Universidad Nacional de Córdoba

Facultad de Ciencias Exactas, Físicas y Naturales

Tesis Doctoral

Minería de Datos en Análisis Ontológico-Funcionales

Autor: Bioing. Cristóbal Fresno Rodríguez

Director: Dr. Elmer Andrés Fernández

Marzo de 2014

Minería de Datos en Análisis Ontológico-Funcionales

por

Bioing. Cristóbal Fresno Rodríguez

Dr. Elmer Andrés Fernández
Director

Comisión Asesora

Dr. Elmer Andrés Fernández
FCEFYN-UNC / FI-UCC

Dra. Cristina Noemí Gardenal
FCEFYN-UNC

Dra. Andrea Sabina Llera
Fundación Instituo Leloir - CONICET

Esta Tesis fue enviada a la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de Córdoba para cumplimentar los requerimientos de obtención del grado académico de Doctor en Ciencias de la Ingeniería.

Córdoba, Argentina
Marzo de 2014

Dedicado a Carolina,
mi compañera incondicional,
y a mi familia

Agradecimientos

Este trabajo no habría sido posible sin el apoyo y el estímulo del *Dr. Elmer Andrés Fernández* quien dio un salto al abismo, cuando me propuso como becario doctoral, con sólo haber estado en contacto pocos momentos durante el dictado de curso de postgrado en la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos. También me gustaría agradecerle a la *Dra. Andrea Llera* por permitirme formar parte de su proyecto al igual que por su apoyo, consejos y paciencia, cada vez que necesité una explicación biológica cuando no compartimos el diccionario.

A todos los compañeros del Grupo de Minería de Datos en Bio-ciencias de la Facultad de Ingeniería de la Universidad Católica de Córdoba, en especial a *Diego Arab*, *Germán González*, *Anibal Olivera* y *Gabriela Merino*; al igual que a los integrantes del Laboratorio de Biometría de la Facultad de Agronomía de la Universidad Nacional de Córdoba, en especial a *Mariano Córdoba*, *Andrea Peña*, *Mónica Picardi*, *Mónica Balzarini* y *Julio Di Rienzo*; así también como a los miembros del Laboratorio de Terapia Molecular y Celular de la Fundación Instituto Leloir, en especial a *Lorena Benedetti*, *Romina Girotti*, *Edgardo Salvatierra* y *Osvaldo Podhajcer*; y a los miembros del Laboratorio de Investigación y Desarrollo en Informática Aplicada de la Universidad Nacional de Villa María, en especial a *Laura Prato*, por haberme soportado durante estos años y con los cuales he compartido muy buenos momentos.

A las diferentes fuentes de financiamiento entre ellas a la Universidad Católica de Córdoba, la Agencia Nacional de Promoción Científica y Tecnológica, a través del Fondo para la Investigación Científica y Tecnológica, y al Consejo Nacional de Investigaciones Científicas y Técnicas.

No puedo terminar sin agradecer a *Carolina Kestler*, por su apoyo incondicional y a mi *familia* (biológica y política), a quienes dedico esta tesis.

Resumen

Palabras claves: *Bioingeniería/Bioinformática - Inteligencia Artificial - Tecnologías de alto rendimiento - Reconocimiento de patrones - Integración de información.*

El análisis ontológico-funcional determina las funciones o procesos biológicos modificados en un experimento genómico/proteómico de alto rendimiento. Si bien existen herramientas para su abordaje, la exploración del experimento implica consultar diversas bases de datos y utilizar programas que no siempre son compatibles entre sí, o sólo se encuentran disponibles vía web. Esta situación conlleva a una serie de problemas como lo son la carencia metodológica del análisis, falta de validación de los resultados, un análisis disgregado al utilizar herramientas con distinto formato en la presentación de sus resultados y dificultad en la integración de los resultados parciales o de distintos experimentos en forma simultánea. Éstos generan una serie de desafíos metodológicos, computacionales y analíticos desde el procesamiento, integración, visualización y validación de los resultados. En este contexto, la Minería de Datos brinda un sustento adecuado, proporcionando no sólo un flujo de trabajo ordenado, sino también estrategias y conceptos de Inteligencia Artificial para su abordaje.

Esta tesis propone diferentes metodologías que permiten un análisis más estructurado y completo de los datos obtenidos en uno o más experimentos, facilitando y enriqueciendo así el análisis ontológico-funcional. Se desarrolló un análisis de calidad de datos que permite detectar tendencias y evaluar su impacto, al igual que herramientas programáticas (automáticas) para integrar y validar computacionalmente los procesos funcionales alterados por el experimento. La exploración de los resultados puede ahora realizarse de forma centralizada a través de una interfaz visual amigable, que facilita la interpretación, búsqueda y extracción de información. Las metodologías propuestas han sido aplicadas a diversos experimentos, demostrando su utilidad en la identificación de patrones de información funcional de interés.

Abstract

Keywords: *Bioengineering/Bioinformatics - Artificial Intelligence - High-throughput Technologies - Pattern Recognition - Information Integration.*

Functional-ontology analysis determines the functions or biological processes modified in a genomic/proteomic high throughput experiment. Several tools exist to address it but, experiment exploration involves querying many databases and programs that are not always compatible with each other or only are web available. This situation leads to numerous problems such as lack of methodological analysis, results validation, a disaggregated analysis due to tool dependent results format and difficulty in the partial or separate experiments results integration. These problems generate a series of methodological, computational and analytical challenges from processing, integration, visualization and results validation. In this context, the Data Mining field provides an adequate livelihood, not just an ordered workflow, but also strategies and concepts from Artificial Intelligence. This thesis proposes different methodologies that allow a structured and complete data analysis from one or more experiments, thus, facilitating and enriching the functional-ontology analysis. It has been developed a data quality control methodology to detect trends and assess their impact, as well as, programmatic (automatic) pipelines to integrate and computational validate functional processes modified by the experiment. Results exploration can now be centrally performed through a friendly visual interface, which facilitates the interpretation, pattern search and information extraction. The proposed methodologies have been applied to several experiments, evidencing its usefulness on functional pattern detection.

Resumo

Palavras-chave: *Bioengenharia/Bioinformática - Inteligência Artificial - Tecnologias de alto desempenho - Reconhecimento de padrões - Integração das informações.*

A análise ontológico-funcional determina as funções ou processos biológicos modificados em uma pesquisa genomica / proteômica de alto desempenho. Ao mesmo tempo em que não há ferramentas para sua abordagem, a exploração da pesquisa envolve consultar vários bancos de dados e utilizar programas que nem sempre são compatíveis uns com os outros, ou só estão disponíveis através da web. Esta situação leva a uma série de problemas como a falta de caminho metodológico da análise, a falta de validação dos resultados, a análise disgregada quando usar ferramentas com formato diferente na apresentação dos seus resultados e a dificuldade no processo de integração dos resultados parciais ou várias pesquisas ao mesmo tempo. Estes geram uma série de desafios metodológicos, recursos computacionais e analíticos da transformação, a integração, a visualização e a validação dos resultados. Neste contexto, a "Minería de Dados" fornece um adequado sustento, não só por proporcionar um bom fluxo de trabalho, mas também as estratégias e conceitos de Inteligência Artificial para a sua abordagem.

