnology Information" (NCBI por sus siglas en Inglés, 2001): "Bioinformática es un campo de la ciencia en el cual confluyen varias disciplinas tales como: biología, computación y tecnología de la información. El fin último de este campo es facilitar el descubrimiento de nuevas ideas biológicas así como crear perspectivas globales a partir de las cuales se puedan discernir principios unificadores en biología. Al comienzo de la "revolución genómica", el concepto de bioinformática se refería sólo a la creación y mantenimiento de base de datos donde se almacena información biológica, tales como secuencias de nucleótidos y aminoácidos. El desarrollo de este tipo de base de datos no solamente significaba el diseño de la misma sino también el desarrollo de interfaces complejas donde los investigadores pudieran acceder los datos existentes y suministrar o revisar datos.

Luego toda esa información debía ser combinada para formar una idea lógica de las actividades celulares normales, de tal manera que los investigadores pudieran estudiar cómo estas actividades se veían alteradas en estados de una enfermedad. De allí viene el surgimiento del campo de la bioinformática y ahora el campo más popular es el análisis e interpretación de varios tipos de datos, incluyendo secuencias de nucleótidos y aminoácidos, dominios de proteínas y estructura de proteínas.

El proceso de analizar e interpretar los datos es conocido como biocomputación. Dentro de la bioinformática y la biocomputación existen otras sub-disciplinas importantes:

El desarrollo e implementación de herramientas que permitan el acceso, uso y manejo de varios

tipos de información.

El desarrollo de nuevos algoritmos (fórmulas matemáticas) y estadísticos con los cuales se pueda relacionar partes de un conjunto enorme de datos, como por ejemplo métodos para localizar un gen dentro de una secuencia, predecir estructura o función de proteínas y poder agrupar secuencias de proteínas en familias relacionadas.

La Medicina Molecular y la Biotecnología constituyen dos áreas prioritarias científico tecnológicas como desarrollo e Innovación Tecnológica. El desarrollo en ambas áreas están estrechamente relacionadas. En ambas áreas se pretende potenciar la investigación genómica y postgenómica así como de la bioinformática, herramienta imprescindible para el desarrollo de estas. Debido al extraordinario avance de la genética molecular y la genómica, la Medicina Molecular se constituye como arma estratégica del bienestar social del futuro inmediato. Se pretende potenciar la aplicación de las nuevas tecnologías y de los avances genéticos para el beneficio de la salud. Dentro de las actividades financiables, existen acciones estratégicas, de infraestructura, centros de competencia y grandes instalaciones científicas. En esta área, la dotación de infraestructura se plasmará en la creación y dotación de unidades de referencia tecnológica y centros de suministro común, como Centros de Bioinformática, que cubran las necesidades de la investigación en Medicina Molecular. En cuanto a centros de competencia, se crearán centros de investigación de excelencia en hospitales en los que se acercará la investigación básica a la clínica, así como centros distribuidos en red para el apoyo a la secuenciación, DNA microarrays y DNA chips, bioinformática, en coordinación con la red de centros de investigación genómica y proteómica que se proponen en el área de Biotecnología. En esta área la genómica y proteómica se fundamenta como acción estratégica o instrumento básico de focalización de las actuaciones futuras.

Las tecnologías de la información jugarán un papel fundamental en la aplicación de los desarrollos tecnológicos en el campo de la genética a la práctica médica como refleja la presencia de la Bioinformática médica y la Telemedicina dentro de las principales líneas en patología molecular. La aplicación de los conocimientos en genética molecular y las nuevas tecnologías son necesarios para el mantenimiento de la competitividad del sistema sanitario no sólo paliativo sino preventivo. La identificación de las causas moleculares de las enfermedades junto con el desarrollo de la industria biotecnológica en general y de la farmacéutica en particular permitirán el desarrollo de mejores métodos de diagnóstico, la identificación de dianas terapéuticas y desarrollo de fármacos personalizados y una mejor medicina preventiva.

2.2 Microarrays

Un chip de ADN (del inglés DNA microarrays)[1] es una superficie sólida a la cual se unen una serie de fragmentos de ADN. Las superficies empleadas para fijar el ADN son muy variables y pueden ser vidrio, plástico e incluso chips de silicio. Los arreglos de ADN son utilizadas para averiguar la expresión de genes,

1 http://es.wikipedia.org/wiki/Chip_de_ADN

monitorizándose los niveles de miles de ellos de forma simultanea.

La técnica consiste en extraer el RNAm de una célula buena y de otra experimental mediante isolation RNA, una vez extraído los dos RNAm se marca cada uno de ellos con un color distinto y se combinan los dos. Acto seguido se vierte el combinado en la superficie del chip de tal modo que cada RNAm se unirá o no a los cDNA de cada gen del chip. Finalmente, aplicando técnicas de análisis de imágenes es posible generar una matriz de datos numéricos, a partir de los patrones de intensidades de cada celda y discriminando la señal informativa de ruido que pudiera haber en segundo plano. Estos datos numéricos corresponderán al nivel de expresión de cada gen expuesto a una serie de condiciones muestrales. Por ejemplo, si el ARN de la célula experimental se coloreó de color rojo, los genes con un color más cercano al rojo tienen un nivel de expresión más elevado en las células experimentales.

Actualmente existen diferentes bases de datos a nuestro alcance a través de internet que unifican y facilitan toda esta información genética además de ofrecer diversas herramientas para el análisis de esta gran cantidad de información. Algunas de estas bases de datos por ejemplo son las que hay en el EMBL (European Molecular Biology Laboratory), el SIB (Swiss Institute of Bioinformatics), el EBI (European Bioinformatics Institute) o el NCBI (National Center for Biotechnology Information). El EBI y el NCB son los que más información contienen y por lo tantos los más utilizados.

Por lo tanto las microarrays son una potente fuente de obtención de perfiles de expresión de genes sometidos a diferentes condiciones, identificar los patrones de los niveles de expresión será muy útil para compararlos y poder estudiar las respuestas de los genes.

Aplicando una serie de procesos experimentales y computacionales sobre la microarray se obtiene una matriz numérica bidimensional que consta de los genes de poblaciones distintas como individuos y de las condiciones muestrales a las que se expusieron las células como variables en el caso que se quiera estudiar a los genes, o a la inversa, si es que se quiere realizar un estudio comparativo de las condiciones a que se somete(este estudio es en el que se basa el trabajo realizado). Cada uno de los valores de la matriz representa el nivel de expresión de un determinado gen bajo una cierta condición muestral. Es posible que en algunos casos se produzcan errores en el proceso y se generen huecos, estos huecos tiene valor 0 dentro de la matriz.

Estas matrices son de grandes dimensiones puesto que existen una gran cantidad de genes y de condiciones muestrales.

Dado que realizar una análisis de esta matriz de gran dimensionalidad es una tarea prácticamente imposible se hacen necesarias técnicas computacionales que permitan agrupar todos estos datos y a partir de el agrupamiento realizar el análisis biológico. La técnica mas utilizada es el agrupamiento o agrupación.

2.2.1 Agrupamiento o clustering

El objetivo de los algoritmos de agrupamiento (clustering en inglés) es , dada una matriz de individuos y variables, encontrar un grupo (clúster) de un conjunto de individuos, de tal forma que los clústers resultantes sean homogéneos y/o estén bien separados. El punto clave es reducir la gran cantidad de datos caracterizándolos en grupos más pequeños de individuos similares. Esto implica que los individuos pertenecientes a un mismo clúster son lo más similares posibles entre ellos, mientra que los individuos de clústers distintos son lo más disimilares posibles[16]. La agrupación de los individuos se realizará de acuerdo a la separación entre ellos determinada por una medida de distancia dada, llamada medida de disimilaridad, este tipo de clustering es estadístico. En cambio, cuando la agrupación de los individuos se realiza de acuerdo a un criterio biomédico por parte del investigador, es un tipo de clusterings por criterios biomédicos.

2.2.2 Genes marcadores

Otra tipo de análisis de microarrays consiste en determinar aquellos genes que se sobrexpresan en unas condiciones experimentales pero no en otras y por extensión, los genes que se sobreexpresan en unos clústers de condiciones experimentales pero no en otros. Estos genes son los llamados genes marcadores. Los genes marcadores se caracterizan por destacar el comportamiento de la microarray, permitiendo decidir en la comparación con otros genes marcadores si estos pertenecen o no a la microarray.

3. Fases

A continuación se muestra una tabla comparativa entre la planificación inicial y la planificación final del trabajo desarrollado:

Planificación inicial	Planificación final
Conocimientos previos en el ámbito de la bioinformática y del proyecto	
Construcción de la base de datos local de genes marcadores de microarrays a partir de una consulta E-Utils a GEO Profiles.	
Diseño e implementación del programa de cruce en tiempo real de genes marcadores entre la microarray de interés y los genes marcadores de la base de datos de microarrays	Diseño e implementación del programa de cruce en tiempo real de genes marcadores entre la microarray de interés y los genes marcadores de la base de datos de microarrays. Optimizado para conseguir un tiempo de cómputo mínimo.
Descarga mediante FTP de los valores de expresión de las microarrays y posterior tratamiento de los genes marcadores.	
Generación de imágenes a partir de los valores de expresión de los genes marcadores tratados anteriormente	
Diseño e implementación de la interfaz web que muestra los resultados	
	Insertar nuevas funcionalidades a la interfaz web
	Borrado periódico de archivos temporales producidos por la interfaz web
Automatización mensual de la base de datos de genes marcadores de microarrays	

Como se aprecia en la tabla, la planificación inicial resulto ser bastante acertada. Sólo ha sido modificada por la búsqueda de un algoritmo realmente óptimo para el cruce de genes marcadores así como por la incorporación de nuevas prestaciones o funcionalidades a la aplicación web.

Los requisitos para realizar el proyecto son:

- Un servidor que debe tener:
  - Conexión óptima a Internet
  - Servidor web Apache
  - Entorno PHP
  - Base de datos MySQL
  - Interprete Perl
  - Compilador GCC
  - Espacio libre en disco suficiente
- Acceso a las base de datos remotas que contengan la información sobre las microarrays que necesito: genes marcadores y valores de expresión de los genes para las condiciones muestrales de la microarray.

3.1 Conocimientos previos en el ámbito de la bioinformática y del proyecto

Antes de empezar ha realizar el trabajo, tuve que adquirir ciertos conocimientos sobre la bioinformática. Anteriormente no había tenido la posibilidad de adentrarme en esta rama de la informática. También ha sido necesario familiarizarme y comprender el entorno del NCBI, lugar escogido como base de datos remota por ser un referente a nivel mundial.

3.1.1 Adquirir conocimientos sobre la bioinformática

Para poder realizar correctamente el trabajo es necesario aprender los conceptos fundamentales con la intención de comprender en su plenitud el proyecto escogido. Los conceptos adquiridos fueron en referencia al análisis de microarrays, donde se engloba la aplicación web implementada. Los conceptos propiamente dichos se muestran en la anterior sección de fundamentos teóricos. Abordando el concepto de una microarray, que es una tabla con gran cantidad de datos que representa las condiciones muestrales (columnas) de un experimento aplicados a genes (filas). También se explica que según el resultado obtenido en un experimento a un gen, este es un gen marcador o no. Quiere decir, que dicho gen resalta claramente el comportamiento o no del experimento aplicado, a partir de sobreexpresarse sobre un clúster respecto a los otros.

Por ello, en el trabajo realizado sólo nos interesan aquellos genes que sean genes marcadores dentro de la microarray. Porque al ser genes representativos de una microarray o experimento, nos sirven perfectamente en el momento de cruzar dos bases de datos de microarrays. Utilizando los genes marcadores para desmarcar mejor el comportamiento de la microarray y por consiguiente los clusters por los que esta formado.

Otro concepto necesario para el desarrollo del trabajo es el de clúster. Un clúster es la agrupación de las condiciones muestrales aplicadas a una microarray por tener un comportamiento parecido o similar, ya sea estadístico o por criterios biomédicos.

3.1.2 Familiarizarme con el entorno del NCBI

Estando ya situado en la bioinformática, el siguiente punto a tratar es el como y donde obtener la información necesaria para construir la base de datos local. La respuesta estaba clara, del National Center for Biotechnology Information (NCBI)[3]. El NCBI es el Centro Nacional para la Información Biotecnológica de los Estados Unidos donde se almacena y constantemente se actualiza la información referente a la

biología molecular. Siendo así una fuente importante de información para todo el mundo, ya que tienen disponibles todas sus bases de datos de manera gratuita y muy accesible.

Teniendo claro donde obtener la información para construir la base de datos local, tenía que descubrir como era la mejor manera de descargar dicha información. El NCBI contiene gran cantidad de bases de datos[2]. Pero la base de datos que más me interesaba para el trabajo ha realizar es la GEO Profiles [10][11][12][13][14]. GEO Profiles es una enorme base de datos formada por 2,710 microarrays donde se guardan los valores de expresión de los genes.

3.1.2.1 Página web (GEO Profiles)

Seleccionada la base de datos del NCBI, tenía que familiarizarme, comprender y hacerme con la sección de GEO Profiles[3] de su página web. Para comprender la estructura y a que corresponde cada parte página web, dan facilidades mediante su ayuda[4].

En la figura 3.1 se puede ver el aspecto y como esta marcada cada parte de la página web. Donde tiene una entrada de texto (A), para poder introducir los requerimientos de búsqueda que se quiera realizar dando la posibilidad de agregar diversas opciones para filtrar mejor la información que se necesita. Más adelante se hará hincapié sobre las opciones de filtrado. Permite configurar (B) la información ha mostrar según el formato, los elementos por cada página y el método de ordenación. Por cada perfil mostrado tiene un checkbox (C) de selección para profundizar más adelante con los elementos seleccionados. Cada perfil de expresión resultado de la búsqueda (D) se muestra la Annotation, que corresponde al símbolo del gen, el nombre completo y los alias, el Report, es la referencia de la secuencia original, y el Experiment, que es el título de la microarray. En la parte E, muestra el número de samples (condiciones muestrales del experimento) y diferentes links relacionados con los perfiles vecinos, cromosomas vecinos, secuencia vecinos, relación por homología del gen y un último link que relaciona recíprocamente con otras bases de datos. Por cada perfil se muestra una imagen (F) representativa de los niveles de expresión del gen para las condiciones muestrales de la microarray, así como la división de clústers de la microarray. El botón de descarga de perfiles (G) permite al usuario descargar los valores y anotaciones para cada perfil de la página o en su defecto sólo aquellos que se hayan seleccionado en el checkbox (C). En la parte H, se encuentran los datos relacionados y características parecidas a la parte E.

2 http://www.ncbi.nlm.nih.gov/guide/all/#databases_
3 http://www.ncbi.nlm.nih.gov/geoprofiles/
4 http://www.ncbi.nlm.nih.gov/geo/info/profiles.html

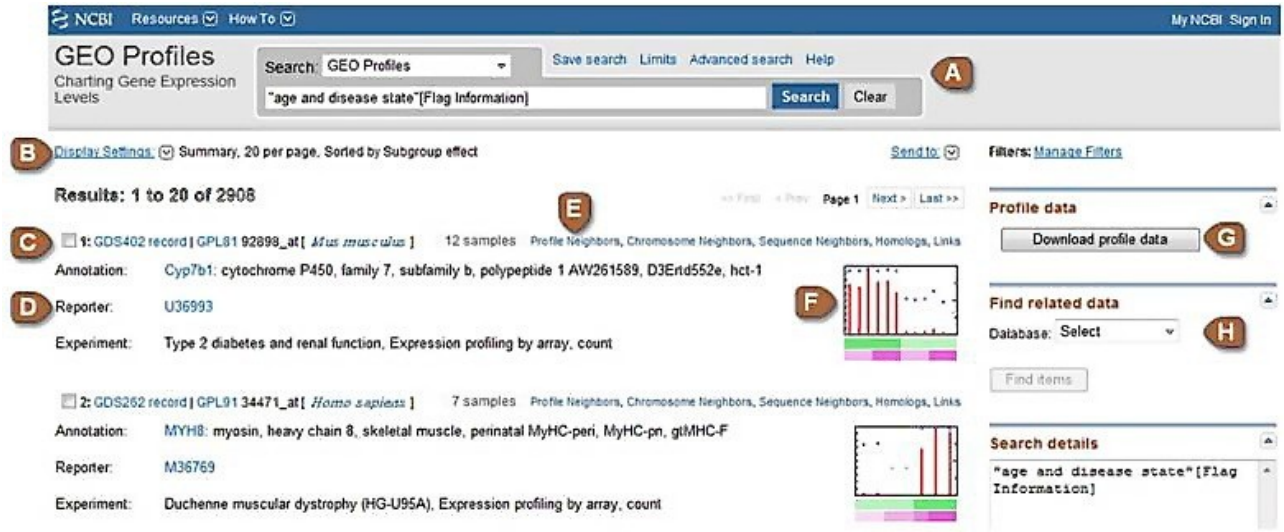


Figura 3.1: Vista general de los genes de diferentes microarrays proporcionados por una consulta a GEO Profiles. En este caso se trata de una consulta requiriendo microarrays con sus condiciones muestrales agrupadas en clusters basados en la edad (sin importar la especie del sujeto) y diferentes estados de una patología (sin importar de que tipo). (A) permite introducir los requerimientos de búsqueda. (B) configurar opciones del mostrado de resultados. (C) checkbox para seleccionar cada perfil. (D) perfil de expresión resultado de la búsqueda. (E) link relacional. (F) imagen representativa de los niveles de expresión del gen para las condiciones muestrales de la microarray. (G) descargar los datos de perfil. (H) búsqueda de datos relacionados.