Esta tese propõe diferentes metodologias que permitam uma análise mais estruturada e completa dos dados obtidos em uma ou mais pesquisas, facilitando e enriquecendo a análise ontológica-funcional. Desenvou-se uma análise da qualidade dos dados que lhe permite detectar tendências e avaliar o seu impacto, bem como ferramentas programáticas (automáticas) para integrar e avalidar processos funcionais computacionalmente alterados pela pesquisa. A exploração dos resultados já pode ser realizada de forma centralizada através de uma interface visual amigável, o que facilita a interpretação, a busca e extração de informações. As metodologias propostas têm sido aplicadas a diferentes pesquisas, provando a sua utilidade na identificação de padrões de informações funcionais de interesse.

Abreviaturas

2D-DIGE : 2D - Difference In Gel Electrophoresis.

ANOVA-PCA o APCA : ANOVA-Principal Component Analysis.

ANOVA-SCA o ASCA : ANOVA-Simlutaneous Component Analysis.

ANOVA : ANalysis Of VAriance.

AP o MO : Aceite de Pescado o Menhaden Oil.

API : Application Programming Interface.

AV o VO : Aceite Vegetal o Vegetal Oil.

CC : Componentes Celulares o Cellular Componente.

CD : Colina Deficiencia.

CORBA : Common Object Request Broker Architecture.

CS : Colina Suplementada.

DABG : Detection Above Background.

DAVID : Database for Annotation, Visualization and Integrated Discovery.

DIBD : Descubrimiento de Información en Bases de Datos.

DM : Data Mining.

DWS : DAVID Web Service.

EASE : Expression Analysis Systematic Explorer.

EC : Enzyme Commission number.

FDR : False Discovery Rate.

FM o MF : Función Molecular o Molecular Function.

FSH : Folicule Stimulant Hormone.

FSHrh : Folicule Stimulant Hormone-recombinante humana.

FSHrh-AC : FSHrh-ÁCida.

FSHrh-BA : FSHrh-BÁsica.

FSHrh-DR : FSHrh-Débilmente Retenidos.

FSHrh-FR : FSHrh-Fuertemente Retenidos.

FSHrh-NR : FSHrh-No Retenidos.

GDA : Gráfo Dirigido Acíclico

GEO : Gene Expression Omnibus.

GI : Protein_GI_Accession number.

GO : Gene Ontology.

GSEA : es el acrónimo de Gene Set Enrichment Analysis.

ID : Identificador.

IRA : Insuficiencia Renal Aguda.

KDD : Knowledge Discovery in Data bases.

KEGG : Kyoto Encyclopedia of Genes and Genomes.

LMDME : Linear Model Decomposition for designed Multivariate Experiments.

LR-I : LR del genoma de la especie bajo estudio.

LR-II : LR de genes presentes en el chip para experimentos de microarreglos.

LR-III : LR especificada a criterio del usuario.

LR : Lista de Referencia.

MAD : Meadian Absolute Deviation.

MD : Minería de Datos.

MEA : es el acrónimo de Modular Enrichment Analysis.

MRCM : Multi-Reference Contrast Method.

NA : Not Available.

NCBI : National Center for Biotechnology Information.

NGS : Next Generation Sequencing.

PB o BP : Procesos Biológicos o Biological Process.

PCA : Principal Component Analysis.

PLSR : Partial Least Squares Regression.

R-T PCR : Real-Time Polymerase Chain Reaction.

RMA : Robust Multi-array Average.

RMI : Remote Method Invocation.

SEA : Singular/Set Enrichment Analysis.

SOAP : Simple Object Access Protocol.

SVG : Scalable Vector Graphics.

TCP : Transmission Control Protocol.

URL : Uniform Resource Locator.

XML : eXtensible Markup Language.

Prefacio

El análisis ontológico-funcional es actualmente uno de los pasos cruciales en el procesamiento de experimentos proteómicos/genómicos de alto rendimiento (del inglés *high throughput*). Usualmente se lleva a cabo para relacionar una lista de genes con conceptos/categorías/términos de relevancia biológica, a los efectos de determinar las funciones y/o vías metabólicas modificadas (*enriquecidas*) por el experimento.

Esta tarea se lleva a cabo consultando grandes bases de datos que poseen vocabulario controlado (conocidas como *ontologías*), donde se almacena la información *funcional* a nivel de genes. En las ontologías se puede encontrar el nombre de los genes, en qué procesos biológicos participan, donde actúan, cuales son las publicaciones asociadas, etc. Una vez identificadas la información ontológica-funcional, se aplican metodologías estadísticas para evaluar si la relación que se observa en el experimento es un evento azaroso o no, cuando se lo compara con un comportamiento de referencia o basal (Rivals et al., 2007).

Existen diferentes herramientas para realizar el análisis ontológico-funcional (Huang et al., 2009a). En este contexto, es habitual que el investigador consulte varias de ellas para explotar al máximo las fortalezas de unas, frente a las debilidades de otras, con el fin de sacar el mayor provecho a los resultados experimentales. Por esta razón, se requiere de un elevado dominio por parte del usuario, dado que muchas veces tendrá que exportar la información de una a otra, con el consecuente reformateo de los datos que ello implica. Tal situación produce en ciertos casos, frustración a los usuarios y dificultan su utilización conjunta en un solo paso. A su vez, algunas técnicas son dependientes de la tecnología utilizada y deben ser adaptadas para poder ser utilizadas. Por ejemplo cuando se quiere aplicar sobre datos de proteínas, es necesario obtener su identificador equivalente a nivel de gen, o incluso construir con

una tecnología diferente una referencia apropiada para el contexto experimental.

Por otro lado, la mayoría de las metodologías solo permiten analizar diseños experimentales simples (tipo caso-control), no pudiendo analizar diseños de mayor complejidad, como tampoco incluir información temporal, clínica, etc. Incluso para el caso simple, el usuario es el único responsable de integrar las extensas tablas de salidas obtenidas de la aplicación de diferentes herramientas. De manera que la propia complejidad de integración de resultados, al igual que la falta de técnicas de resumen visual de información que se pueda realizar sobre ellas, limita la capacidad de análisis. Adicionalmente, no existe un patrón de oro (del inglés *gold standard*) para validar los resultados en este tipo de metodologías, recurriendo de forma habitual a una validación mediante literatura científica. De manera que los problemas nombrados anteriormente, impactan negativamente en la extracción de patrones que pueda realizarse sobre la información que pudiese estar disponible, donde la aplicación de técnicas de minería de datos sería de gran provecho en este campo.

Esta tesis proporciona metodologías que facilitan el análisis ontológico-funcional de experimentos genómicos/proteómicos, desde la perspectiva de la minería de datos. En particular, se propone un análisis más estructurado y completo de los datos proporcionados por distintas fuentes de información. Se abordan los problemas que se suscitan al utilizar herramientas desarrolladas para el análisis genómico en el estudio de la proteómica, en lo que respecta a la utilización de una lista de referencia. También se propone una estrategia para la validación de los resultados mediante simulación numérica. Adicionalmente, se facilita y automatiza la indagación de distintas fuentes de información, presentándola de una manera amigable. Esto permite la extracción de patrones, sobre nuevas relaciones inferidas del contraste visual de los resultados obtenidos, en experimentos que presenten dos o más condiciones.