Al realizar un click sobre la imagen que aparece en la figura 3.1 sección F, aparece la imagen de una manera más detallada, como se puede ver en la figura 3.2. Donde se puede apreciar más claramente, dando un un resumen de la microarray (I), la imagen ampliada (J) niveles de expresión del gen para las diferentes muestras de la microarray, los clusters de naturaleza biomédica (K), los valores de expresión (L) y cada condición muestral (M).

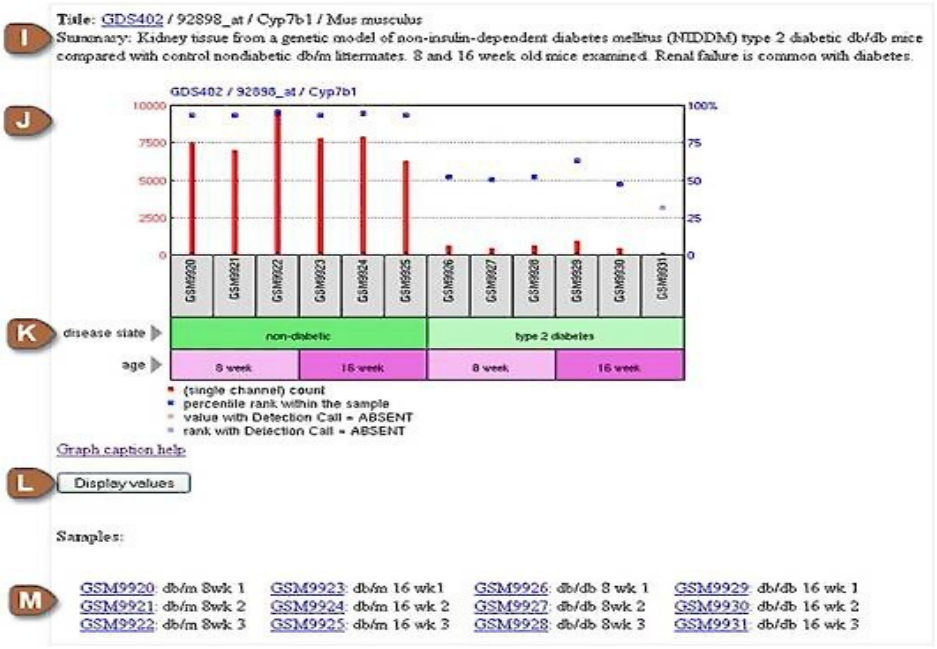


Figura 3.2: Vista detalle de la expresión del gen Cyp7b1 para las condiciones muestrales de la microarray GDS402 que estudia el riñón de ratas diabéticas. Las agrupan en muestras de ratón diabético y no diabético, así como en muestras de ratones de 8 semanas o 18 semanas de edad. Estos serían clusters de muestras de origen biomédico definidos por los creadores de la microarray. (I) resumen de la microarray. (J) niveles de expresión del gen para las diferentes muestras de la microarray. (K) clusters de origen biomédico definidos por los autores de la microarray. (L) muestra los valores de expresión.(M) condiciones muestrales.

La imagen mostrada en la figura 3.2 permite al investigador poder observar los valores de expresión por cada condición muestral de la microarray y las agrupaciones de condiciones muestrales en clusters de naturaleza biomédica (K). Por ello, considero necesario poder mostrar al usuario de mi aplicación web una imagen representativa de los clusters de la microarray. La imagen de mi aplicación web en lugar de mostrar los valores de expresión como sucede en la figura 3.2, muestro los niveles de expresión del gen para cada clúster. Más adelante retomaré el tema y lo explicaré con detenimiento.

3.1.2.2 Método de descarga de información de la base de datos de GEO

Al comprender la base de datos de GEO Profiles y que posibilidades daba su página web, me centré en la manera de obtener la información. La información tendría que ser de un gran número de genes (cuantos más mejor, ya que será más útil en el momento de realizar el cruce entre genes marcadores).

El NCBI al permitir acceder a su información, tiene diversos modos para poder obtener sus datos. Las opciones permitidas para el caso de la base de datos GEO Profiles son:

- A través de los links de descarga por cada microarray (DataSet).

- Descarga por FTP.
- Un acceso programado mediante consultas Entrez Programming Utilities (E-Utils).
- A través de los links de descarga de perfiles de expresión que aparecen en la página web, correspondiente a la figura 3.1, G.
- Realizando una consulta a Entrez GEO DataSets y Entrez GEO Profiles, y luego exportando el resultado a un documento o archivo.

La opción escogida tiene que cumplir dos requisitos adecuados al trabajo que se va ha realizar sobre estos datos, que son los siguientes:

- Descargar de manera automática, es decir, mediante un programa.
- Descargar sólo genes marcadores.

La opción que cumple los requisitos descritos es el acceso programado mediante consultas E-Utils. Porque la opción links de descarga por microarray, link de descarga de perfiles de expresión de la página web y mediante consulta Entrez son opciones manuales que no se pueden realizar mediante un programa. Entonces quedan dos opciones, FTP[5] y E-Utils[6], donde FTP realiza la descarga total de la microarray causa que no queremos para formar la base de datos y E-Utils tiene un comportamiento parecido a la página web de GEO Profiles mediante consultas online. Por ello, la opción más adecuada es E-Utils, aunque cabe decir que posteriormente será necesario utilizar la opción FTP para poder obtener los valores de expresión de cada gen marcador de la base de datos de genes marcadores de microarrays.

3.1.3 Construcción de consultas E-Utils necesarias

Decidida la forma de realizar la descarga de información, llega el momento de averiguar como se construye y la mejor manera de obtener los datos.

E-Utils[15] es una herramienta que facilita el acceso a los datos de Entrez fuera de la interfaz web mediante consultas y puede ser útil para obtener datos a partir de una búsqueda. La respuesta de la herramienta es un archivo con formato XML[7]. Hay ciertos requisitos por parte del NCBI para utilizar dicha herramienta:

- Ejecutar programas los fines de semana o entre las 21:00 y 5:00 del Este durante la semana para una serie de mas de 100 solicitudes.
- Enviar consultas a E-Utils, no a la página estándar del NCBI.

5 http://www.ncbi.nlm.nih.gov/Ftp/
6 http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
7 XML es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium(W3C). La tecnología XML busca dar solución al problema de expresar información estructurada de la manera más abstracta y reutilizable posible

- No hacer más de tres peticiones por cada segundo.
- Utilizar el parámetro email y tool para el software distribuido, de manera que puedan analizar el proyecto y ponerse en contacto si hubiese algún problema.

Cualquier incumplimiento de los requisitos del NCBI para utilizar la herramienta E-Utils provocaría un corte de la conexión con el servidor donde se ejecuten las consultas. Por esta razón hay que tener en cuenta las condiciones descritas en el momento de ejecutar el programa de descarga de información.

La consulta que se realiza esta formada por cuatro partes:

- Una parte base, común a todas las consultas que se realizan. Porque permite el acceso a E-Utils: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/ .
- La segunda parte determina la aplicación ha utilizar con E-Utils..
  Opciones:
  - eSearch: a partir de una consulta de texto responde con la lista de identificadores únicos (UID) dada una base de datos.
  - eSummary: a partir de una lista de UIDs responde con el resumen de ellos.
  - eFetch: a partir de una lista de UIDs responde con los registros.
  - ePost: acepta una lista de UIDs guardándolos como conjunto en el historial del servidor y responde con la correspondiente “query key” y “web environment”. Estos valores hacen referencia al conjunto o consulta asociada, así el servidor recuerda el conjunto o consulta anteriormente realizada. Las opciones también sirven para el resto de aplicaciones.
  - eLink: a partir de una lista de UIDs de una base de datos permite relacionar dicha lista con los IDs correspondientes a otra base de datos de Entrez.
  - eInfo: proporciona el número de registros indexados en un campo de una base de datos, la fecha de la última actualización y los enlaces que están disponibles con otras bases de datos de Entrez.
- La tercera parte consiste en determinar a que base o bases de datos se realiza la consulta. El NCBI consta de muchas bases de datos, pero en este trabajo sólo me centro en dos: la GEO Profiles y la Gene. Con la base de datos GEO Profiles se obtiene la lista de identificadores únicos de GEO Profiles(UIDs) asociados a la consulta y con la base de datos Gene se obtiene la lista de identificadores de los genes (IDs) a partir de los UIDs.
- La última parte es propiamente la consulta ha realizar. Consiste en el término de entrada a la base de datos, opciones opcionales al término, opciones opcionales a la consulta y otras opciones obligatorias. Más adelante se explicará en detalle las opciones para la consulta.

Sabiendo la estructura de la consulta ha realizar, tengo que decidir la forma de construir dicha consulta. Selecciono las aplicaciones necesarias para poder descargar la información.

En la segunda parte de la consulta me interesa primero la aplicación eSearch, porque mediante un término introducido responde con una lista de UIDs. Cada UID corresponde con un perfil de expresión mostrado en la página web. Enlazando la salida de la consulta eSearch con la entrada de eSummary. Además de saber el UID de cada perfil se obtiene información sobre el UID. También se necesita la aplicación eLink, para poder averiguar el ID de la base de datos Gene a partir del UID de la base de datos GEO Profiles. Más adelante explico porque es necesario utilizar esta última aplicación de E-Utils (eLink).

Para acabar de construir las consultas necesarias para obtener la información del NCBI, hay que determinar en la cuarta parte de la solicitud que término se introduce y las diversas opciones adicionales para la correcta utilización.

Para escoger el término adecuado, en la sección de información de la página web de GEO Profiles[8] aparece la explicación de los campos y ejemplos de búsqueda. Aparece una gran cantidad de campos disponibles ha utilizar, pero como interesa aquellos genes más representativos de una microarray. Selecciono el campo “Flag Type” y “Ranked Standard Deviation”. El primero hace referencia a los perfiles que exhiben específicamente el tipo de efectos del subconjunto. El segundo campo está relacionado con el percentil del rango de la desviación estándar del perfil en comparación con todos los otros perfiles en un conjunto de datos.

Antes de acabar de construir la consulta para ser lanzada a E-Utils, realizo diversas pruebas en la página estándar de GEO Profiles. Utilizando el “Ranked Standard Deviation” con valores de 50,75 y 100[9] y utilizando el “Flag Type”, probando el ejemplo de la ayuda, con rank subset effect[10].

Me doy cuenta que el resultado es muy similar, aparte del número de genes resultante cercano a 700 mil. Los primeros genes mostrados en la página web son semejantes. Al ver esta coincidencia contacté con el staff de ayuda referente a GEO Profiles del NCBI vía correo electrónico. Donde explicaba el motivo de mi e-mail, dado dos consultas diferentes que mostrarán un resultado parecido y como podría obtener una consulta similar a la provocada cuando un usuario introduce un texto en la página web de GEO Profiles.

Me resolvieron la duda, en la primera parte de la pregunta contestaron que aunque fueran consultas diferentes tenían un comportamiento similar y por ello el resultado. Referente a la segunda parte, conseguir una salida como la suya, con ese orden no es posible. Porque los cálculos referentes al efecto de un subconjunto se realizan con un método ad-hoc. Aunque me aconsejaron que utilizase el campo “Flag Type” con value subset effect[11], que obtendría una salida similar a la producida por la página web de GEO Profiles.

Al probar dicha consulta, el resultado era semejante a los dos anteriores casos. Haciendo caso a la respuesta del e-mail, acabe escogiendo esta última opción para formar mi consulta.

8 http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html
9 Ejemplo con valor 100: http://www.ncbi.nlm.nih.gov/geoprofiles?term=100[Ranked+Standard+Deviation]
10 http://www.ncbi.nlm.nih.gov/geoprofiles?term=rank+subset+effect[Flag+Type]
11 http://www.ncbi.nlm.nih.gov/geoprofiles?term=value+subset+effect[Flag+Type]

Referente a las opciones a insertar en la consulta para la primera parte de la descarga, simplemente inserto la opción “usehistory”. Insertando dicha opción permite al servidor almacenar la consulta realizada. En la respuesta de la consulta en formato XML muestra 20 UIDs del total de la lista de UIDs de la consulta y se añade un par de campos o "claves" mencionados anteriormente “query key” y “web environment”. Como campo obligatorio en cada consulta, el NCBI recomienda introducir el e-mail por posibles errores o uso indebido de la herramienta E-Utils.

3.1.4 Compresión del aplicativo PCOPGene, concretamente NCR-PCOPGene

Comprender el aplicativo PCOPGene[4], concretamente NCR-PCOPGene[5], es importante porque es el lugar donde se integra mi aplicación web dentro del servidor[2].

La herramienta PCOPGene es una aplicación web para el estudio de microarrays. Hay dos vertientes para su utilización. En la primera vía, a partir de un gráfico de genes marcadores permite seleccionarlos con la intención de poder estudiar dichos genes seleccionados. Dando la posibilidad de relacionar los genes con publicaciones donde aparece el gen, permite mostrar información relevante sobre el gen, etc. En la segunda vía, tiene una serie de herramientas que permite analizar la microarray que contiene los genes marcadores mencionados en la anterior vía y su comportamiento, la relación entre los diferentes clústers y como se comportan, etc.

En PCOPGene hay una herramienta web llamada NCR-PCOPGene, donde recae mi interés, ya que mi aplicación web se enlaza con ella. La aplicación tiene la funcionalidad de encontrar los genes marcadores de la microarray del usuario, a partir del criterio del usuario en sobreexpresar o infoexpresar unos clusters respecto a otros, como se puede observar en la figura 3.3.

En la figura 3.3 se determinan las condiciones de búsqueda de genes marcadores utilizando como base matemática los intervalos de confianza creados a partir de la distribución T de Student. El nivel de confianza define la cantidad de condiciones muestrales que deben estar dentro del intervalo, según el usuario determine el valor de alfa. El nivel de confianza será 99.7%, 99.1% y 95.1% dependiendo del valor de alfa que se selecciona en la parte inferior izquierda de la figura 3.3, con valores 0.003, 0.009 y 0.049, respectivamente.

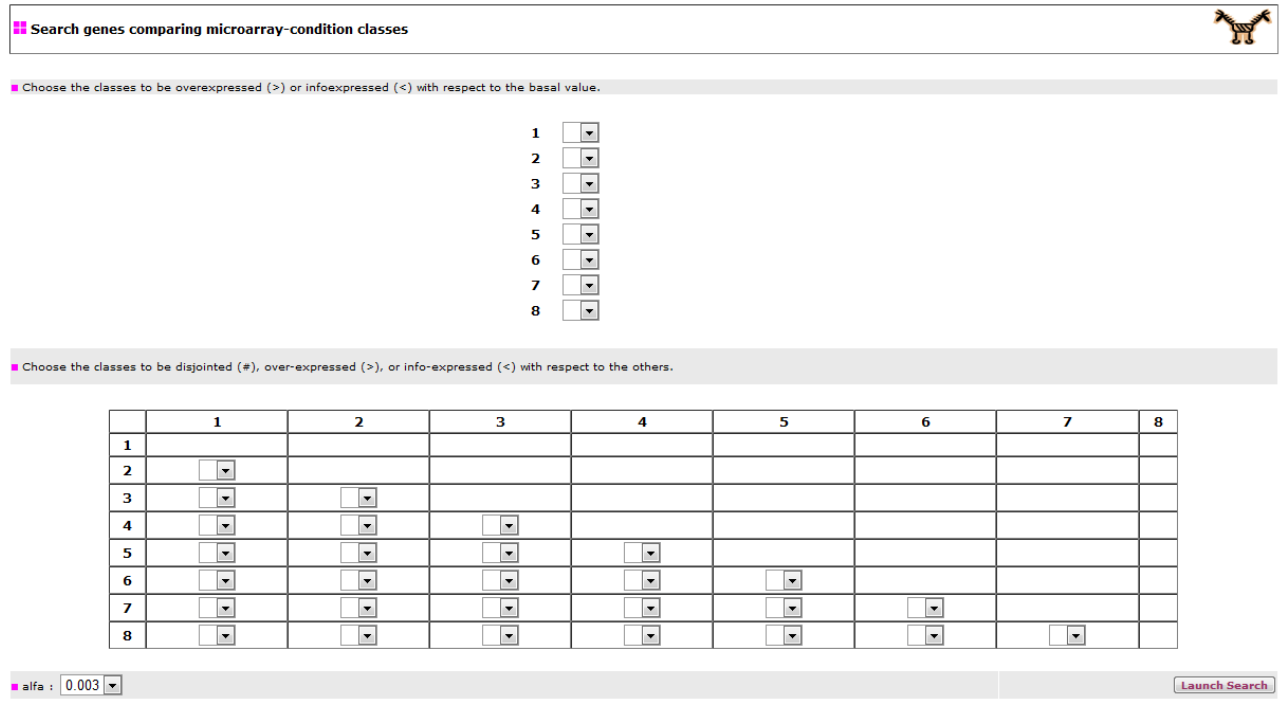
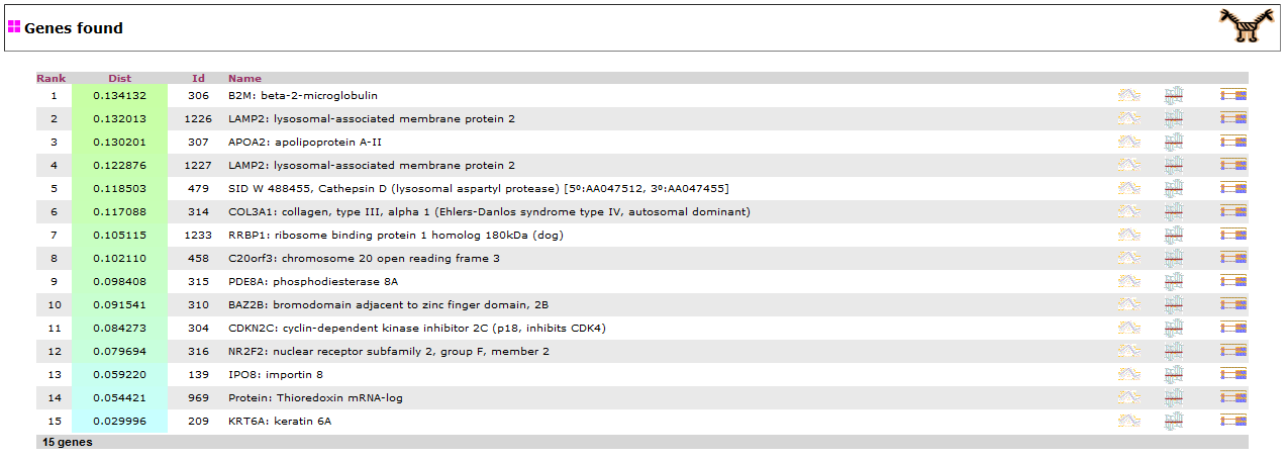


Figura 3.3: PHP de selección de parámetros de búsqueda para la aplicación NCR-PCOPGene. En el primer apartado se ajustan los requerimientos de expresión de cada clúster respecto al valor basal 0, es decir si se busca el cluster sobreexpresando o infoexpresando. En el segundo apartado se ajustan los requerimientos de expresión de cada clúster respecto el resto, es decir si un cluster está sobreexpresado, infoexpresado o diferentemente expresado (cualquiera de las opciones anteriores) respecto a otro cluster. El parámetro alfa selecciona el nivel de confianza. El botón Launch Search lanza la búsqueda de los genes marcadores para la distribución de clusters requerida.

El resultado obtenido de cumplir las condiciones requeridas por el usuario, muestra los genes marcadores que cumplen dichas condiciones, como se observa en la figura 3.4. El resultado obtenido en la figura es un ejemplo de cumplir las siguientes condiciones: sobreexpresar el clúster 2, el clúster 3 sin solapamiento del clúster 2 y sobreexpresar el clúster 4 sobre el 3.

A partir de los genes marcadores resultantes de la aplicación NCR-PCOPGene comenzará el inicio de mi aplicación. Siendo los genes marcadores que cumplen las condiciones del usuario la entrada de mi aplicación, donde se comparan los genes marcadores obtenidos con la base de datos de genes marcadores de microarrays. De esta manera ambas aplicaciones se fusionan.



Genes found

Rank	Dist	Id	Name
1	0.134132	306	B2M: beta-2-microglobulin
2	0.132013	1226	LAMP2: lysosomal-associated membrane protein 2
3	0.130201	307	APOA2: apolipoprotein A-II
4	0.122876	1227	LAMP2: lysosomal-associated membrane protein 2
5	0.118503	479	SID W 488455, Cathepsin D (lysosomal aspartyl protease) [5':AA047512, 3':AA047455]
6	0.117088	314	COL3A1: collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)
7	0.105115	1233	RRBP1: ribosome binding protein 1 homolog 180kDa (dog)
8	0.102110	458	C20orf3: chromosome 20 open reading frame 3
9	0.098408	315	PDE8A: phosphodiesterase 8A
10	0.091541	310	BAZ2B: bromodomain adjacent to zinc finger domain, 2B
11	0.084273	304	CDKN2C: cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
12	0.079694	316	NR2F2: nuclear receptor subfamily 2, group F, member 2
13	0.059220	139	IPO8: importin 8
14	0.054421	969	Protein: Thioredoxin mRNA-log
15	0.029996	209	KRT6A: keratin 6A

15 genes

Figura 3.4: PHP que muestra el resultado de la consulta a la aplicación NCR-PCOPGene. Lista los genes que cumplen las condiciones requeridas por el usuario. La columna Dist muestra la distancia de separación entre clusters. Esta distancia entre clusters mide la separación en términos de expresión de los clusters para cada gen marcador en la microarray del usuario (a mayor diferencia a la hora de expresar uno u otro clúster más marcador será el gen). La lista se ordena de mayor a menor distancia, pues los genes con mayor distancia serán los genes que más marcan la separación entre los clusters. El degradado de color de la columna Dist es proporcional al valor de la distancia entre clusters.

3.2 Construcción de la base de datos local de genes marcadores de microarrays

En esta segunda fase del trabajo se realiza la parte del preproceso del trabajo, ya que al partir desde cero en el proyecto, se necesita conseguir previamente la información. La información ha conseguir consiste en los genes marcadores resultantes de realizar la consulta construida mediante E-Utils en la sección 3.1.3 de fases del trabajo. La información descargada a modo de genes marcadores tienen clusters de información biomédica, que permiten enriquecer la información de clusters para la aplicación NCR-PCOPGene.

3.2.1 Descargar la información del NCBI mediante E-Utils

La primera fase de la descarga consiste en conseguir la lista de UIDs , que se realiza con eSearch hacia la base de datos Geoprofiles. A causa de que el lenguaje de programación utilizado en E-Utils es el Perl[12], opto por escoger Perl[19] como lenguaje ha utilizar para descargar y tratar la información. La consulta final queda de la siguiente manera:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=geoprofiles&term=value+subset+effect[Flag+Type]&retmode=xml&usehistory=y

12 Perl es un lenguaje de programación diseñado por Larry Wall en 1987. Perl toma características del lenguaje C, del lenguaje interpretado shell (sh), AWK, sed, Lisp y, en un grado inferior, de muchos otros lenguajes de programación.

A partir de activar la opción usehistory en la consulta, se pueden obtener el par de “claves” relacionadas (anteriormente comentadas: web environment y query key) con la consulta para relacionar con otra consulta, realizando un “pipeline”[13] entre ellas. De esta manera realizo una consulta sobre eSummary sin necesidad de descargarme primero la lista de UIDs de eSearch por completo. Este proceso se permite realizar gracias al par de “claves” de la primera consulta, así con una descarga consigo la lista de UIDs e información relevante sobre cada UID. La segunda consulta a E-Utils es:
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=geoprofiles&WebEnv=$web&query_key=$key&retstart=$retstart&retmax=$retmax&email=allen34@gmail.com

El par de "claves" que hacen referencia a $web(web environment) y $key(query key), son los valores necesarios para poder recordar al servidor la información requerida en la primera consulta. Los campos retstart y retmax son el intervalo indicativo de los genes marcadores ha descargar en una consulta del total de genes producidos por la primera consulta.

Como resultado de la descarga, se generan 1250 archivos (en formato XML). Donde cada archivo alberga 500 genes marcadores (marcado en el intervalo comentado) y su información asociada. El proceso de descarga tiene una duración aproximada de 2 horas.

El único problema que se podría encontrar en este método de descarga sería que la información descargada no sigue ningún orden, ya que como se ha comentado anteriormente se hace con un método ad-hoc del NCBI. Pero como los genes marcadores descargados son agrupados por la microarray a la que pertenecen, no carece de tanta importancia el orden entre ellos.

3.2.2 Tratar la información descargada de los genes marcadores

En esta segunda fase, el objetivo agrupar los genes marcadores descargados (alrededor de 600 mil) a la microarray a la que pertenece, construyendo la base de datos de genes marcadores de microarrays.

Como la información descargada esta en formato XML, es necesario instalar un módulo de Perl que permita leer un XML cómodamente. Perl tiene infinidad de módulos y librerías para solucionar el problema[14], como XML::Parser, XML::LibXML, XML::Twig, XML::Simple,etc. Incluso dispone de un módulo específico para bioinformática (BioPerl), pero que yo no lo utilicé. Finalmente me decidí por el módulo XML::Simple[21]. Porque como su nombre indica, es sencillo de utilizar para leer un XML. En el momento de instalarlo en el

13 Un pipeline consiste en un flujo de datos transmitido en un proceso comprendido por varias fases secuenciales, siendo la entrada de cada una la salida de la anterior.
14 http://search.cpan.org/

servidor tuve algún problema. No por el módulo, que se instaló correctamente con CPAN[15], sino por las librerías asociadas al módulo. Las librerías no acababan de instalarse correctamente en el servidor por problemas de dependencias y versión de Perl en el servidor. La solución conseguida fue instalar una versión más antigua de la librería, ya que inicialmente obtuve la más reciente.

A parte de agrupar cada gen marcador a la microarray correspondiente, se obtendrá su ID de la base de datos Gene. El ID es importante obtenerlo porque es el identificador de gen que permite relacionar cualquier base de datos del NCBI.

Para obtener el ID, primero se consulta sobre la base de datos local "geneinfo" del servidor implementada en MySQL[16] con los campos taxid(identificador de la especie) y el nombre del gen. Si esta consulta no es satisfactoria, entonces se realiza una petición a E-Utils mediante eLink, comentado anteriormente. Ya que con el UID de la base de datos GEO Profiles del gen se puede obtener el ID de la base de datos Gene, gracias a eLink[17].

He optado por este método para lograr el ID porque si existe dentro de la base de datos del servidor, la respuesta es más rápida que realizar una consulta a un servidor externo y tratar dicha información. Sino no hay otra opción de poderlo conseguir.

La manera de agrupar la información por microarray es también en formato XML, otras opciones estudiadas fueron archivos de texto y realizar una base de datos con MySQL. La opción de crear una base de datos no tenía mucho sentido, ya que se construiría una tabla por cada microarray existente. Al ser un contenido no relacional, no se aprovecharía la función básica de una base de datos en MySQL, relacionando por un o varios IDs de las tablas. La siguiente opción, almacenar la información en archivos de texto es semejante a tener la información en formato XML. Pero la lectura de un archivo de texto es secuencial y se tendrían que leer línea por línea el archivo en cuestión. Provocando un aumento de tiempo en los procesos posteriores cada vez que se necesitará realizar la lectura de un fichero de la base de datos de genes marcadores de microarrays. Por ello, la solución escogida para almacenar las microarrays en formato XML permite reducir el tiempo de lectura respecto a los archivos de texto, ya que se acceden por tags[18].

Inicialmente creaba un archivo por microarray existente, conteniendo por cada entrada un perfil de expresión (representado por UID). Los campos para cada perfil de expresión de la microarray son:

- Geneid, identificador de la base de datos Gene del NCBI.
- UID, identificador de la base de datos GEO Profiles del NCBI.
- GeneName, nombre del gen.

15 CPAN (Comprehensive Perl Archieve Network) es una gran colección de software y documentación PERL que, permite de un modo extremadamente sencillo la instalación de módulos Perl.
16 MySQL es un sistema de gestión de bases de datos relacional, multihilo y multiusuario.
17 Ejemplo con el UID 9239179: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=geoprofiles&db=gene&id=9239179&email=allen34@gmail.com
18 Un tag o etiqueta es una marca con tipo que delimita una región en los lenguajes basados en XML.

- Title, título de la microarray.
- Summary, resumen de la microarray.
- Taxid, identificador de la microarray.
- Idref, identificador del perfil en la microarray.

Contener la información de esta manera, no fue una buena opción. Los campos title, summary y taxid tienen la misma información en cada perfil de expresión de la microarray. De esta manera, tampoco podía saber la cantidad de genes por microarray que almacenaba el XML. Entonces decidí dividir la información de una microarray en dos ficheros XML, uno almacena los perfiles de expresión de la microarray e información única y el otro fichero guarda la información referente a la propia microarray, es decir, el title, el summary y el taxid. Además de esta manera, por cada perfil de expresión guardaba más de un nombre del gen en el campo GeneName y decidí separar los perfiles de expresión por los nombres del gen.

Observando los campos existentes en la información descargada, decidí completar la información añadiendo un par de campos más por cada archivo de microarray. El resultado final de los datos en una microarray es:

- Fichero XML que contiene los perfiles de expresión o genes marcadores con los campos:
  - Geneid, identificador de la base de datos Gene del NCBI.
  - UID, identificador de la base de datos GEO Profiles del NCBI.
  - GeneName, nombre del gen.
  - Idref, identificador del perfil en la microarray.
  - Alias, campo nuevo que contiene otros posibles nombres del gen.
  - GeneDesc, campo nuevo que describe de forma breve el gen.
- Fichero XML que contiene información referente a la propia microarray, con los campos:
  - Title, título de la microarray.
  - Summary, resumen de la microarray.
  - TaxId, identificador de la especie de la microarray.
  - Taxon, campo nuevo que contiene el nombre de la especie.
  - Count, campo nuevo que alberga la cantidad de genes marcadores del anterior fichero.

La información descrita permite mostrar todo lo necesario por cada microarray y cada gen marcador en la aplicación web realizada.

La información descargada en números a modo de estadística son 630 mil perfiles descargados divididos en 2510 microarrays. Por cada microarray hay 250 genes marcadores de media. El proceso de agrupar la información y tratarla tiene una duración aproximada de 5 horas, este tiempo es debido a la gran transmisión de datos que se realiza con ficheros.

3.3 Cálculos online

Una vez obtenidos los datos, llega el momento de realizar el programa que cruza genes marcadores en tiempo real (online) en la aplicación web.

La función principal es cruzar los genes marcadores de la microarray del usuario como entrada del programa con la base de datos de genes marcadores de microarrays construida en la fase 3.2. La salida del programa es la compatibilidad de los genes marcadores de entrada con los genes marcadores de cada una de las microarrays de la base de datos, mostrando también cuales son los genes marcadores en común.

El programa trata ficheros de texto, como los genes marcadores de entrada o de salida, y ficheros en formato XML por parte de la base de datos de microarrays. Por ello se ha implementado en lenguaje C [19], porque permite tratar cómodamente el tipo de ficheros mencionados y se caracteriza por su velocidad de ejecución (necesario en un programa "online"). Para poder leer un XML se utiliza la librería libxml[122] de C.

El cruce entre genes marcadores se realiza a partir del nombre del gen. Para mejorar o filtrar el resultado obtenido, el programa permite una serie de opciones de búsqueda:

- Búsqueda simple: compara el nombre del gen marcador de entrada con el nombre del gen de la microarray de la base de datos. Se realiza para la totalidad de la base de datos de genes marcadores de microarrays.
- Búsqueda por alias: además de incluir la búsqueda simple, también compara el nombre del gen marcador de entrada con todos los nombres asociados al perfil de expresión de la microarray (alias).
- Búsqueda por filtro: realiza una búsqueda simple o por alias, pero sólo para un conjunto de microarrays de la base de datos de microarrays.

El proceso del programa se observa en el siguiente diagrama de flujo, figura 3.5. Donde primero se realiza una lectura del fichero del conjunto de genes marcadores que se cruzan con los genes marcadores de la base de datos de microarrays. Los nombres de los genes marcadores introducidos son almacenados en un array para poder realizar más adelante la comparación. Seguidamente, comienza el cálculo para encontrar compatibilidades. Este proceso es un bucle por cada elemento del listado de microarrays. Por cada microarray se leen sus genes marcadores y por el nombre del gen se compara con el array descrito anteriormente, teniendo en cuenta las opciones de búsqueda mencionadas. Si está activada la opción de alias, además se compara cada alias con el array de entrada. Si el resultado es positivo se escribe en el fichero que lista las compatibilidades de microarrays y los genes marcadores en común en otro fichero. Así

19 C es un lenguaje de programación creado en 1972 por Dennis M.Ritchie en los Laboratorios Bell como evolución del anterior lenguaje B. Es un lenguaje orientado a la implementación de Sistemas Opertaivos, concretamente Unix.

sucesivamente para cada elemento del listado de microarrays.