La organización del documento de tesis es como sigue:

Capítulo 1: brinda una visión global del **análisis ontológico-funcional**, las diferentes metodologías y herramientas existentes. Adicionalmente se introduce al lector en los diferentes problemas asociados a este tipo de análisis.

Capítulo 2: introduce al lector al concepto de **minería de datos** en el contexto del análisis ontológico-funcional. Se describen las diferentes etapas involucradas en el análisis (entendimiento del problema y datos, modelado, evaluación y

reporte).

Capítulo 3: presenta los **aportes realizados** en este trabajo de tesis, en el contexto del análisis ontológico-funcional. En este sentido, se profundiza sobre las diferentes contribuciones realizadas en cuanto a la *consistencia e integridad de identificadores*, *exploración multivariada* y su aplicación para el *control de calidad* de los datos, al igual que *integración, visualización, exploración y validación* de los resultados obtenidos. Se presentan tres bases de datos donde se aplicó la metodología propuesta.

Capítulo 4: muestra la **aplicación** de las diferentes estrategias desarrolladas en la presente tesis, sobre dos experimentos genómicos. En el primero se contrasta el impacto funcional de las diferentes configuraciones de la hormona folículo estimulante (FSH) en humanos, es decir, la integración y exploración de las diferentes configuraciones a nivel funcional. En el segundo ejemplo, se profundiza sobre la exploración multivariada y control de calidad de datos, en el estudio en dos órganos, bajo el efecto protector de aceite de pescado en insuficiencia renal aguda inducida por la dieta.

Capítulo 5: presenta las **conclusiones y trabajos futuros** producto de la presente tesis. Se destacan los diferentes aportes realizados al estado del arte, así también como las posibles líneas que se pueden continuar a partir de lo realizado a lo largo del doctorado.

Índice general

Capítulo 1

Análisis Ontológico Funcional

La era de las ciencias "*ómicas*", como la *genómica* (estudio de los genes) y la *proteómica* (estudio de las proteínas), ha dado lugar a grandes revoluciones y avances en la biología. Parte de ello se debe a la incorporación de tecnologías de alto rendimiento (del inglés *high throughput*). Estas tecnologías han permitido pasar del análisis clásico de una única variable (gen o proteína), hacia una evaluación masiva de todos los genes (genoma) o proteínas (proteoma) en forma simultanea, sobre una variedad de diseños experimentales.

Estas tecnologías no sólo producen una gran cantidad de datos, de forma y estructura compleja, sino que también generan grandes bases de datos con la intención de almacenar el conocimiento adquirido como por ejemplo en PubMed (`www.ncbi.nlm.nih.gov/pubmed`). La gran cantidad y variedad de datos con las que estas bases cuentan, obligan a implementar flujos de análisis/procesamiento que presentan diversas complejidades, lo que motiva a utilizar distintas herramientas bioinformáticas, dependiendo de la tecnología utilizada, para aprovechar al máximo la/s particularidad/es de cada una de ellas (Gentleman et al., 2005). Por ejemplo, en cáncer, normalmente se utilizan *microarreglos de ADN* para evaluar la expresión de genes y *geles en diferencias de electroforesis bidimensional* en proteínas, para el descubrimiento de marcadores moleculares en diagnóstico y terapia (Phan et al., 2009). En estas aplicaciones se identifican genes o proteínas *candidatas*, que se expresan diferencialmente en distintas condiciones experimentales, las cuales pueden ir desde el caso más simple en un ensayo tipo *caso-control*, hasta diseños experimentales de *ma-*

yor complejidad. También el investigador puede utilizar diferentes algoritmos, para buscar genes que se expresen de manera similar (coexpresan) en diferentes condiciones experimentales, o incluso utilizar un criterio *ad hoc* para seleccionar los genes o proteínas candidatas.

En este contexto, no solo es posible estudiar genes o proteínas candidatas de forma individual, sino que es posible evaluar cómo responde todo el sistema biológico cuando participan todos los candidatos en su conjunto, sobre una variedad de procesos y/o funciones biológicas conocidas. Esta tarea se lleva a cabo consultando grandes bases de datos que poseen vocabulario controlado (conocidas como *ontologías*), donde se almacena la información *funcional* a nivel de genes. En las ontologías se puede encontrar el nombre de los genes, en qué procesos biológicos participan, donde actúan, cuales son las publicaciones asociadas, etc. Una vez identificada la información ontológica-funcional, se aplican metodologías estadísticas para evaluar si la relación que se observa en el experimento es un evento azaroso o no, cuando se lo compara con un comportamiento de referencia o basal (Rivals et al., 2007). De esta manera, es posible determinar qué funciones y/o vías metabólicas se ven *enriquecidas* (modificadas) por el experimento, lo que se conoce como *enriquecimiento funcional* o *análisis ontológico-funcional*. A tales efectos, el presente capítulo introduce a las diferentes ontologías utilizadas en biología (sección 1.1), metodologías existentes para el análisis (sección 1.2) y herramientas de mayor renombre (sección 1.3), donde se presenta la problemática que se pretende abordar en esta tesis.

1.1. Ontologías

El término *ontología* (del griego *ονταος* "*del ente*", genitivo del participio del verbo *εἰμί* "*ser, estar*" y *λόγος* "*ciencia, estudio, teoría*"), se define como la rama de la filosofía que se ocupa de la naturaleza y organización de la realidad, es decir de lo que "existe". En el campo de la Inteligencia Artificial, una ontología define el vocabulario de un área, mediante un conjunto de términos básicos y relaciones entre dichos términos, así como reglas que combinan términos y relaciones que amplían las definiciones dadas en el vocabulario (Guarino, 1995).

En Biología las ontologías son grandes bases de datos de anotación, que poseen vo-

cabulario controlado para almacenar de forma estructurada, la información existente o conocida. Por ejemplo las ontologías que se utilizan para el análisis ontológico-funcional contienen información sobre las funciones, procesos o locaciones donde actúan, etc. de cada gen. Desafortunadamente no se incluye en ellas información que permita diferenciar de forma apropiada las funcionalidades de isoformas de la misma proteína o considerar splicing alternativo del mismo gen, lo que deriva en una anotación incompleta. Pese a esto, las ontologías son la principal fuente de conocimiento biológico. Las mismas se utilizan con éxito para encontrar rápidamente información relevante de lo que a la fecha se conoce, para el experimento bajo estudio. Más aún, el propio uso favorece la incorporación de nuevo conocimiento, al igual que a su perfeccionamiento (curación).

Existe un gran número de ontologías, dependiendo básicamente del organismo bajo estudio y/o de la problemática que motiva la investigación. Sin embargo, en esta tesis se utilizaron dos de las ontologías de mayor difusión en la comunidad científica, que permiten realizar un análisis de genómica/proteómica funcional: GO (Gene Ontology, Ashburner et al. (2000)) y KEGG (Kyoto Encyclopedia of Genes and Genomes, Kanehisa y Goto (2000)).

1.1.1. Gene Ontology

El consorcio de Gene Ontology (GO, `www.geneontology.org`) fue creado hacia fines de los noventa y su misión inicial fue recopilar la información biológica dispersa de organismos *eucariotas*, en lo que respecta a las funciones asociadas a sus genes (Ashburner et al., 2000). Con el tiempo, pasó de ser una mera herramienta para unificar el vocabulario biológico, a transformarse en la ontología de mayor popularidad. Esto no hubiera sido posible sin que el consorcio por un lado, permitiera la incorporación de nuevos organismos y por otro, liberara el conocimiento a la comunidad científica. Cabe destacar que de esta manera, la propia comunidad participa activamente en la curación (revisión y actualización) de la información disponible.

En esta ontología, la información se encuentra estructurada en tres *categorías* principales:

Procesos biológicos (PB) se refieren a un objetivo biológico al cual un gen o alguno de los productos genéticos asociados a él contribuye. Un proceso se lleva

a cabo a través de uno o más conjuntos ordenados de funciones moleculares. Los procesos, usualmente involucran una transformación química o física, en el sentido de que algo ingresa a un proceso y algo diferente se obtiene a su salida. Esta categoría comprende PB generales como por ejemplo "*crecimiento celular y mantenimiento*" o "*transducción de señales*", hasta más específicos como "*metabolismo de pirimidinas*" o "*biosíntesis de cAMP*".

Funciones moleculares (FM) se definen como la actividad bioquímica de un producto genético, incluyendo la unión a ligandos específicos o estructuras. En este sentido se describe sólo lo que hace, sin precisar dónde ni cuándo se produce el evento en realidad. Ejemplos de FM comprenden desde "*enzimas*", "*transporte*" o "*ligando*" hasta más específicos como "*adenilato ciclasa*" o inclusive "*transporte o transferencia de electrones dentro del ciclo de transporte de electrones en la vía de la fotosíntesis*".

Componentes celulares (CC) se refieren al lugar en la célula donde un producto genético es activo. Estos términos reflejan la comprensión de la estructura celular. En CC se incluyen términos como "*ribosoma*", "*membrana nuclear*" o "*aparato de Golgi*".

Por lo tanto en GO, se tiene la información de los genes que "participan" en un determinado *concepto/término* biológico perteneciente a PB y/o FM, y en qué CC "actuan". A su vez, un mismo gen puede participar en más de un término dentro de la misma categoría principal (PB, FM o CC), e incluso en más de una de ellas.

En esta ontología la información se almacena utilizando estructuras jerárquicas en forma de grafos dirigidos acíclicos (GDA), para cada una de las categorías principales. En cada uno de los GDA, los conceptos/términos biológicos se representan como *nodos* en la estructura. Cada nodo tiene asociado los genes que se relacionan con el concepto que éste representa, como se aprecia en la figura 1.1. Adicionalmente, cada GDA se encuentra organizado de manera jerárquica mediante relaciones entre nodos del tipo "*es un*" o "*es parte de*". En este sentido el nodo de más arriba del grafo (nodo raíz), representa el concepto/término más genérico posible (PB, FM o CC). A medida que se desciende por la jerarquía (se recorre el grafo), los nodos poseen un grado de especificidad mayor en el proceso en sí mismo y por ende, la cantidad

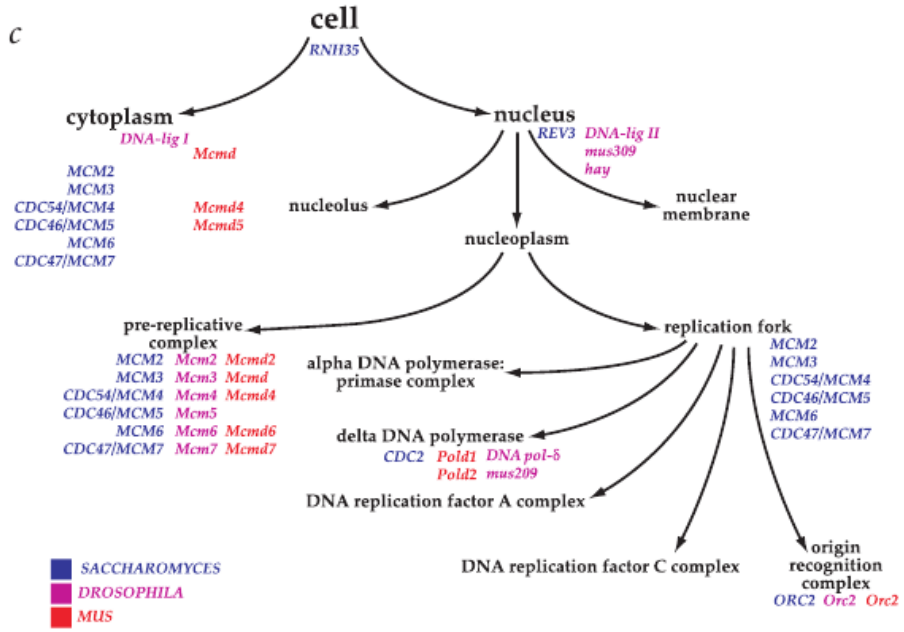


Figura 1.1: Esquema simplificado de la ontología de *Componentes Celulares* (CC) para tres organismos (en colores diferentes). Por simplicidad no se han incluido todos los genes. Imagen extraída de Ashburner et al. (2000).

de genes asociados a él es menor. Por ende, un gen que se encuentre anotado en un nodo dado, también se encontrará presente en todos los nodos de sus ancestros de la misma rama del grafo. Adicionalmente, dicho gen puede estar asociado a nodos de diferentes ramas del mismo grafo, e incluso en diferentes grafos (PB, FM o CC).

1.1.2. Kyoto Encyclopedia of Genes and Genomes

El proyecto de la enciclopedia de genes y genomas de Kyoto, conocida por sus siglas del inglés KEGG (Kyoto Encyclopedia of Genes and Genomes), comenzó en mayo de 1995 bajo el programa del genoma humano, fomentado por el ministerio de educación, ciencia, deportes y cultura de Japón (Kanehisa y Goto, 2000). Este proyecto se impulsó con la finalidad de formar una base de conocimientos para el análisis sistemático de las funciones de los genes y vincular la información genómica con funciones de mayor orden, en lo que se conoce como *vías metabólicas*.

En esta ontología, la información se encuentra almacenada en tres grandes bases de datos. La primera de ellas almacena la información de los "**genes**" para todos los genomas completamente secuenciados y algunos parcialmente secuenciados. En esta base se puede encontrar para cada gen: el nombre, la secuencia de nucleótidos o aminoácidos, posición en el genoma, organismos en los cuales se puede encontrar, al igual que los identificadores del mismo gen en otras bases de datos. Además se puede indagar sobre aquellos genes muy similares a él en diferentes organismos, que provienen de un ancestro común (ortólogos), y genes que sufrieron una duplicación en el mismo organismo evolucionando de manera independiente (parálogos). Cabe destacar que existen muchos genes para los cuales no se conoce información de su función biológica ni localización. En estos casos, solo se dispone de información parcial en la base de datos.