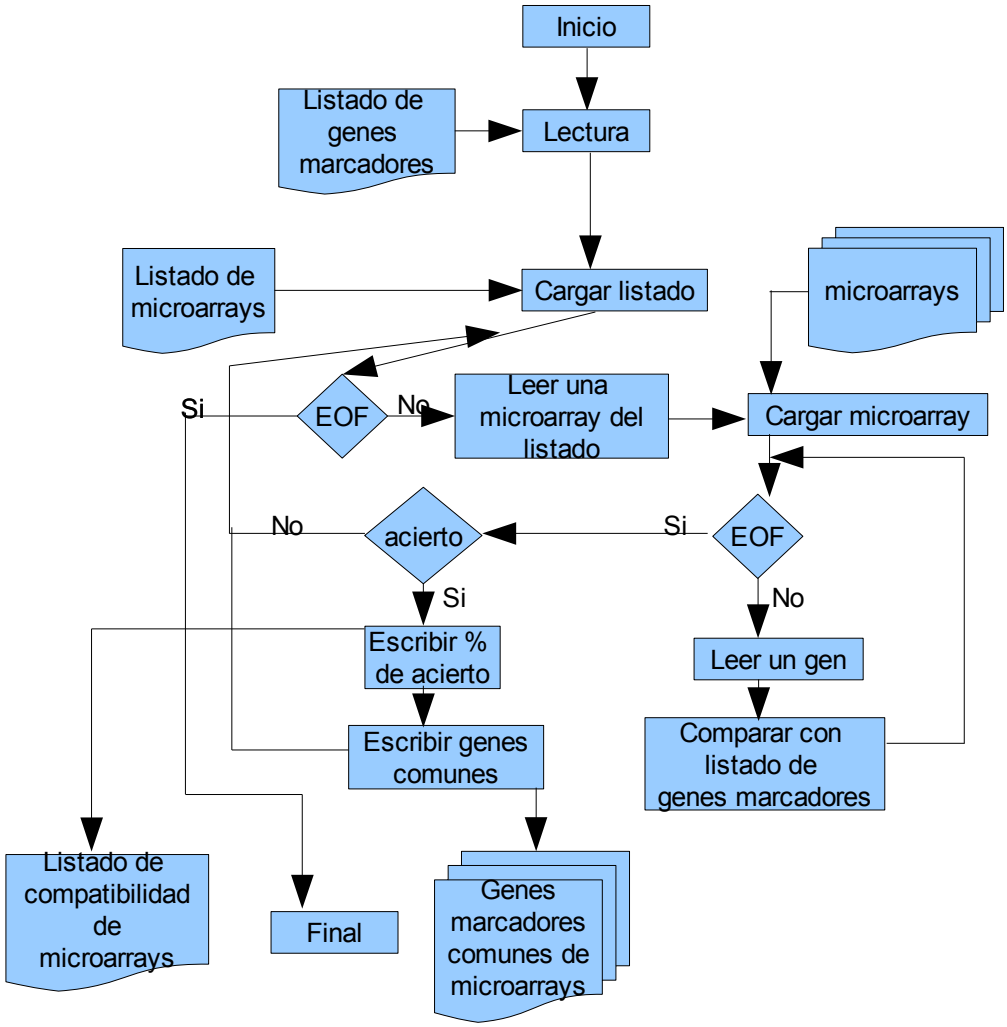


Figura 3.5: Diagrama de flujo del comportamiento del programa para realizar el cruce entre genes marcadores. Primero se leen los genes marcadores y el listado de microarrays de la base de datos. Por cada microarray del listado se comparan sus genes marcadores con los genes marcadores de la consulta sobre la microarray del usuario. Si supera el treeshold se escribe en un fichero el porcentaje de genes marcadores comunes entre la microarray del listado y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario. Además, por cada comparación que supere el treeshold se escribe en ficheros aquellos genes marcadores comunes.

Puntualizar sobre el fichero de salida llamado listado de compatibilidad de microarrays que almacena *el porcentaje* de genes marcadores comunes entre la microarray del listado y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario y también el porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray listada. Así se puede interpretar el nivel de acierto de ambos una vez mostrados en la interfície web.

El método para comunicar el programa con el aplicativo web, referente a los genes marcadores, son

los ficheros de salida mediante el UID (identificador único de GEO Profiles) y el índice del array que almacena el conjunto de genes marcadores que son cruzados con la base de datos. El array creado en el programa para guarda los genes marcadores de entrada tiene su homólogo en la aplicación web. Por eso, guardando el índice del array permite posteriormente a la aplicación web seleccionar directamente el gen marcador.

El proceso desarrollado es un proceso final, ya que en el transcurso de la elaboración se fue modificando para poder optimizar el tiempo de respuesta del programa y eficacia. Las modificaciones son:

- Mejora del tiempo de respuesta del programa: Inicialmente tenía una duración de 35 segundos en el servidor[2], pero a causa de reducir el número de arrays a tratar, de rectificar los bucles de búsqueda de las microarrays y de los genes marcadores logré reducir el tiempo alrededor de una séptima parte del tiempo inicial.
- Eliminación de las microarrays inferiores a un 75% de acierto en el resultado: A priori introduje un corte en el resultado. El corte consistía en mostrar como resultado aquellas microarrays superiores al 75% de acierto en la base de datos de microarrays. Posteriormente me di cuenta que era un porcentaje muy restrictivo observando los resultados mostrados al aplicar un test de pruebas. Por ello considere suprimirlo.

A partir de las modificaciones aplicadas y el desarrollo descrito, se aprecia un resultado bastante satisfactorio para un programa en tiempo real.

3.4 Aplicación Web

La aplicación web se ha desarrollado en un servidor diferente, llamado http://devresearch.uab.es. Es el servidor de pruebas, donde se realizan e implementan las interfícies web antes de integrarlo en el servidor principal, http://revolutionresearch.uab.es.

3.4.1 Generación de imágenes a partir de los valores de expresión de los genes marcadores

En la aplicación web, aparece una imagen por cada gen marcador que representa los niveles de expresión del gen para cada clúster de una microarray concreta. Para poder generar dicha imagen utilicé una herramienta implementada por un compañero que me genera los genes marcadores de la base de datos de microarrays y los genes marcadores de la microarray de interés por el usuario. La herramienta genera la imagen a partir de introducir los valores de expresión de cada uno de los clústers. Con la imagen

se puede observar los niveles de expresión del gen para cada cluster de la microarray, permitiendo ver como se expresa cada uno de los clusters a modo de imagen.

El proceso de generación de imágenes necesita previamente descargar del FTP[20] del NCBI las microarrays existentes en la base de datos de genes marcadores de microarrays. Las microarrays del FTP contienen los valores de expresión de todos los genes de la microarray. Por ello, una vez descargadas se tienen que seleccionar sólo los genes que aparecen en la base de datos de genes marcadores de microarrays.

La microarray del FTP se tiene que parsear debidamente para obtener las diversas distribuciones de clusters que contiene, con la intención de generar un archivo colors[21] por distribución de clusters. Un archivo colors representa la distribución de clústers respecto las condiciones muestrales del experimento.

En la implementación para parsear la microarray descargada desde el FTP tuve algunas dificultades para generar el archivo colors. El problema consistía en realizar un código preparado para la infinidad de distribuciones de clusters que puede haber en las diferentes microarrays.

Por cada microarray puede haber diversas distribuciones de clusters, formadas por condiciones muestrales muy dispersas dado un orden determinado. Tampoco todas las distribuciones de clusters son una buena representación, ya que un clúster con menos de 4 condiciones muestrales no se considera una agrupación válida. Pero conseguí obtener el resultado esperado y de manera eficiente con una serie de bucles y arrays que entrelazados entre sí obtenía crear los archivos colors.

Seleccionados los genes, sus valores de expresión que aparecen en la microarray de la base de datos de microarrays y construido el archivo colors para la microarray, son introducidos a un programa que cálcula la distribución t-Student[22] con valores alfa de 0.009 y probabilidad del 0.991 aplicando un intervalo de confianza[23] del 95%. Obteniendo como respuesta los valores de mínimo (min), máximo (max) y media (avg) de cada clúster existente por cada gen de la microarray.

El programa que calcula los valores (min, max y avg), ha sido la reutilización de otro programa existente de un compañero que se utiliza para la aplicación web NCR-PCOPGene. El programa se encarga de calcular la distribución t-Student aplicando un intervalo de confianza para poder seleccionar que clusters se sobreexpresan o infoexpresan según el criterio del usuario que utiliza el aplicativo para una microarray, obteniendo los genes que cumplían dichas condiciones con los clusters. Del programa original se ha extraído el cálculo referente a distribución t-Student y intervalo de confianza, modificándolo para obtener los valores de mínimo,máximo y media de los clusters y para todo los genes de la microarray introducida. A causa de algunos clusters formados por valores de expresión en las condiciones muestrales dispersas y que

20 http://ftp.ncbi.nih.gov/pub/geo/DATA/SOFT/GDS/

21 Un archivo colors se representa con dos columnas. La primera columna indica el número de la condición muestral, empezando por el 1. La segunda columna es el número del clúster que pertenece la condición muestral.

22 La distribución T (de Student) es una distribución de probabilidad que surge del problema de estimar la media de una población y ésta debe ser estimada a partir de los datos de una muestra.

23 Se llama intervalo de confianza a un par de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto.

no se focaliza correctamente con el cálculo comentado, se ha restringido el intervalo de confianza para estos casos concretos. Estos casos se producen cuando los valores (min y max) obtenidos no reflejan adecuadamente el intervalo de concentración de muestras, es decir, los valores (min y max) que corresponden al límite inferior y superior son menores que el valor mínimo de las muestras y mayores que el valor máximo de las muestras. Entonces los valores de las muestras mínimo y máximo pasan a ser los límites del clúster, realizando y focalizando mejor el intervalo del conjunto de condiciones muestrales en un clúster.

Elaborados los valores de una microarray, se estructuran para ser introducidos en la herramienta que genera las imágenes que muestran los niveles de expresión del gen para cada clúster de la microarray. Además la herramienta permite indicar el punto inicial, el valor basal. Otra opción ha introducir son los centros de las distribuciones de los clusters. Los centros corresponden a la cantidad de condiciones muestrales agrupadas por cluster respecto al total de condiciones muestrales de la microarray. La medida de centros tomada no conseguía resaltar claramente los centros en las distribuciones de clusters, con lo que decidí sumar un factor de corrección para mejorar la interpretación visual del usuario.

Un ejemplo de imagen es la mostrada en la figura 3.6. Donde la imagen representa la distribución de dos clusters a lo largo del rango de expresión de un gen. Un clúster pintado de color rojo donde se puede apreciar una mayor dispersión de los valores dada la largura de la línea. Otro clúster pintado de color lila que muestra una concentración mayor de sus muestras. El cluster lila se sobreexpresa respecto al cluster rojo.

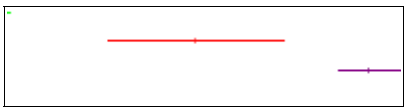


Figura 3.6: Imagen que representa la distribución de dos clusters a lo largo del rango de expresión de un gen. Un clúster pintado de color rojo donde se puede apreciar una mayor dispersión de los valores dada la largura de la línea. Otro clúster pintado de color lila que muestra una concentración mayor de sus muestras. El clúster lila se sobreexpresa respecto al cluster rojo para este gen y microarray en cuestión. En la parte superior izquierda del imagen se puede ver el valor basal o valor 0 (de color verde).

3.4.2 Diseño e implementación de la interfaz web

Una vez construida la base de datos de genes marcadores de microarrays, realizado el programa de cruce entre genes marcadores, generado las imágenes asociadas a la base de datos de genes marcadores de microarrays, elaboré la herramienta que relaciona lo mencionado anteriormente a modo de aplicación web.

El proceso de diseño e implementación de la interfaz web contiene tres subsecciones. Consisten en conectar dos aplicaciones, una vista general donde se muestran las microarrays compatibles y una vista

detalle donde aparecen los genes marcadores de la microarray seleccionada.

3.4.2.1 Conexión entre aplicaciones web

Antes de realizar la aplicación propiamente dicha, hay que conectar la aplicación NCR-PCOPGene con mi aplicación. La aplicación NCR-PCOPGene busca los genes marcadores de la microarray del usuario con uno u otro clúster sobreexpresado o infoexpresado respecto al resto, a petición del usuario. Mostrando los genes marcadores resultantes, donde modifico la página web en formato php[24] para enlazar ambas amplicaciones.

La modificación consiste en escribir en ficheros aquella información relevante para mi aplicativo e insertar el enlace entre ambas. La información relevante son el listado de genes marcadores para la distribución de clusters requerida por el usuario y la distancia que mide la separación en términos de expresión de los clusters para el gen marcador y la microarray del usuario (a mayor diferencia a la hora de expresa uno u otro clúster más marcador será el gen). Como se puede observar en la figura 3.7 que muestra la página web resultante del NCR-PCOPGene, donde aparece un listado de genes marcadores para la distribución de clusters requerida por el usuario. El listado está ordenado de mayor a menor según la distancia que mide el mejor candidato a ser gen marcador. También se aprecia como se ha incorporado el enlace entre ambas aplicaciones continuando la línea del servidor en la parte inferior de la figura 3.7.

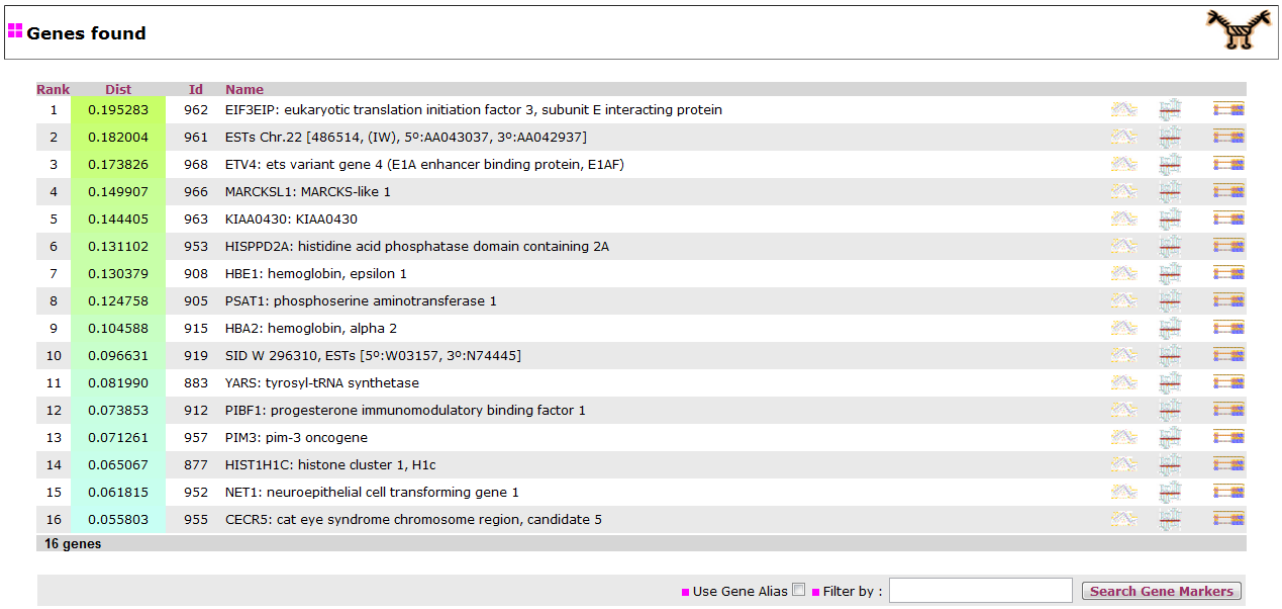


Figura 3.7: PHP que muestra el resultado de la consulta a la aplicación NCR-PCOPGene. Este listado muestra los genes marcadores para la distribución de clusters requerida por el usuario. En la parte inferior aparece el botón para lanzar el cruce de estos genes marcadores resultado y los genes marcadores de la base de datos de genes marcadores de microarrays. Al lado aparecen las opciones de alias, para realizar un cruce de genes marcadores considerando todas las posibles nomenclaturas de cada gen, y de filtrado de

24 PHP es un lenguaje de programación interpretado, diseñado originalmente para la creación de páginas web dinámicas.

microarrays por palabra clave, lo que solo compararía los genes resultado con genes marcadores para microarrays que estudien del cáncer de colon (si como palabra clave usásemos "colon cáncer").

El enlace a parte de conectar las dos aplicaciones, permite al usuario poder activar otros métodos de búsqueda explicados en la sección 3.2.1 correspondiente al programa de cálculo online. El método de búsqueda por alias, se habilita mediante un checkbox y el método de filtrado, consiste en un cuadro de texto donde poder escribir la palabra o palabras claves para filtrar las microarrays.

La búsqueda para filtrar microarrays se realiza con una búsqueda de la palabra o palabras claves introducidas en el cuadro de texto, por los campos Title y Summary de la base de datos de genes marcadores de microarrays, con la intención de reducir el conjunto de microarrays. Inicialmente proveé de realizar la búsqueda a partir de E-Utils, que consistía en realizar una consulta con la palabra o palabras claves. La respuesta era una lista con los nombres de la microarrays que coincidían con dicha búsqueda. El proceso de E-Utils lo descarté por el excesivo tiempo de cómputo que requería para un cálculo en tiempo real. El proceso de la consulta tenía una duración de 40 segundos (descargar y tratar), un resultado excesivo. Por ello opté por la opción de búsqueda sobre la información de la base de datos de genes marcadores de microarrays, ya que es más rápido (4 segundos como máximo).

Para permitir que un usuario pudiese realizar diferentes pruebas sobre la aplicación utilicé un distintivo entre consulta y consulta. El distintivo es la fecha, cogiendo hora,minuto y segundo de la consulta realizada. Así en la posterior generación de archivos temporales para el uso de la aplicación web no produce ningún problema entre consulta y consulta.

Al lanzar mi aplicación mediante el enlace descrito se ejecuta el programa de cálculo online y se generan las imágenes de las distribuciones de clusters de condiciones muestrales correspondientes a los genes marcadores de la microarray del usuario. Para amenizar la espera de la ejecución de los programas en la página web, se muestra un gif animado de carga de resultados como se observa en la figura 3.8.

Figura 3.8: Gif animado que se muestra mientras se realiza el cruce de genes marcadores.

La herramienta de un compañero que genera la imagen de distribuciones de clusters de condiciones muestrales de un gen, inicialmente estaba implementa en Python[25]. Dicha herramienta se lanzaba desde un

25 Python es un lenguaje de programación de alto nivel cuya filosofía hace hincapié en una sintazis muy limpia y que favorezca un código legible.

programa en el servidor web (Apache[26]) para generar las imágenes de cada gen. Pero al lanzar un programa en Python desde Apache no se ejecutaba dicho programa. Soluciones intentadas para resolver el problema de ejecución de Python desde Apache:

- Intenté solucionarlo instalando el módulo correspondiente de Python (mod_python) en el servidor web.
- Comprobé los permisos del módulo de Python instalado, eran los mismo que el resto de módulos instalados.
- Comprobé los permisos del programa, de propietario, de grupo y de ejecución, estaban correctos.
- Comprobé el archivo de configuración del servidor web Apache, situado en /etc/httpd/conf/httpd.conf. La ruta de carga del módulo de Python era incorrecta, la corregí. Comprobé que estuviera habilitado los archivos .py en el servidor. Comprobé la ruta de situación del intérprete de Python.
- En el programa forcé la ruta del intérprete de Python.

Después de intentar solucionar el problema con las posibles soluciones mencionadas, no conseguí lanzar el programa en Python desde el servidor web Apache. El compañero que realizó la herramienta, también intentó encontrar alguna solución pero sin éxito. Al final el codirector decidió que el compañero cambiase el programa de Python a C, lenguaje que se ejecuta perfectamente desde el servidor web.

3.4.2.2 Vista general de la interfaz web

Al finalizar la ejecución de los programas de cruce de genes marcadores y generador de imágenes de los genes marcadores de la microarray del usuario, se muestran los resultados obtenidos. Los resultados se presentan en una tabla. La tabla contiene las microarrays compatibles con los genes marcadores de la microarray del usuario.

La vista general muestra los resultados obtenidos en el programa de cruce a nivel de microarrays. Aunque la comparación se realiza entre genes marcadores, los genes marcadores de la base de datos están agrupados por microarrays. En el siguiente punto se explica la vista detalle, donde se muestran los genes marcadores.

La tabla almacena toda la información relevante sobre las microarrays resultantes del cruce online. A continuación se describe el diseño y funcionamiento de la vista general con la ayuda de las figuras 3.9, 3.10, 3.11 y 3.12, donde se observa la estructura de la tabla y la apariencia global de la vista general. Manteniendo la armonía del servidor y los colores característicos como el lila y el azul.

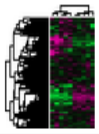
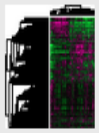
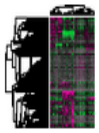
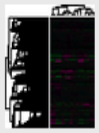
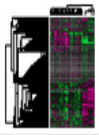
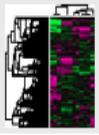
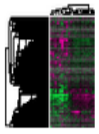
26 El servidor HTTP Apache es un servidor web HTTP de código abierto para plataformas Unix, Microsoft, Macintosh y otras, que implementa el protocolo HTTP/1.1 y la noción de sitio virtual.

En la vista general se recoge el resultado obtenido en el programa de cálculo online y la información de la base de datos de microarrays, dividiendo la información por columnas y cada una de las microarrays por filas.

La tabla contiene un total de 7 columnas, se puede ver más claramente en la figura 3.9. Las columnas de izquierda a derecha son:

- GDS: El nombre de la microarray.
- % user gds: El porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario (la lista de microarrays se ordenará inicialmente por este campo)
- %matching gds: El porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray listada
- Title & Summary: El título de la microarray, un breve resumen sobre la microarray y las distribución de clústers de la microarray.
- GDS Analysis: Imagen de la distribución de clusters de la microarray, figura 3.11.
- % gene markers: El porcentaje de genes marcadores de la microarray de la base de datos de microarrays entre el total de genes de esta microarray
- Link: La última columna lanza la vista detalle de la microarray en cuestión

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
GDS3257	37.500000	2.678571	Cigarette smoking effect on lung adenocarcinoma *[Homo sapiens]* Analysis of different tumor stage adenocarcinoma and paired normal lung tissues of current, former and never smokers. To date, tobacco smoking is responsible for over 90% of lung cancers. Results provide insight into the molecular basis of lung carcinogenesis induced by smoking. **Subsets: 2 tissue ,3 individual ,4 disease state ,2 gender ,2 other sets.**		1.005251
GDS2255	37.500000	4.687500	Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung *[Homo sapiens]* Analysis of neutrophils that transmigrate to the alveolar space. Transmigration induced by bronchoscopic instillation of volunteers with endotoxin (LPS). Results identify differences between pulmonary and circulating neutrophils that occur early in endotoxin-induced lung inflammation. **Subsets: 3 cell type ,2 agent ,18 individual sets.**		0.574429
GDS2214	37.500000	2.857143	Septic neutrophil response to lipopolysaccharide and high mobility group box 1 protein *[Homo sapiens]* Analysis of septic neutrophils treated with lipopolysaccharide or high mobility group box 1 (HMGB1) protein. Neutrophils isolated from patients with sepsis-induced acute lung injury (ALI). HMGB1 is a late mediator of endotoxin lethality, and neutrophils play a role in endotoxemia-associated ALI. **Subsets: 3 agent ,8 individual sets.**		0.942422
GDS1673	37.500000	2.083333	Non-diseased lung tissue *[Homo sapiens]* Analysis of non-diseased lungs from 23 multi-organ donors. Upper and lower lobe peripheral sections of the lung were examined. Donors varied in age, sex, smoking history, and ethnicity. Results provide a reference for microarray studies of pulmonary disease. **Subsets: 2 tissue ,2 gender ,3 stress ,6 age ,3 other sets.**		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%

Figura 3.9: PHP correspondiente a la vista general. El PHP lista las micrarrays que comparten genes marcadores con la microarray de interés del usuario. La primera columna muestra el identificador de la microarray y es un link a GEO Datasets para mostrar todo lo referente a la microarray en cuestión. La segunda columna corresponde al porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario (la lista de microarrays se ordenará inicialmente por este campo). La tercera columna muestra el porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray listada. La cuarta columna muestra el título y un resumen de la microarray. La quinta columna muestra la distribución de clusters de la microarray. La sexta columna es el porcentaje de genes marcadores de la microarray de la base de datos por el total de genes de esta microarray. El icono de la última columna lanza la vista detalle de la microarray en cuestión.

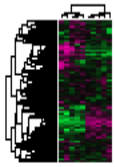
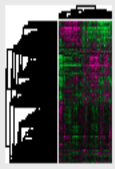
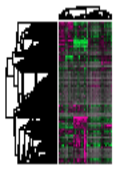
GDS	% user gds	% matching gds	Title & Summary	GDS Analysis	% gene markers
GDS1312	43.750000	0.664137	Squamous lung cancer *[Homo sapiens]* Expression profiling of squamous lung cancer biopsy specimens and paired normal specimens from 5 patients. Differentially expressed genes integrated with protein interaction maps. Results suggest that differentially expressed genes are highly connected through protein interactions. **Subsets: 2 disease state ,5 individual sets.**		4.730063
GDS3257	37.500000	2.678571	Cigarette smoking effect on lung adenocarcinoma *[Homo sapiens]* Analysis of different tumor stage adenocarcinoma and paired normal lung tissues of current, former and never smokers. To date, tobacco smoking is responsible for over 90% of lung cancers. Results provide insight into the molecular basis of lung carcinogenesis induced by smoking. **Subsets: 2 tissue ,3 individual ,4 disease state ,2 gender ,2 other sets.**		1.005251
GDS2255	37.500000	4.687500	Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung *[Homo sapiens]* Analysis of neutrophils that transmigrate to the alveolar space. Transmigration induced by bronchoscopic instillation of volunteers with endotoxin (LPS). Results identify differences between pulmonary and circulating neutrophils that occur early in endotoxin-induced lung inflammation. **Subsets: 3 cell type ,2 agent ,18 individual sets.**		0.574429

Figura 3.10: Imagen ampliada de las tres primeras microarrays mostradas en la figura 3.9. Donde se aprecia mejor el contenido de la tabla. La primera columna muestra el identificador de la microarray y es un link a GEO Datasets para mostrar todo lo referente a la microarray en cuestión. La segunda columna corresponde al porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray del usuario (la lista de microarrays se ordenará inicialmente por este campo). La tercera columna muestra el porcentaje de genes marcadores comunes entre la microarray listada y la microarray del usuario respecto el total de genes marcadores de la microarray listada. La cuarta columna muestra el título y resumen de la microarray. La quinta columna muestra la distribución de clusters de la microarray (tanto de genes como de condiciones muestrales). Esta imagen se amplía al pasar el ratón por encima. La sexta columna es el porcentaje de genes marcadores de la microarray de la base de datos de microarrays entre el total de genes de esta microarray. La flecha de la última columna lanza la vista detalle de la microarray en cuestión.

Las columnas descritas permiten ordenar la tabla según el criterio del usuario, con la función tablesorter[27] de Javascript[28] de la librería jQuery[29]. Al cargar la página web se ordena por la segunda columna, % user gds reflejando el porcentaje de genes marcadores de la microarray del usuario encontrados en la microarray de la base de datos de genes marcadores de microarrays. Se permite ordenar por el resto de columnas descritas.

Cada una de las filas o celdas de la tabla alberga una microarray compatible con los genes

27 http://tablesorter.com/docs/
28 JavaScript es un lenguaje de scripting basado en objetos, utilizado para acceder a objetos en aplicaciones. Principalmente, se utiliza integrado en un navegador web permitiendo el desarrollo de interfaces de usuario mejoradas y páginas web dinámicas.
29 http://jquery.com/

marcadores de la microarray del usuario. Las filas de la tabla se muestran en dos colores para facilitar la visión del usuario, intercalando ambos colores como una zebra. Aunque se ordene la tabla por otra columna a la inicial, la visión de colores en zebra se mantendrá gracias a la función tablesorter de Javascript.

Si el investigador necesitase más información sobre la microarray, el nombre de la microarray correspondiente a la columna GDS es un link[30] que abre una nueva pestaña hacia el NCBI sobre la microarray seleccionada. También la columna GDS Analysis (figura 3.11) que corresponde a la imagen de la microarray es un link[31] hacia el NCBI para apreciar con más claridad las distribución de clusters de la microarray y la concentración de muestras.

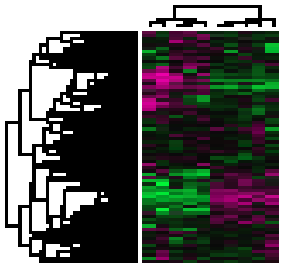


Figura 3.11: Imagen ampliada de la distribución de clusters de la microarray GDS1312. Correspondiente a la imagen mostrada en la columna GDS Analysis de las figuras 3.9 y 3.10. Las filas son los genes y las columnas las condiciones muestrales. En la figura aparece un hierárchical clustering tanto de genes como de condiciones muestrales. Estos son unos clusters de origen estadístico. Clicando sobre la figura accedes a una interfaz de GEO Datasets que te permite navegar por la figura gen a gen.

Para enriquecer el aspecto visual de la interfaz web, en la imagen de la microarray se ha incorporado una funcionalidad (figura 3.11). La funcionalidad sucede con el ratón situado encima de la imagen y se aumenta con un efecto suave para una mejor visualización. El efecto pertenece a la librería Dojo[32] en JavaScript.

En figura 3.12, aparece una línea que muestra el número de microarrays resultantes con un porcentaje superior al 10% y también las inferiores. Se realiza esta separación entre las microarrays resultantes para mostrar inicialmente aquellas microarrays superiores y si el usuario cree conveniente ver las inferiores, darle la posibilidad de hacerlo.

El botón que permite mostrar las microarrays inferiores al 10% se encuentra en la parte inferior derecha de la figura 3.10, llamado Show/Hide %user-gds < 10%. Como su nombre indica, muestra y oculta las microarrays inferiores. Dicha acción se realiza también con una función implementada en Javascript, teniendo en cuenta sólo las filas que corresponden a las microarrays inferiores.

30 Ejemplo para la microarray GDS3179: http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3179
31 Ejemplo para la microarray GDS3179: http://www.ncbi.nlm.nih.gov/projects/geo/gds/analyze/analyze.cgi?ID=GDS3179
32 www.dojotoolkit.org/

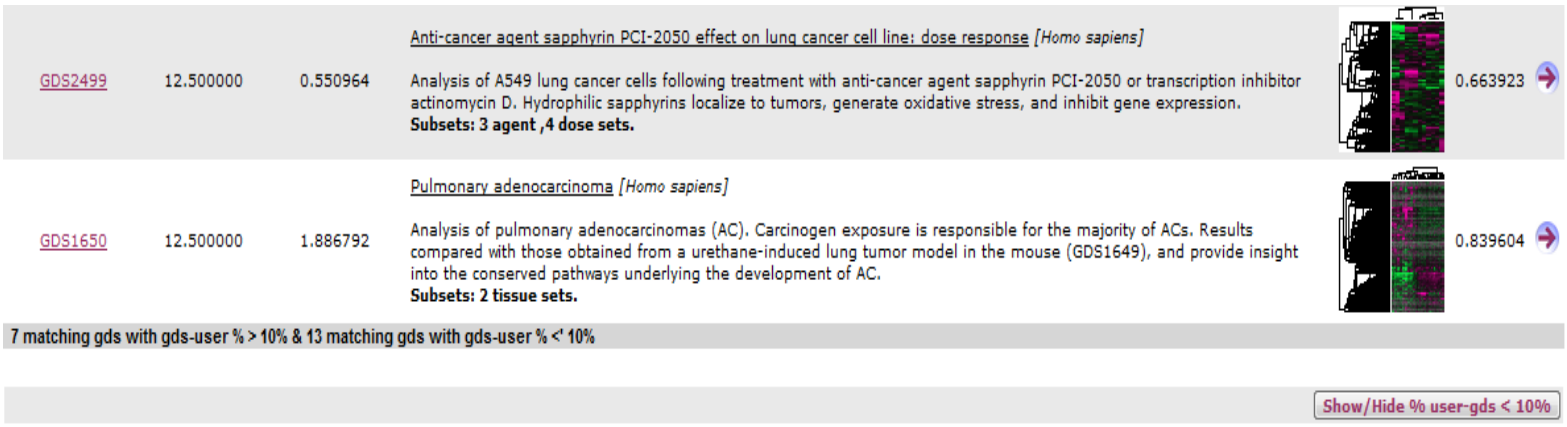


Figura 3.12: Imagen ampliada de la parte inferior de la figura 3.9. La última línea de la tabla indica el número de microarrays superiores al 10% de compatibilidad entre los genes marcadores de la microarray del usuario y sus genes. El botón "Show/Hide % user-gds < 10%" permite mostrar o ocultar las microarrays inferiores al 10% de compatibilidad.