La segunda base de datos, estructura y vincula las funciones de diferentes genes que participan en determinado proceso o vía metabólica. Estas vías se representan mediante gráficos como por ejemplo para el "*ciclo de Krebs* o *ciclo del ácido cítrico*" (figura 1.2), "*transporte de membrana*", "*transducción de señales*" o "*ciclo celular*". En estos gráficos se identifican tres tipos de elementos: *cajas rectangulares* codificados por cuatro números (EC, del inglés Enzyme Commission number) para representar productos de genes, *flechas* para el flujo de las reacciones y *cajas con bordes redondeados* para vincular a otras vías que participan en el proceso. A su vez, esta base de datos se complementa con un conjunto de tablas de grupos ortólogos, donde se encuentra la información de subvías conservadas, las cuales son de especial utilidad para la predicción de funciones de genes.

Las dos bases anteriores se complementan con una tercera base de datos de "**ligandos**". En ella se encuentra la información acerca de los compuestos químicos, moléculas y reacciones enzimáticas asociadas a los genes y vías metabólicas involucradas. Adicionalmente KEGG proporciona de herramientas para explorar la información de genes y visualizar vías metabólicas, mapas de genomas, entre otras. Desafortunadamente, a partir de julio de 2011, el consorcio decidió dejar la política de acceso público, para cobrar diferentes tarifas por el acceso a la información disponible. No obstante, algunas herramientas permiten acceder a versiones previas de forma gratuita.

1.2. El análisis de enriquecimiento funcional

El análisis de **enriquecimiento ontológico-funcional** permite evaluar el impacto a nivel sistémico, de un grupo de genes *candidatos* o *activos* que en su conjunto alteran, modifican o *enriquecen* por ejemplo un proceso biológico, función molecular, vías metabólicas, etc. debido al experimento/tratamiento. Para ello se recurre a diferentes bases de datos como GO o KEGG, para obtener la información ontológica-funcional relacionada a dichos genes candidatos. Una vez seleccionada la fuente de información ontológica, el investigador debe optar entre algunas de las metodologías

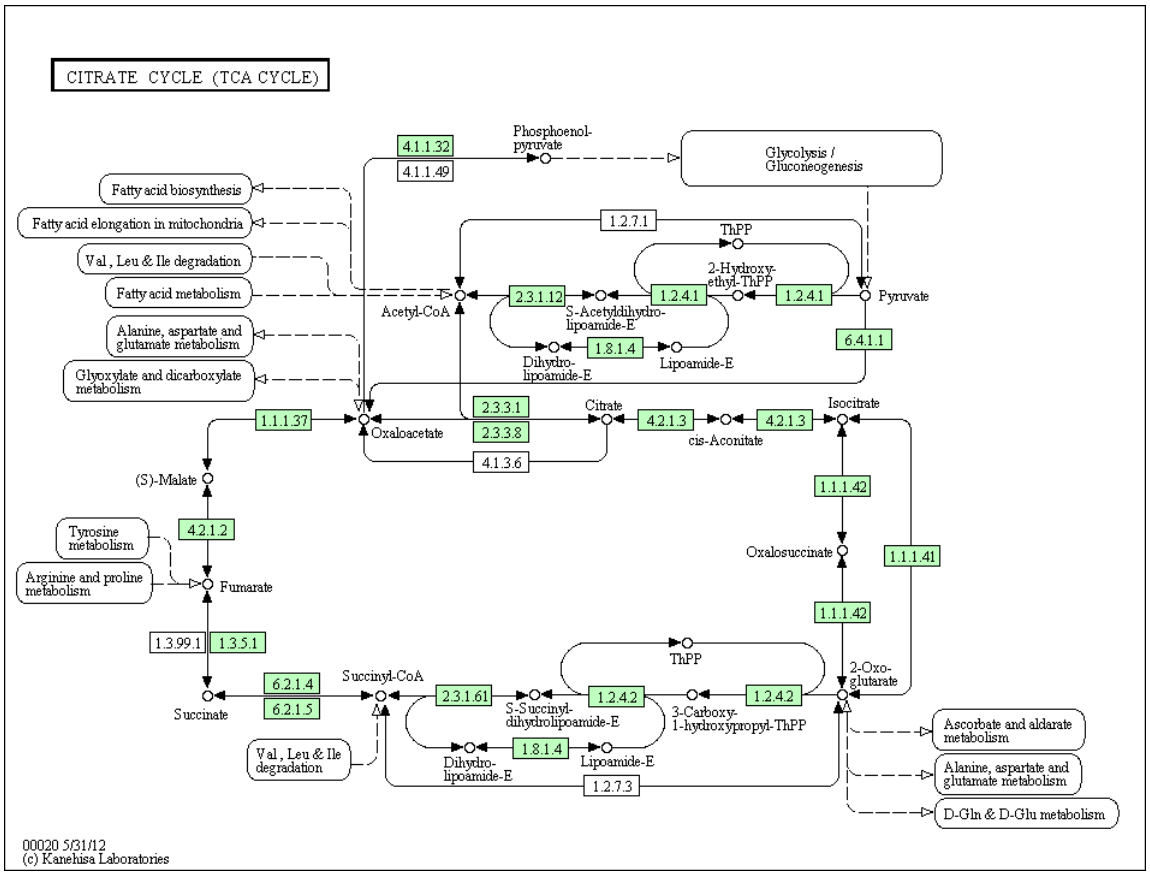


Figura 1.2: Ejemplo de una vía de KEGG en este caso es el *ciclo del ácido cítrico* para humanos. Las cajas rectangulares representan productos de genes y en cajas con bordes redondeados otras vías que participan en el proceso. Imagen extraída de www.genome.jp/kegg-bin/show_pathway?hsa00020

disponibles para realizar el análisis funcional.

1.2.1. Metodologías de análisis de enriquecimiento funcional

Existen tres grupos bien definidos para realizar el análisis (Huang et al., 2009a):

SEA es el acrónimo de *Singular/Set Enrichment Analysis*. Tradicionalmente utilizado debido a su trayectoria al igual que disponibilidad en herramientas/portales digitales (Khatri y Draghici (2005) y Huang et al. (2009a)). Esta metodología necesita definir una segunda lista adicional (de *referencia*), que servirá para especificar lo que es esperable del comportamiento basal del modelo biológico. Una vez definida, se realiza para cada concepto/término de interés, en forma independiente, un test de hipótesis para comparar las proporciones observadas sobre los candidatos respecto a la referencia. Así, un término resultará enriquecido si en dicha comparación existe evidencia de que las proporciones son diferentes, a un nivel de significancia definido por el usuario.