Los diferentes efectos y funcionalidades aplicadas a la página web mediante Javascript me provocó problemas al complementar las funciones diferentes en una misma página web. Cada función utiliza una librería diferente de JavaScript, por lo que hay que tener en cuenta el orden en que se ejecutan y cual de ellas se actualizan al suceder algún evento sobre la página. Primero se carga con la página web el efecto de aumentar la imagen y una vez cargada la información que contiene la página web a modo de tabla, se ocultan las microarrays inferiores y finalmente se utilizan las funciones de ordenar para concluir con el aspecto que se muestra en la figura 3.9.

Sino se carga en primera instancia el efecto de aumentar la imagen con la página web, provoca un mal funcionamiento del efecto en algunos navegadores (por ejemplo, Google Chrome o Explorer) y afecta al resto de funciones JavaScript. Al realizar alguna acción sobre la página produce un evento y requiere actualizar la función de ordenar la información y realizar el zebrado en las filas, porque sino se deriva en un mal funcionamiento del comportamiento de la función. Como por ejemplo, ordenar incorrectamente la columna seleccionada o no realizar el zebrado sobre las filas.

El diseño mostrado en la figura 3.9 es el diseño final de la interfaz web. Inicialmente carecía de algunos aspectos que se introdujeron para mejorar el aspecto informativo, funcional y visual. A continuación describo aquellos aspectos que inicialmente no se incorporaban en la página web:

- Ordenar por cualquier columna: A priori opté por mostrar la información ordenada solamente por la segunda columna (% user gds, la ordenación inicial en este momento), siendo una columna estática. Pero pensé que sería más dinámico y de más utilidad poder permitir al usuario poder ordenar la información según su criterio.
- Efecto de aumento de la imagen: Estimé que era un efecto atractivo visualmente para la página

web, aportando más dinamismo.

- Mostrar las microarrays inferiores al 10% de compatibilidad: Inicialmente sólo se mostraban las microarrays con un resultado superior al 10%, pero consideré que quizá sería de curiosidad e interés para el investigador poder ver aquellas microarrays inferiores. Dando así más información de microarrays al usuario.
- Incluir la columna % gene markers: Consideré oportuno incluir más información sobre las microarrays, ya que la nueva columna muestra el porcentaje de genes marcadores de la microarray de la base de datos de microarrays entre el total de genes de esta microarray
- Completar la información mostrada en la columna Title & Summary: Continuando con la finalidad del punto anterior, dar mas información al investigador sobre las microarrays. Opté por acompañar el título de la microarray con el nombre de la especie a la que pertenece (en cursiva) la microarray y añadir la distribución de clusters (en negrita) de la microarray.

3.4.2.3 Vista detalle de la interfaz web

A partir del enlace mostrado en la última columna de la vista general se accede a la vista detalle. En la vista detalle se muestran los genes marcadores comunes (o no) entre la microarray seleccionada y los genes marcadores de la microarray de interés.

Continuando con la estructura de la vista anterior, la información referente a los genes marcadores se almacena en una tabla, siguiendo la misma línea como se observa en la figura 3.13. La tabla esta formada por columnas que dividen la información ha mostrar y las filas corresponden a los genes marcadores de la microarray.

La tabla contiene un total de 7 columnas, de izquierda a derecha son:

- Gene Name: El nombre del gen marcador y breve descripción del gen. En el caso de los genes marcadores comunes, muestra el nombre del gen marcador de la microarray del usuario. En los genes marcadores no comunes, muestra el nombre del gen marcador de la base de datos.
- Matching-gds clúster distribution: La imagen de la distribución de clusters de condiciones muestrales de la microarray seleccionada.
- Dist: La separación en términos de expresión de los clusters para el gen marcador y la microarray del usuario (a mayor diferencia a la hora de expresa uno u otro clúster más marcador será el gen)
- user-gds clúster distribution: La imagen de la distribución de clusters de condiciones muestrales de la microarray del usuario.
- Correlation factor: El grado de correlación del gen marcador
- Gene marker: La relación del gen marcador con la expresión del gen.
- Gene: Información del gen de la base de datos Gene.

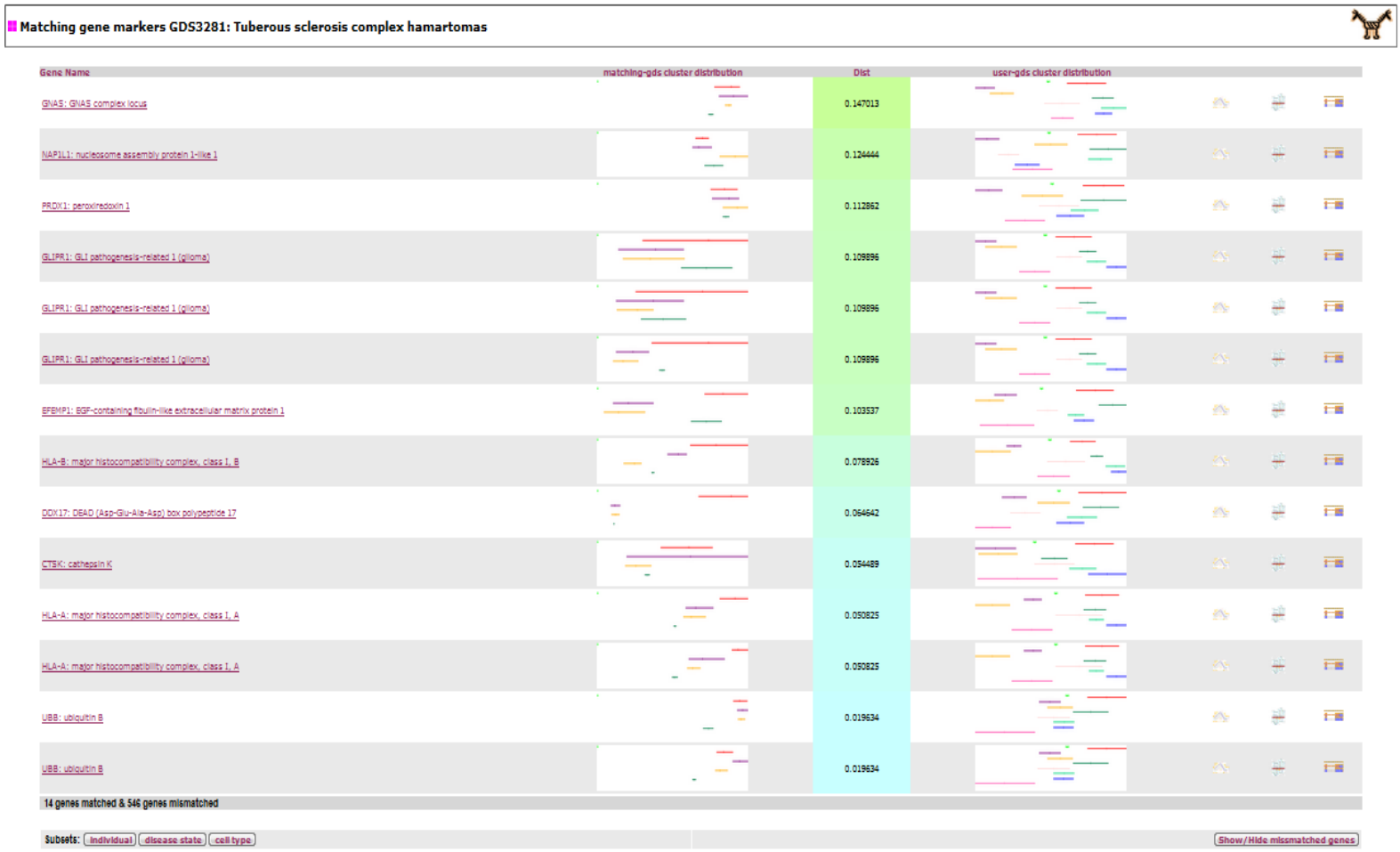


Figura 3.13: PHP correspondiente a la vista detalle. El listado muestra los genes marcadores de la microarray de la base de datos seleccionada en la vista general. La primera columna corresponde al nombre del gen marcador. La segunda columna es la distribución de clusters a lo largo del rango de expresión del gen marcador para la microarray de la base de datos (puede haber varias distribuciones). La tercera columna, corresponde a la distancia de separación en términos de expresión de los clusters en la microarray del usuario para el gen marcador (a mayor distancia entre la expresión de uno u otro clúster más marcador será el gen). La cuarta columna muestra la distribución de los clusters a lo largo del rango de expresión del gen marcador en la microarray del usuario (sólo aparece en los genes marcadores comunes entre ambas microarrays). Las tres últimas columnas corresponden al grado de correlación del gen marcador y el resto de genes de la microarray del usuario y información del gen de la base de datos Gene del NCBI. Notar que los clusters de la microarray del usuario y de la microarray de la base de datos no tienen que ser los mismos, pero si un mismo gen es marcador para un cluster de una y otra microarray, querrá decir que estos clusters mantendrán una equivalencia. En esta linea podrán extrapolarse los atributos del cluster biomédico de la microarray de la base de datos al cluster de la microarray del usuario.

La tabla se ordena inicialmente con la columna Dist, es la distancia de separación en términos de expresión de los clusters en la microarray del usuario y para el gen marcador y sólo es visible para aquellos genes marcadores comunes entre ambas bases de datos. El degradado de color de la columna Dist indica si las distancias son muy similares, cuanto más degradado más variabilidad de distancias. También se

permite al usuario ordenar por el nombre del gen, primera columna.

Las filas albergan información sobre el gen marcador. Las imágenes mostradas en las columnas matching-gds clúster distribution y user-gds clúster distribution son las generadas en la sección 3.4.1. La primera de ellas corresponde con la distribución de clústers del gen marcador de la base de datos de microarrays y la segunda corresponde con la distribución de clusters del gen marcador de la microarray del usuario. En ambas imágenes se ha incorporado la funcionalidad de aumentar su tamaño con un efecto suave, como sucedía en la vista general. Además de haber realizado el zebrado sobre las filas, para una mejor visualización de los genes marcadores.

Si el investigador necesitase más información sobre el gen marcador, el nombre del gen es un link[33] directo con el perfil de expresión correspondiente a ese gen marcador en la base de datos GEO Profiles del NCBI. Las tres últimas columnas corresponden a la correlación de los genes de la microarray con el gen marcador, la relación del gen marcador con la expresión del gen y información del gen.

En la parte inferior de la tabla de la figura 3.14 aparece el número de genes marcadores comunes y el número de genes marcadores no comunes. Al final de la figura 3.14 se puede observa la distribución de clusters de la microarray a modo de botones, permitiendo al usuario poder seleccionar la imagen de cada una de las distribuciones de clusters. A partir de una función JavaScript implementada, por cada selección de distribución se va cambiando la imagen correspondiente a la segunda columna (matching-gds clúster distribution). Un ejemplo de las diferentes distribuciones que puede tener una microarray se muestran en las figuras 3.15, 3.16 y 3.17.

Las distribuciones de clusters son determinadas por el experimento aplicado a la microarray y la posibilidad de agrupar condiciones muestrales con un comportamiento similar. Por eso, la cantidad de distribuciones varia por cada microarray.

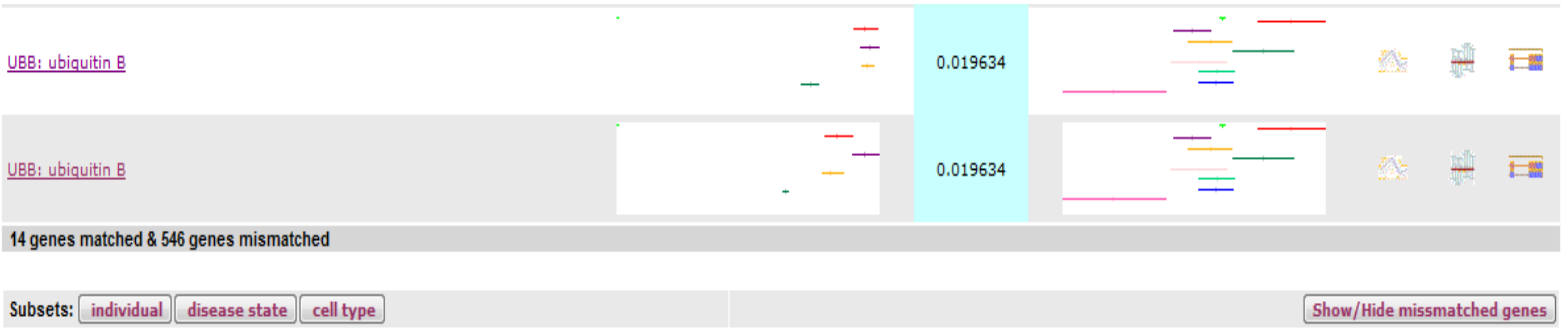


Figura 3.14: Imagen ampliada de la parte inferior de la figura 3.11. En la última línea de la tabla se muestran el número de genes comunes y no comunes al cruce de los genes marcadores de la microarray del usuario por la microarray listada.

33 Ejemplo con el gen UBC: http://www.ncbi.nlm.nih.gov/geoprofiles?term=53195319

También se da la posibilidad al usuario de ver los genes marcadores no comunes de la microarray con el botón Show/Hide missmatched genes situado en la parte inferior derecha de la figura 3.12. Al realizar la acción de mostrar los genes no comunes, se despliega la tabla con los genes marcadores restantes de la microarray. Para estos genes marcadores no se tiene ni la información Dist, ni la imagen de la distribución de clusters de la microarray del usuario. Debido a no formar parte de la microarray del usuario.

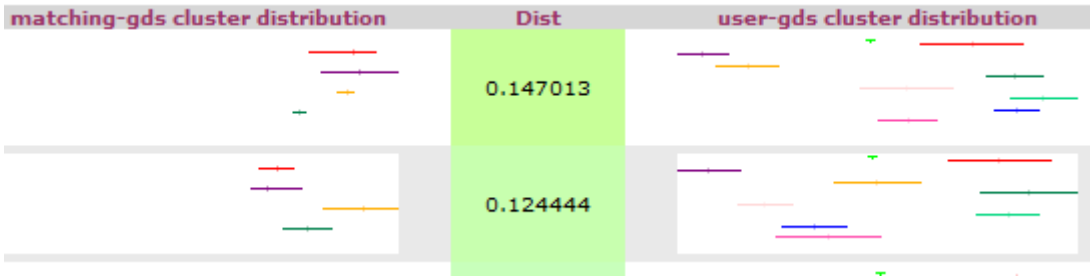


Figura 3.15: Imagen ampliada de las dos primeros genes (GNAS y NAP1L1) de la figura 3.11. La columna matching-gds cluster distribution muestra la distribución de clusters a lo largo del rango de expresión del gen marcador para la microarray de la base de datos (puede haber varias distribuciones), en este caso la distribución corresponde a individual. La columna user-gds cluster distribution muestra la distribución de los clusters a lo largo del rango de expresión del gen marcador en la microarray del usuario (sólo aparece en los genes marcadores comunes).

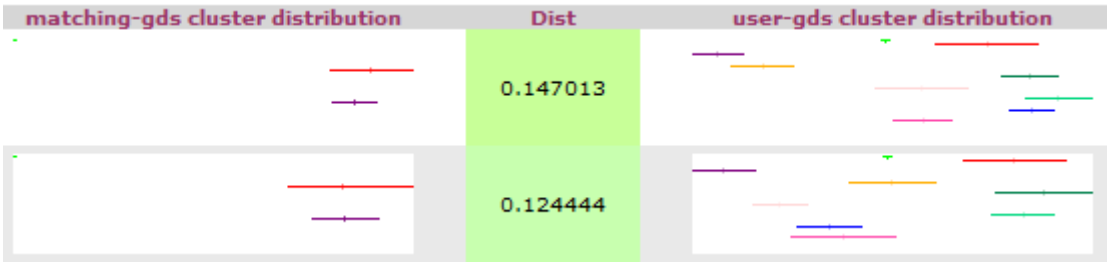


Figura 3.16:Imagen ampliada de las dos primeros genes (GNAS y NAP1L1) de la figura 3.11. La columna matching-gds cluster distribution muestra la distribución de clusters a lo largo del rango de expresión del gen marcador para la microarray de la base de datos (puede haber varias distribuciones), en este caso la distribución corresponde a disease state. La columna user-gds cluster distribution muestra la distribución de los clusters a lo largo del rango de expresión del gen marcador en la microarray del usuario (sólo aparece en los genes marcadores comunes) .Este caso el gen no es marcador no se puede extrapolar información del cluster.

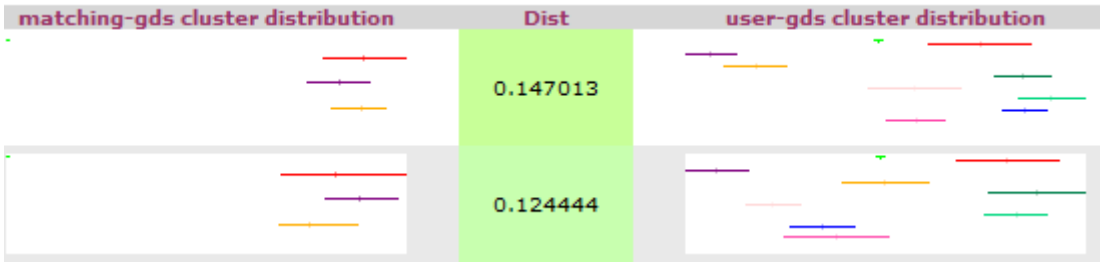


Figura 3.17:Imagen ampliada de las dos primeros genes (GNAS y NAP1L1) de la figura 3.11. La columna matching-gds cluster distribution muestra la distribución de clusters a lo largo del rango de expresión del gen marcador para la microarray de la base de datos (puede haber varias distribuciones), en este caso la distribución corresponde a cell type. La columna user-gds cluster distribution muestra la distribución de los clusters a lo largo del rango de expresión del gen marcador en la microarray del usuario (sólo aparece en los genes marcadores comunes).

Observando las figuras 3.15, 3.16 y 3.17 se pueden observar como dependiendo del tipo de agrupación la expresión del gen marcador para la microarray varia. Esto se puede observar en la columna matching-gds cluster distribution de las 3 figuras. En la figura 3.15 tiene una agrupación de individual, la figura 3.16 se agrupa según el disease state y la figura 3.17 se agrupa según el cell type. La columna user-gds cluster distibution no varia, ya que corresponde a la microarray del usuario y contiene una única distribución.

3.4.2.4 Borrado de archivos temporales

Finalizada la interfaz web pensé en mantener en buen estado el servidor. A lo largo del proceso de la aplicación web, se generan archivos temporales necesarios en la página web para mostrar la información requerida.

Los archivos temporales generados corresponden a la salida del programa de cruce online y a la generación de gifs de los genes marcadores de la microarray del usuario.

Debido a la basura generada por la interfaz web decidí realizar un script programado en el cron[34] para que cada 2 días limpiase el directorio que almacena los archivos temporales en el servidor. Una vez realizada la consulta, mostrada y trabajada por el usuario, cuando finaliza el estudio sobre la consulta los archivos temporales dejan de ser útiles y pueden ser eliminados. Así se mantiene en buen estado el servidor y no se almacena información innecesaria.

Inicialmente el proceso de limpieza se realizaba automáticamente después de cerrar la pestaña o el navegador donde se hubiese hecho la consulta. Al cerrar la pestaña del navegador se lanzaba otro php que

34 El cron es un administrador regular de procesos en segundo plano, de nominados demonios, que ejecuta procesos a intervalos regulares

realizaba el borrado de los archivos temporales, pero provocaba que se abriese otra pestaña vacía. Desestimé esta opción en consecuencia del programa en el cron, porque podría llegar a ser molesto para el usuario que cada vez que cerrase la consulta hecha se abriese otra pestaña que posteriormente tendría que cerrar.

3.5 Automatización mensual de la base de datos de genes marcadores de microarrays

La actualización de la base de datos de genes marcadores de microarrays es esencial para la aplicación, manteniendola en buen estado y aumentando la cantidad de microarrays de la base de datos de microarrays.

La idea de actualizar la base de datos es reutilizar en gran medida el preproceso iniciado al principio del trabajo. Es decir, ahorrarse en lo posible la generación de gifs para los genes marcadores.

Al almacenar los datos en ficheros XML, la primera parte del preproceso se realiza en su totalidad. Descargando y agrupando por microarrays los genes marcadores del NCBI. Corresponde al apartado 3.2 de fases, aprovechando los scripts necesarios. A partir de la figura 3.16 se puede observar el proceso de este punto (corresponde a la descarga de genes marcadores y agrupar la información) y el resto de la actualización.

A partir de ese momento, se tiene que comprobar los nuevos datos descargados con los antiguos, y comparar que genes marcadores son nuevos. Este proceso se corresponde con el proceso de comparar de la figura 3.18. Para el caso de que sean nuevos, se genera su gif correspondiente. Si la microarray por completa es nueva, se realiza por completo la generación de imágenes (implica la necesidad de descargar mediante el FTP, la microarray correspondiente). La generación de gifs se realiza directamente en el actual sistema de ficheros.

Cuando finalicé la etapa de generación de los nuevos datos a partir de los genes marcadores, se reemplaza la base de datos de microarrays actualizadas por la actual. Finalizando el proceso de actualización con la eliminación de la base de datos de microarrays antigua y archivos temporales creados para la actualización.

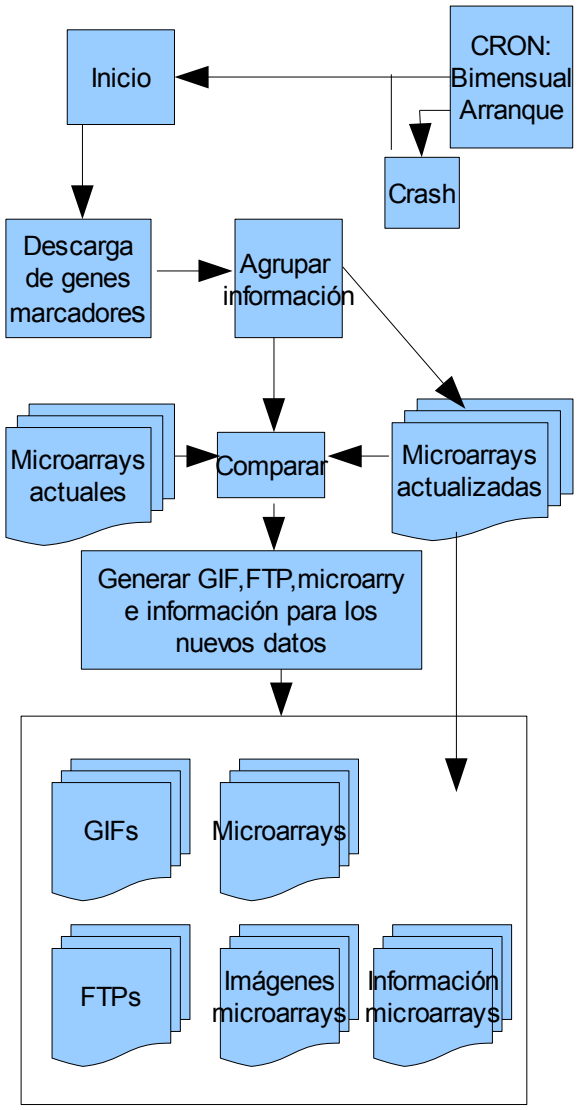


Figura 3.18: Esquema de la actualización de la base de datos local de genes marcadores. La actualización se inicia con el cron cada 2 meses. Primero se descargan los genes marcadores y se agrupan por la microarray a la que pertenecen, construyendo la base de datos de microarrays pero completamente actualizada. Seguidamente se comparan las microarrays actuales y las actualizadas. Por cada gen marcador nuevo de la base de datos de microarrays actualizada se realiza la generación de la imagen correspondiente y se guarda en el directorio correspondiente. Las imágenes generadas se almacenan directamente sobre el actual sistema de ficheros (el recuadro inferior de la figura).

La actualización se realiza bimensualmente mediante la utilización del cron del sistema operativo, que ejecuta el programa de actualización de la base de datos de microarrays. En el cron también se programa para el caso de apagado de la máquina en el proceso de actualización. Comprobando al inicio de la máquina, si la actualización finalizó correctamente. En caso negativo, se reinicia el proceso de actualización. La actualización tiene una duración aproximada de 8 horas.

4. Informe técnico

A continuación describo de manera más técnica por secciones los programas implementados y la estructura de directorios seguida para el desarrollo de la aplicación web. En los programas implementados realizo una descripción del funcionamiento, su ubicación, cuales son los parámetros de entrada y de salida y los módulos o librerías necesarias para su utilización.

4.1 Estructura de directorios

Entender la estructura de directorios es básico para poder situar correctamente los programas, la base de datos de genes marcadores de microarrays y las páginas web.

Los ficheros se dividen en dos directorios:

- Directorio /var/www/cgi-bin/gds/:
  - Base de datos de genes marcadores de microarrays almacenada en el directorio “GDSs/”.
  - Programa de cruce en tiempo real, llamado matching.
  - Imágenes de los genes marcadores almacenados en el directorio “GIFS/”. Por cada distribución de clusters de la microarray hay otro subdirectorio.
  - Programas y ejecutables, referentes a la actualización y programas online de la aplicación (filtro.pl y onlinegifs.pl).
  - Archivos temporales se generan en el directorio “online/”. Dentro de este directorio se almacena por el subdirectorio del id del usuario.
  - Programa de borrado de archivos temporales, llamado borrar.pl
  - Imágenes de las distribuciones de microarrays almacenadas en el directorio “images/”.

- Directorio /var/www/html/applic/gexp/microarray/:
  - Páginas web.

4.2 Construcción de la base de datos de genes marcadores de microarrays

La construcción de la base de datos local se realiza en 2 etapas:

- Descarga de genes marcadores a través de E-Utils.
- Agrupación de los genes marcadores por microarrays.

Cada etapa la he realizado con un script en Perl. Para poder utilizar ciertas funciones implementadas en Perl se necesita incorporar al script el correspondiente módulo que contiene dichas funciones. Los módulos básicos ya están instalados con la instalación de Perl en el sistema operativo. Algunos módulos es necesario instalarlos, una vía cómoda para ello es CPAN.

El modo de instalación es bastante sencillo. Primero hay que lanzar como root el siguiente comando para entrar la consola de CPAN:

perl -MCPAN -e shell

Una vez en la consola de CPAN, hay que instalar el módulo requerido. En mi caso, necesito un módulo para poder leer y tratar XML, es decir, el módulo XML::Simple[21]:

cpan> install XML::Simple

Una vez hecho esto, CPAN automáticamente compilará el módulo. Si fuera necesario la instalación de algún otro módulo, se repite el último paso con el módulo correspondiente. Para poder realizar consultas MySQL en Perl se necesitan los módulos DBI y DBD::mysql, que también son instalados en el sistema.

Para el caso concreto del módulo XML::Simple también es necesario instalar una librería en el sistema operativo del servidor, que corresponde a libxml con el comando:

sudo apt-get install libxml-simple-perl

4.2.1 Descarga de genes a través de E-Utils

En esta etapa el programa se encarga de realizar una consulta (construida adecuadamente a los requisitos), explicada en la sección de fases 3.2, con la herramienta E-Utils del NCBI. Concretamente a la base de datos GEO Profiles del NCBI. Con la respuesta de la primera consulta en forma de lista de identificadores de perfiles de expresión de GEO Profiles (UID), enlazo la información de la primera consulta con una segunda consulta para obtener además del UID, información sobre el perfil. Este proceso se realiza con un bucle que va descargando archivos XML. Cada archivo XML contiene 500 genes marcadores desordenados.

No requiere parámetros de entrada. El script crea un directorio "out/" que contiene cada XML descargado. También se crea un archivo txt para almacenar el número de ficheros XML generados para la posterior etapa.

4.2.2 Agrupación de los genes marcadores por microarrays

A partir de la información conseguida en la etapa anterior se agrupa cada gen marcador en su

correspondiente microarray o gds, ya que la información descargada no esta agrupada por microarray. Además de seleccionar la información necesaria de cada perfil de expresión, se obtiene el ID de la base de datos de Gene (GeneID). El GeneID se obtiene por la base de datos local geneinfo del http://revolutionresearch.uab.es, si no esta disponible se consigue a partir de una consulta a E-Utils.

No requiere parámetros de entrada, pero utiliza como entrada el archivo txt generado anteriormente con el número de ficheros XML y los XML generados en el directorio "out/". La salida del script se crea en el directorio "GDSs/" donde se almacenan todos los genes marcadores agrupados en su correspondiente microarray y un archivo txt que contiene todos los nombres de las microarrays contenidas llamado listagds.out. Cada microarray esta formado por dos ficheros XML, uno contiene los genes marcadores y la información de cada perfil de expresión y el otro fichero contiene información referente a la propia microarray.

Estructura del fichero XML que contiene los genes marcadores de una microarray:

```
<gds>
	<gen>
		<geneid></geneid>
		<uid></uid>
		<geneName></geneName>
		<alias></alias>
		<geneDesc></geneDesc>
		<idref></idref>
	</gen>
	<gen>
		<geneid></geneid>
		<uid></uid>
		<geneName></geneName>
		<alias></alias>
		<geneDesc></geneDesc>
		<idref></idref>
	</gen>

	....

</gds>
```

Estructura del fichero XML que contiene la información referente a la microarray:

```
<gds>
	<taxid></taxid>
	<titlee></titlee>
	<summary></summary>
	<taxon></taxon>
	<count></count>
</gds>
```

4.3 Cruce en tiempo real

El programa compara los genes marcadores de la microarray del usuario con la base de datos de genes marcadores de microarrays. Esta implementado en lenguaje C (compilador gcc de Linux) y para poder analizar los ficheros XML es necesario la librería libxml2.

El funcionamiento del programa consiste en recorrer cada microarray de la base de datos y por cada gen marcador que contiene compararlo con los genes marcadores de la microarray del usuario.

El programa se llama cruce.c. Para compilarlo y ejecutarlo se hace de la siguiente manera:

1. gcc -c cruce.c `xml2-config --cflags`
2. gcc -o matching `xml2-config --libs` cruce.o
3. ./matching ARG1 ARG2 ARG3 ARG4 ARG5 ARG6

Para poder ejecutar el ejecutable generado en el paso 2 y lanzado en el paso 3, es necesario dar permisos de ejecución al ejecutable matching con chmod[35].

El programa tiene varios argumentos de entrada, en orden son:

- Ruta del archivo de entrada que contiene los genes marcadores de la microarray del usuario.
- Ruta del archivo de entrada que contiene la lista de nombres de microarrays de la base de datos.
- Ruta del archivo de salida que contendrá una lista de microarrays resultantes con el porcentaje de acierto respecto a los genes marcadores de la microarray del usuario y el total de genes marcadores de la microarray de la base de datos.
- Ruta del archivo de salida que contendrá los genes marcadores comunes del cruce para una microarray. Se repetirá esta ruta para cada microarray de la lista generada en el punto anterior.
- Número entero que designa si se utiliza la búsqueda por alias o no. Si se realiza es un 1, sino un 0.
- Número entero del número total de microarrays existentes en la lista de nombres de microarrays de la base de datos (segundo argumento) en el caso de utilizar una búsqueda filtrada por nombre (no contiene todas las microarrays de la base de datos). En caso contrario es un 0.

La salida del programa corresponde con el argumento 3 y 4. En la ruta del argumento 3 se genera un archivo txt con la lista de microarrays compatibles y el porcentaje de acierto de los genes marcadores de entrada (argumento 1) sobre la base de datos de genes marcadores de microarrays. Para la ruta del argumento 4, se generarán tantos archivos txt como microarrays existentes en archivo del argumento 3, cada archivo se diferencia por el nombre de la microarray y contienen los genes marcadores comunes entre

35 Comando de linux que permite cambiar permisos de acceso de un archivo o directorio. http://es.wikipedia.org/wiki/Chmod

los genes marcadores de entrada y los genes marcadores de la microarray compatible.

4.4 Generación de imágenes que representan la distribución de clusters

La generación de imágenes consta de 2 programas, un programa que genera todos las imágenes de los genes marcadores de la base de datos de microarrays y otro programa para generar de manera online las imágenes correspondientes a los genes marcadores de la microarray de interés.

Para poder generar las imágenes correspondientes a los genes marcadores de la base de datos de microarrays es necesario obtener los valores de expresión de cada gen marcador. Los valores de expresión son almacenados en las microarrays del FTP del NCBI.

Entonces es necesario obtener las microarrays del FTP con el script geoftp.pl. Este script lee el archivo listagds.out que contiene los nombre de las microarrays para descargar y mediante el módulo de Perl Net::FTP me permite obtener las microarrays en el directorio "FTP/".

Descargadas las microarrays ya se puede comenzar con la generación de imágenes. El script para generar por completo las imágenes de los genes marcadores de la base de datos de microarrays se llama generacion.pl, está implementado en Perl. No necesita parámetros de entrada, porque ya se encarga el script de obtener los ficheros que necesita para su funcionamiento. La salida del script es un directorio llamado "GIFS/" donde se almacenan los gifs de los genes marcadores. En el directorio "GIFS/" se generan subdirectorios para cada distribución de clusters de una microarray. Por ejemplo si una microarray tiene 3 distribuciones de clusters, se generan 3 directorios dentro del directorio "GIFS/" correspondientes a cada una de las distribuciones de clusters de la microarray.