GSEA es el acrónimo de *Gene Set Enrichment Analysis*. Esta metodología construye un ordenamiento inducido en la totalidad del perfil de expresión de las muestras pertenecientes a dos clases: control y tratamiento. El objetivo de la misma es determinar si todos los miembros de un concepto/término de interés, están distribuidos al azar o no (en algún extremo) a lo largo del mismo (Subramanian et al., 2005). Para ello, los genes se disponen en base a un criterio de ordenamiento como la correlación entre su expresión y la distinción de clases, o alguna otra métrica plausible para generar un ranking. Luego se recorre el ordenamiento para calcular el máximo del enriquecimiento inducido. A cada gen se aplica una función de costo que aumenta (disminuye) proporcional a la correlación de su nivel de expresión con el fenotipo de las clases, cada vez que encuentre un gen que pertenezca (o no) a la lista de miembros de la categoría de interés. Luego, este enriquecimiento se compara contra la distribución nula generada a partir de permutaciones en las etiquetas de las clases, a los efectos de evaluar si el ordenamiento observado es esperable o no por azar. En caso de que difiera del azar, existen distintos criterios para definir cuáles son los genes del experimento de interés en el término. Usualmente se utilizan los más

próximos al máximo o los del segmento más pequeño entre el máximo del costo y el comienzo o fin de la lista.

MEA es el acrónimo de *Modular Enrichment Analysis*. Este método toma como punto de partida cualquiera de los anteriores (usualmente SEA), a los efectos de incluir "la redundancia de la red biológica" en el análisis, dada las relaciones existentes entre los términos explorados. Una posibilidad es aplicar un agrupamiento a los resultados de SEA. Para ello, cada término se puede codificar como un vector binario representando la pertenencia (o no), de cada gen de la lista de interés anotado a él. Luego se realiza un agrupamiento de los términos utilizando algún estadístico como por ejemplo Kappa (Cohen et al., 1960). Posteriormente, el enriquecimiento asociado a cada grupo, se define mediante alguna operación realizada sobre los valores p de cada término perteneciente al agrupamiento. Por ejemplo, Huang et al. (2007) utiliza la media geométrica medida en escala logarítmica a tales efectos.

Los dos primeros métodos se usan para saber si un término se encuentra enriquecido (o no) en la condición analítica estudiada. No obstante, la formulación del problema es diferente para cada caso. Para realizar el análisis vía SEA es necesario proporcionar dos listas, una de ellas de referencia y la otra de los genes candidatos. Esta última suele estar conformada por aquellos genes que hayan sido identificados como expresados diferencialmente, entre distintas condiciones experimentales (e.g. caso-control) para un umbral definido. En GSEA se utiliza una única lista que contiene la totalidad de genes disponibles por ejemplo en un chip, si fuera el caso de utilizar microarreglos de ADN, para luego utilizar el criterio de ordenamiento propuesto para medir el enriquecimiento. En este sentido GSEA respecto a SEA, no utiliza un umbral para definir la lista de candidatos. En general, los resultados con SEA y GSEA son similares, pese a que no existe un estándar para su comparación (Hung et al., 2012). Sin embargo, la gran debilidad que presentan ambos métodos se debe a la aplicación en forma independiente, a cada concepto/término biológico, perdiéndose así la relación entre ellos.

En esta tesis se utilizó SEA como motor de cálculo del enriquecimiento funcional, y MEA para integrar/explorar los resultados. En este sentido, SEA se centra en comparar una **lista de genes candidatos** contra los genes de una **lista de referencia**

para encontrar términos enriquecidos, conociendo *a priori* quiénes son los integrantes de los distintos términos a evaluar. Formalmente el iésimo término ($Término_i$) responde a una distribución **Hipergeométrica**, a la cual se le realiza una prueba de hipótesis de homogeneidad o independencia (Walpole et al., 1999). Según la herramienta bioinformática elegida, esta prueba de hipótesis puede realizarse de diferentes maneras (Rivals et al., 2007). Hay herramientas que utilizan la propia distribución **Hipergeométrica** (BINGO, Maere et al. (2005); CLENCH, Shah y Fedoroff (2004); GeneMerge, Castillo-Davis y Hartl (2003)), una aproximación con una distribución **Binomial** (CLENCH, Shah y Fedoroff (2004); GFINDER, Masseroli et al. (2004); GOToolBox, Martin et al. (2004)) o aquellos basados en **tablas de frecuencias** observadas de tamaño 2x2 con totales marginales fijos, para dos categorías mutuamente excluyentes: en filas *Candidatos* o no ($Candidatos^c$) y en columnas $Término_i$ de interés o no ($Término_i^c$):

Tabla 1.1: tabla de contingencia 2x2 para el i-ésimo término de interés

	Término$_i$	**Término$_i^c$**	**Total**
Candidatos	n_i	$N_{Candidatos} - n_i$	$N_{Candidatos}$
Candidatos^c	$n_{Término} - n_i$	$(N - N_{Candidatos}) - (n_{Término} - n_i)$	$N - N_{Candidatos}$
Total	$n_{Término}$	$N - n_{Término}$	N

El total de genes de la lista de referencia (N), se encuentra dividido en filas en caso de pertenecer o no a la lista de candidatos (Candidatos o $Candidatos^c$); mientras que las columnas determinan la pertenencia (o no) de los genes al término de interés ($Término_i$ o $Término_i^c$).

En la tabla 1.1 se muestra cómo los $N_{Candidatos}$ genes pertenecientes a la lista de *Candidatos*, se encuentran divididos en n_i genes pertenecientes al $Término_i$ de interés y aquellos que no pertenecen al término de interés ($Término_i^c$), es decir, $N_{Candidatos} - n_i$ genes. A su vez, la lista de referencia determina la cantidad total de genes de la tabla (N) y la cantidad de genes que pertenecen al término de interés ($n_{Término}$). Consecuentemente, $Candidatos^c$ es el conjunto complementario de genes, que no pertenece al conjunto de *Candidatos*, es decir, el remanente de genes de la lista de referencia. Este conjunto contiene $N - N_{Candidatos}$ genes, distribuidos en $n_{Término} - n_i$ genes sobre el término de interés que no pertenecen a la lista de candidatos, y aquellos que no se encuentran en el término de interés $(N - N_{Candidatos}) - (n_{Término} -$

n_i), dejando completamente determinada la tabla.

Usualmente estas tablas de frecuencias se analizan mediante una prueba exacta de **Fisher** (GOstat, Falcon y Gentleman (2007); GoMiner, Zeeberg et al. (2003); DAVID, Dennis Jr et al. (2003); EASEonline, Hosack et al. (2003)) o una aproximación para grandes muestras con una **prueba** χ^2 de un grado de libertad (GoSurfer, Zhong et al. (2004); Onto-Express, Khatri et al. (2002); CLENCH, Shah y Fedoroff (2004)) para cada uno de los i términos evaluados (cientos a miles). No obstante, en esta metodología el investigador se ve obligado a definir una lista de referencia para completar la tabla 1.1, es decir definir N y $n_{Término}$. Dicha lista impacta sobre los resultados obtenidos, por lo que su definición/elección no es trivial.

1.2.2. Selección de lista de referencia

La mayoría de las herramientas que realizan SEA tales como DAVID (Base de datos para Anotación, Visualización y Descubrimiento Integrado, Dennis Jr et al. (2003) y Huang et al. (2007)), permiten al usuario elegir una lista de referencia (LR) de una lista de posibilidades:

LR-I El genoma de la especie en estudio.

LR-II La lista de genes presentes en el chip para experimentos de microarreglos.

LR-III Una lista especificada a criterio del usuario.