El script comienza con obtener los nombres de las microarrays de la base de datos de las microarrays a partir del archivo creado en "GDSs/" llamado listagds.out. A partir de los nombres de las microarrays se leen las microarrays descargadas en el directorio "FTP/" para obtener los valores de expresión de los genes marcadores.

Por cada microarray descargada se parsea con el objetivo de obtener los valores de expresión de los genes marcadores de la base de datos y la o las distribuciones de clusters de la microarray. Para obtener los valores de expresión de los genes marcadores se compara cada gen de la microarray descargada a partir de los identificadores del gen en la microarray. El identificador de la microarray (idref) se encuentra en el XML que contiene los genes marcadores de la microarray en la base de datos de genes marcadores de microarrays. Para obtener las distribuciones de clusters se utilizan unos arrays para determinar cada distribución y generar los archivos colors para cada distribución. Un archivo colors esta formado por 2 columnas, la primera determina la condición muestral y la segunda determina el clúster al que pertenece.

Un ejemplo de archivo colors seria:

1 1
2 1
3 1
4 2
5 2
6 2

Donde el valor 1 representa un clúster que agrupa las condiciones muestrales 1,2,3 y el valor 2 representa otro clúster que agrupa las condiciones muestrales 4,5,6.

Una vez cumplidos los objetivos para parsear la microarray descargada se realiza una llamada al programa gsTdeStud por cada distribución de clusters de la microarray. El programa gsTdeStud esta implementado en C y es una reutilización de otro programa existente. La entrada del programa son los valores de expresión de los genes marcadores y un archivo colors. A partir de los datos de entrada el programa calcula los valores mínimo, media y máximo de cada clúster de todos los genes introducidos a partir de calcular una distribución t de student y aplicar un intervalo de confianza.

Con los valores mínimo,media y máximo de todos los clústers de un gen marcador, se estructuran en forma de lista para introducirlo en la herramienta de generación de gif. A parte de la lista introduzco el tamaño de la imagen (200x50), el nombre del gif, el punto medio(0) y una lista con los centros de los clusters. Los centros de los clusters corresponden a la cantidad de condiciones muestrales existentes en un clúster por el total de condiciones muestrales de la microarray más un factor de corrección de 0,2 para resaltar más los centros. Este proceso se realiza por todos los genes marcadores existentes en la base de datos de genes marcadores de microarrays y por todas las distribuciones de clusters de las microarrays.

4.5 Interfície web

La sección de interfície web contiene las páginas webs que muestran los resultados obtenidos en el programa de cruce y los scripts que permiten realizar los cálculos online.

Páginas webs en orden de proceso:

- appl_classegsresults2.phtml: Página web que muestra los genes marcadores de la microarray del usuario según los criterios del usuario en sobreexpresar o infoexpresar los clusters. Pertenece a los cálculos previos y enlaza ambas aplicaciones web.

- intermedio3.php: Comunica las variables entre ambas aplicaciones web y muestra un gif para amenizar la espera de los cálculos.
- intermedio2.php: Realiza las llamadas de los programas de cruce entre genes marcadores y generador de gifs para los genes marcadores de la microarray del usuario obtenidos en appl_classegsresults2.phtml.
- resultados.php: Muestra la lista de microarrays compatibles en forma de tabla.
- detalle.php: Muestra los genes marcadores comunes (o no) de la microarray seleccionada y permite ver a modo de gif la distribución de clusters del gen marcador.

Para poder acceder a los ficheros XML de la base de datos de microarrays se requiere de xml2array.php, que permite leer y tratar los XML.

Scripts (en Perl) lanzados desde intermedio2.php:

- filtro.pl: Es el programa que comunica la parte web con la base de datos de microarrays. Realiza la llamada al programa de cruce teniendo en cuenta los parámetros comentados en la sección 4.3 y lanza la generación de gifs mediante el programa onlinegifs.pl. Tiene 5 parámetros en su llamada con la siguiente forma:

  perl fil2.pl -a ARG1 -li ARG2 -fe ARG3 -mt ARG4 -tx ARG5

  ARG1: identifa con un 1 si esta activo el alias o 0 en caso contrario.
  ARG2: id del usuario en el servidor.
  ARG3: la fecha del servidor.
  ARG4: id de la microarray del usuario.
  ARG5: texto introducido para filtrar las microarrays.
- onlinegifs.pl: Es el programa que realiza la generación de gifs de los genes marcadores de la microarray del usuario. Tiene como parámetros de entrada el id de la microarray del usuario, el id del usuario y la fecha. A partir de los parámetros de entrada el script obtiene el archivo colors y los valores de expresión de los genes marcadores para poder generar los gifs.

4.6 Actualización

La actualización de la base de datos de microarrays se realiza cada 2 meses mediante el cron. Permitiendo comprobar si hay nuevos genes marcadores en la base de datos GEO Profiles del NCBI.

El programa esta programado en el cron con:

0 0 10 */2 * bsh <ruta del archivo>

quiere decir, que el dia 10 a las 00.00 cada dos meses se ejecuta el archivo bsh especificado en la ruta. En mi caso la ruta corresponde a: /var/www/cgi-bin/gds/act.bsh, donde se encuentra el archivo que

iniciará la actualización. Este archivo lanza un script en Perl que contiene las llamadas para realizar la actualización, llamado principal.pl.

El script principal.pl contiene las llamadas en orden de:

- xsrc.pl: Descarga de genes marcadores y su información para construir la base de datos local.
- xunion2.pl: Agrupa los genes marcadores en su correspondiente microarray. El directorio donde se generan es "GDSsACT/".
- comparar1.pl: Realiza una comparación de la base de datos de microarrays descargada en la actualización con la base de datos de microarrays actual. Consiste en comparar primero si la microarray descargada existe en base de datos de microarrays actual, si existe se comparan cada uno de los genes marcadores pertenecientes. Sino existe la microarray, se realiza la generación de gifs completa como en la sección 4.4.
- comparar2.pl: Finaliza la comparación con la generación de gifs de los genes marcadores nuevos dentro de una microarray existente en la base de datos de microarrays actual.
- perl rmv.pl: Completa la actualización con el renombrado de las carpetas y eliminación de la base de datos de microarrays antigua y de archivos temporales como los generados por el script xsrc.pl.

Comentar que los scripts xsrc.pl y xunion2.pl son muy semejantes a los descritos en la sección 4.2, pero con pequeñas modificaciones para realizar más comprobaciones. Los scripts compara1.pl y compara2.pl realizan la generación de gifs para los genes marcadores utilizando el código y el mismo sistema de directorios de la sección 4.4, adaptado para la actualización ya que no hace falta realizar la generación de gifs para todos los genes marcadores de las microarrays. Porque si un gen marcador descargado en la actualización es el mismo que en la base de datos de microarrays actual, no hace falta volver a generar su gif correspondiente.

También hay un script llamado arranque.bsh que está programado en el cron para el caso de interrupción de la actualización. Simplemente comprueba al iniciarse el servidor si la actualización se ha realizado satisfactoriamente. En caso contrario, se iniciará de nuevo el proceso de actualización.

4.7 Limpieza del servidor

A causa de los archivos temporales que se generan por el cruce entre genes marcadores y generación de gifs online. Se necesita realizar una limpieza del servidor periódicamente. Porque sino se acumularía basura. Por ello se utiliza el script borrar.pl programado en el cron con la sentencia:

0 0 */2 * * bsh <ruta del archivo>

quiere decir, que cada dos días a las 00.00 se ejecuta el archivo bsh especificado en la ruta. En mi caso la ruta corresponde a: /var/www/cgi-bin/gds/limp.bsh, donde se encuentra el archivo que iniciará la limpieza. Este script llamará a borrar.pl para limpiar el contenido de los directorios de usuario existentes en el directorio "online/", donde se generan los archivos temporales.

5. Conclusiones

Los objetivos marcados han sido alcanzados con creces, tanto los principales como los secundarios. Como resultado ahora se ofrece una nueva herramienta muy útil para los investigadores en el campo de la biología molecular.

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. De esta manera permitimos cruzar los genes marcadores para los clusters de origen estadístico (o no) de la microarray que analiza el usuario con los genes marcadores para clusters de origen no estadístico de las microarrays que hay almacenadas en la base de datos local. El resultado que se consigue es facilitar al usuario extrapolar el conocimiento proporcionado por unas microarrays (un total de 2510 gds a las que mensualmente se añaden otras nuevas) a otras diferentes, concretamente la microarray de estudio del investigador.

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. Está actualmente programada para actualizarse cada 2 meses con las nuevas microarrays subidas al NCBI por grupos de investigación internacionales. Las bases de datos locales también se actualizarán con la identificación de los nuevos genes anteriormente desconocidos, lo que permitirá ofrecer al usuario la información más actual con todo lo referente a los genes marcadores obtenidos por el aplicativo y refinar las búsquedas de genes comunes entre microarrays.

El cruce “on-line” 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. Gracias a la optimización de los algoritmos de cruce, el aplicativo web es fluido en la navegación, con tiempos de espera mínimos para el usuario.

La interfaz web realizada proporciona un aplicativo altamente usable, entendible y con una alta operativilidad, lo que facilitará su estudio a los investigadores.

Por lo tanto, todos los objetivos se han cumplido incluso proporcionando algunas funcionalidades extras, de forma que el usuario podrá recibir más información útil para su estudio y de una forma más organizada.

Finalmente comentar sobre lo gratificante que ha sido realizar este último viaje en la universidad a modo de trabajo final. Dando a luz una aplicación web que en un futuro esperemos que pueda ayudar a descubrir la causa y remedio de algunas patologías.

5.1 Trabajos futuros

El proyecto desarrollado podría continuar las siguientes líneas de ampliación, con el objetivo de enriquecer con más genes marcadores la base de datos y permitir al investigador otras vías de acceso a la aplicación web:

- Aumentar el número de genes marcadores, es decir, realizar otras consultas E-Utils a GEO Profiles con diferentes parámetros de búsqueda para obtener más genes marcadores. En este momento, se realiza una consulta con el término "value subset effect", usando otros parámetros se obtendrían algunos genes marcadores distintos. Esto conseguirá aumentar el numero de genes marcadores de la base de datos de genes marcadores de microarrays, lo que facilitaría encontrar mayores coincidencias en las búsquedas.
- Permitir acceder a mi aplicación desde otras aplicaciones, a parte de la aplicación NCR-PCOPGene. Es decir, realizar otros cálculos para la microarray actual que obtuviese diferentes genes marcadores como salida. Pasar entonces esos genes marcadores a mi aplicación y así conocer que están marcando esos genes marcadores.
- Permitir introducir manualmente aquellos genes marcadores que desee estudiar el investigador., Pasar entonces esos genes marcadores a mi aplicación y así conocer que están marcando esos genes marcadores.
- Realizar un cruce entre los genes marcadores de todas las microarrays de la base de datos con el objetivo de establecer equivalencias entre microarrays o con clusters de condiciones muestrales equivalentes (dado que comparten sus genes marcadores).
- En la página web resultante de la aplicación NCR-PCOPGene, donde se listan los genes marcadores para la microarray del usuario mostrar las imágenes con la distribución de clusters de condiciones muestrales para el rango de expresión del gen marcador.

5.2 Presupuesto

El presupuesto del proyecto esta desglosado en 3 partes que son: mano de obra, software requerido y servidor. A continuación se presenta un cuadro con los costes del proyecto:

Cantidad	Descripción	Precio unitario	Importe
420	Mano de obra	8	3360
5	Software(*):		0
1	Servidor web Apache	0	0
1	Entorno PHP	0	0
1	Gestor de bases de datos MySQL	0	0
1	Intérprete Perl	0	0
1	Compilador GCC	0	0
1	Servidor:	8000	6780
1	Servidor Supermicro SYS-6016T-NTRF		
2	INTEL DP WETMERE 6C X5650 2.66G 12M 6.4GT		
8	DDR3 1333 8GB ECC REGISTERED		
4	HUA722020ALA330 HITACHI 2TB 7200 SATAII 32MB		
1	ADAPTEC RAID 5405 SATA/SAS KIT PCI-E 4 PORT INT.		
	3 años de garantía - RAID 10 con los 4 discos pero solo de 300Gb (150Gb de cada disco) para S.O. - RAID 5 con los 4 discos con el resto del espacio (5.2TB netros) para datos		
		TOTAL (sin IVA)	**10140€**
		TOTAL (SIN SERVIDOR) (sin IVA)	**3360€**

(*): El software utilizado es código libre (open source), por lo que no tiene coste alguno.

Matizar sobre el presupuesto, que el servidor descrito corresponde al servidor que alberga el domino http://revolutionresearch.uab.es/, donde se integra la aplicación web. Este servidor se utiliza para albergar todas las aplicaciones desarrolladas en la área de bioinformática del IBB. Es un servidor potente que cumple exageradamente las necesidades de la aplicación desarrollada en el proyecto.

He dividido los 2 precios, porque no se ha comprado exclusivamente el servidor para mi aplicación, sino para toda la área de bioinformática. El coste total sin servidor representa el presupuesto del desarrollo de la aplicación en el proyecto.

6. Bibliografía:

[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] La página Web oficial del National Center for Biotechnology Information (NCBI) que ofrece de manera pública todas sus bases de datos. http://www.ncbi.nlm.nih.gov/

[4] 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., 9;10:138 Microarray interactive gene networks

[5] 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

[6] 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

[7] Delicado, P.(2001) Another look at principal curves and surfaces. Journal of Multivariate Analysis, 77, 84-116 . PCOP theoretical definition

[8] 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

[9] 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

[10] Barret T (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res, 35 , D760-5. Pubmed

[11] Barrett T, Edgar R. (2006) Gene Expression Omnibus (GEO): Microarray data storage, submission, retrieval, and analysis. Methods in Enzymology, 411:352-369. Pubmed

[12] Barrett T, Edgar R. (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO). Methods Mol Biol, 338: 175-90. Pubmed

[13] Barret T (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res, 33: D562-6. Pubmed

[14] Edgar R. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 30: 207-10. Pubmed

[15] NCBI (2010) Entrez Programming Utilities http://www.ncbi.nlm.nih.gov/books/NBK25501/

[16] http://revolutionresearch.uab.es/downloads/webmicroarray/cluster1_3.html#inici : Microarray clustering

[17] http://www.solociencia.com/ : Portal de la Ciencia y la Tecnología en Español.

[18] http://es.wikipedia.org : Enciclopedia de contenido libre.

[19] http://www.perl.org : Home of the Perl programming language.

[20] http://perlenespanol.com/ : Una de las comunidades más grandes de Perl en habla hispana.

[21] http://www.cpan.org/ : Open source Perl modules ready to download and use

[22] http://search.cpan.org/~grantm/XML-Simple-2.18/lib/XML/Simple.pm : Easy API to maintain XML

[23] http://xmlsoft.org/ : The XML C parser and toolkit of Gnome.

[24] http://php.net : Sitio oficial de PHP con gran cantidad de recursos en ingles, noticias, descargas, documentación, calendario de eventos relacionados.

Firmado: Marc Muñoz Escudero

Bellaterra, 22 de Junio de 2011

Resumen

En la presente memoria se detalla con precisión las diversas fases del trabajo para construir una aplicación web en el servidor http://revolutionresearch.uab.es que permite enriquecer los clusters de la microarray del usuario con información biomédica de una base de datos remota.

Los clusters de origen estadístico (o no) de la microarray del usuario se enriquecen a partir de cruzar sus genes marcadores con la base de datos de genes marcadores de microarrays (base de datos remota) con clusters basados en información biomédica. La base de datos de genes marcadores de microarrays ha sido obtenida a partir de la base de datos de GEO Profiles del NCBI.

Resum

En la present memòria es detalla amb precisió les diverses fases del treball per tal de construir una aplicació web en el servidor http://revolutionresearch.uab.es que permet enriquir els clusters de la microarray de l'usuari amb informació biomèdica de una base de dades remota.

Els clusters d'origen estadístic (o no) de la microarray de l'usuari s'enriqueixen a partir de creuar els seus gens marcadors amb la base de dades de gens marcadors de microarrays (base de dades remota) amb clusters basats en informació biomèdica. La base de dades de gens marcadors de microarrays ha sigut obtinguda a partir de la base de dades de GEO Profiles del NCBI.

Summary

This report details various stages of the work in order to build a web application on this server http://revolutionresearch.uab.es that allows enrich clusters in the microarray's user with biomedical information from a remote database.

Clusters of statiscial origin (or not) of the microarray's user are enriched from the crossing of its genes markers with the database of genes markers of microarray (remote database) with clusters based on biomedical information. The database of genes markers of microarray has been obtained from the database GEO Profiles in NCBI.