Por lo general, el genoma (LR-I) es la opción por defecto en la mayoría de las herramientas. Sin embargo, desde un punto de vista analítico, el uso de diferentes LRs podría producir resultados diferentes. Más aún, una inapropiada definición/elección de la LR podría contradecir los supuestos estadísticos, potencialmente sesgando la interpretación de los resultados (Khatri y Draghici, 2005). Para vislumbrar esto, consideremos por ejemplo un experimento proteómico para estudiar las proteínas que se encuentran en el espacio extracelular (secretoma), bajo un estudio tipo caso-control. En este contexto, no se está accediendo a "todo" el proteoma, sino a un subconjunto de éste que se encuentra fuera de la célula. Una vez definido un término de interés como *apoptosis* (muerte celular programada) e identificadas las proteínas candidatas

($N_{Candidatas} = 80$), el número de ellas pertenecientes al término ($n_{Candidatas} = 10$) será conocido, dejando determinada por completo la primer fila de la tabla 1.2.

Tabla 1.2: tabla de contingencia 2x2 para el ejemplo proteómico sobre *apoptosis*

	Apoptosis	**Apoptosis**^{c}	**Total**
Candidatos	$n_i = 10$	$N_{Candidatos} - n_i = 70$	$N_{Candidatos} = 80$
Candidatos^c	$n_{Term} - n_i$	$(N - N_{Candidatos}) - (n_{Term} - n_i)$	$N - N_{Candidatos}$
Total	n_{Term}	$N - n_{Term}$	N

Note que hasta no definir la lista de referencia, el gran total N y el tamaño del término (n_{Term}) se encuentran indefinidos, por ende, la segunda fila se encuentra indeterminada.

No obstante, para completar el resto de las celdas de la tabla 1.2, es necesario definir N y $n_{Término}$, es decir una LR. Asumiendo que la verdadera LR es conocida (LR_v), será conocido el gran total de proteínas $N_v = 750$, y por lo tanto quedará definida la cantidad de proteínas que pertenecen al término ($n_{Term_v} = 60$). Cabe destacar que la lista de candidatos es un subconjunto contenido dentro de la LR. Esta configuración se encuentra esquematizada mediante líneas en el panel A de la figura 1.3, donde se ha destacado la correspondencia con las celdas de la tabla 1.2. Por consiguiente, una selección de la LR diferente, podría potencialmente modificar el resultado del test estadístico (valores p) de acuerdo a los siguientes escenarios simulados en la figura 1.3.A:

a) La longitud de la LR varía modificando la cantidad de proteínas totales que posee desde: un número menor al caso real N_1, pasando por el caso real N_v, hasta alcanzar un número mayor al real N_2, es decir, $N_1 < \ldots < N_v < \ldots < N_2$, mientras que el tamaño del término, se encuentra compuesto por las mismas proteínas que en el caso real (es constante), por lo que $n_{Term} = n_{Term_v} = cte$.

b) La elección de la LR contiene una cantidad de proteínas en el término de interés, *apoptosis*, diferente a las del caso real. Por lo tanto, varía el tamaño del término pudiendo ser mayor o menor al real, es decir, $n_{Term_1} < n_{Term_V}$ o $n_{Term_2} > n_{Term_V}$, mientras que la cantidad de proteínas totales de la LR, se mantiene constante, por lo que $N = N_v = cte$.

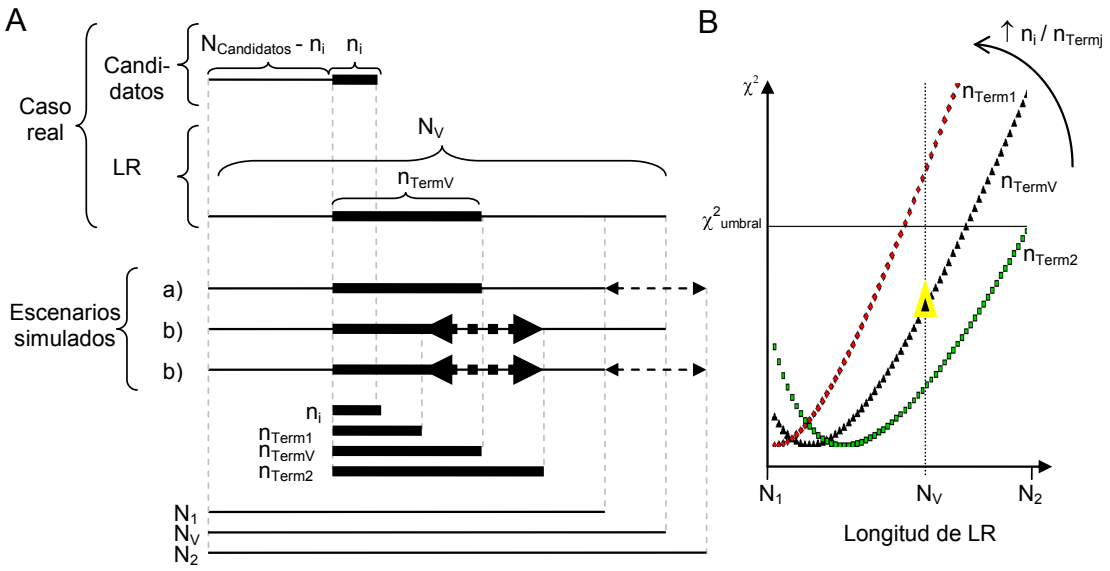


Figura 1.3: Simulación de χ^2 utilizando los siguientes parámetros: $n_i = 10$, $N_{Candidadtos} = 80$, $N_1 = 400$, $N_V = 750$, $N_2 = 1000$, $n_{Term1} = 50$, $n_{TermV} = 60$ y $n_{Term2} = 70$. A) Esquema de diferentes escenarios. B) Gráfico de χ^2 contra la longitud de la lista de referencia (LR) para las diferentes configuraciones.

c) Una combinación de los dos escenarios anteriores. Esto es lo que que realmente sucede cuando se cambia la LR (genoma, lista de genes del chip o una referencia definida por el usuario).

En el panel B de la figura 1.3 se muestra la evaluación del estadístico χ^2 a medida que se varia la longitud de la LR desde $N_1 = 400$, pasando por $N_V = 750$, hasta $N_2 = 1000$ y tres tamaños de término en curvas de colores: $n_{Térm_1} = 50$ (rombos rojos), $n_{Térm_v} = 60$ (triángulos negros) y $n_{Térm_2} = 70$ (rectángulos verdes).

En ella se puede apreciar que a medida que se incrementa la longitud de la LR como en el escenario a), se producen valores más altos de χ^2, independiente del tamaño del término (curvas de colores). Es decir que, el sólo hecho de aumentar la cantidad de miembros de la LR, produce una mayor posibilidad de superar la línea horizontal a la altura χ^2_{umbral}, que delimita el umbral de enriquecimiento, para un determinado nivel de significación (por ejemplo $\alpha = 0{,}05$). No obstante, el escenario biológico real corresponde a un único valor χ^2_v, representado por un triángulo amarillo

con centro negro, en la curva $n_{Térm_v}$ situado a una longitud de LR $N = N_V$. Cabe destacar que este valor es inferior al umbral, es decir, $\chi^2_v < \chi^2_{umbral}$. Por lo tanto, el término apoptosis no se encuentra enriquecido en este experimento.

En el escenario b), para una longitud de la LR similar a la verdadera (línea vertical de puntos con $LR = N_V = 750$) un cambio en el tamaño del término al pasar del caso real (curva de triángulos negros) a uno menor (curva de rombos rojos de $n_{Térm_1} = 50$), aumenta el solapamiento $n_i/n_{Térm_j}$. Esta situación implica valores χ^2 más elevados, pudiendo superar así el umbral de enriquecimiento producto de la reducción del tamaño de término. Por el contrario, para tamaños de términos más grandes (curva de rectángulos verdes de $n_{Térm_1} = 70$), disminuye el solapamiento con la consecuente disminución en la pendiente de la curva. Por ende, el valor estadístico para la misma longitud de la referencia es menor al caso real. Más aún, para este tamaño de término (curva de rectángulos verdes) se necesita una longitud aún mayor al real (curva de triángulos negra), para alcanzar el umbral de enriquecimiento χ^2_{umbral}.

Por otra parte, el escenario c) representa una combinación de modificaciones de tanto la longitud de la cantidad de proteínas en la LR (variación de N) y tamaño de término (cambio de curvas de colores). De manera que es esperable obtener valores del estadístico χ^2 entre los resultados de los dos escenarios anteriores: a) y b).

Como se mostró en SEA, la elección de diferentes LRs puede introducir sesgo en el enriquecimiento funcional. En este contexto, al utilizar referencias muy grandes como el genoma (**LR-I**, *por defecto* en la mayoría de las herramientas), proteínas que no son detectables (por ejemplo, porque sólo estudiamos las proteínas extracelulares) o bien porque la tecnología no tiene la resolución suficiente, van a estar presentes en la LR. En consecuencia, se introduce un sesgo en el análisis por no satisfacer los supuestos estadísticos (Zeeberg et al., 2003). En este caso se impone una distribución condicional, donde no todos los miembros marginales son capaces de formar parte de cualquier celda de la tabla de contingencia 1.2. Es decir, hay proteínas que pertenecen a la segunda fila, que "nunca" podrán ser parte de la lista de candidatos ya que es imposible tener mediciones de ellos.

Una situación similar podría estar presente en los experimentos de *microarreglos de ADN*, cuando se utiliza el genoma (**LR-I**) en un chip personalizado para una enfermedad particular (cáncer, Parkinson, etc.), o al utilizar la lista de genes del

chip entero (**LR-II**) en lugar de la lista de genes definido por el usuario (**LR-III**), teniendo en cuenta aquellos genes detectados en el experimento según los controles de calidad del fabricante (Affymetrix (2004), Archer y Reese (2010) Hackstadt y Hess (2009), McClintick y Edenberg (2006) y Bourgon et al. (2010)). En este sentido, los investigadores deben utilizar sólo aquellos genes que están "sistemáticamente" presentes en el estudio.

Las tecnologías de secuenciamiento de nueva generación (*NGS*, del inglés Next Generation Sequencing), potencialmente podrían detectar todos los genes presentes en la muestra. No obstante, la LR-III dependerá fuertemente de la profundidad de secuenciamiento. Por lo tanto en SEA, cualquier tecnología podría no estar proporcionando la LR adecuada, siendo en cualquier caso un desafío para el investigador. Si bien no existe un estándar para la comparación de resultados obtenidos y en la literatura es usual no encontrar contra cuál de ellos se realizó el análisis, es una buena idea utilizar una LR que contenga todos los posibles candidatos a elegir de la muestra (LR-III). En este contexto frente a diferentes elecciones, Huang et al. (2009a) aseguran que es más importante la estabilidad de los genes encontrados que los valores p del enriquecimiento.

1.3. Herramientas para análisis de SEA y MEA

Existen diferentes herramientas para realizar el análisis de enriquecimiento ontológico funcional mediante SEA y MEA (Huang et al., 2009a). En su mayoría siguen el esquema de procesamiento de la figura 1.4, donde se utiliza como dato de entrada la información de anotación almacenada en las ontologías (sección 1.1) y los identificadores tanto de los genes candidatos, al igual que los pertenecientes a una referencia dada. Luego, se relaciona la información de los identificadores con la almacenada en las bases de anotación funcional, para realizar el análisis propiamente dicho, a los efectos de identificar los términos enriquecidos (modificados) utilizando alguna de las metodologías plausibles (sección 1.2.1). Finalmente, estas herramientas presentan los resultados para la exploración del investigador.

En este contexto, es habitual que el investigador consulte diferentes herramientas para explotar al máximo las fortalezas de unas, frente a las debilidades de otras,

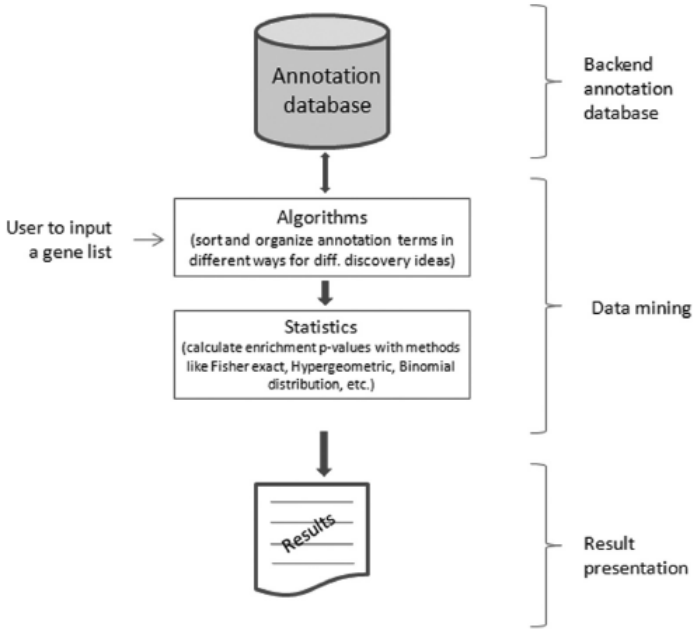


Figura 1.4: Esquema de flujo de trabajo en herramientas de enriquecimiento funcional. Tres grandes bloques se identifican: bases de datos de anotación ontológica; minería de datos, donde los identificadores de los genes candidatos se relaciona con la información de anotación y realiza el procesamiento del enriquecimiento funcional; finalmente se presentan los resultados. Imagen extraída de Huang et al. (2009a).

a los efectos de sacar el mayor provecho a los resultados experimentales. En este sentido, introduciremos tres herramientas contemporáneas como *GoMiner* (Zeeberg et al., 2003, 2005), *DAVID* (Dennis Jr et al., 2003; Huang et al., 2009b, 2007) y *GOstat/s* (Beissbarth y Speed, 2004; Falcon y Gentleman, 2007), donde se presentan los diferentes aspectos/problemática involucrada en cada etapa del análisis funcional y cómo lo abordan cada una de ellas.

1.3.1. Formas de acceder a las herramientas

Las tres herramientas permiten acceso mediante una página web. Sin embargo, para realizar diferentes procesamientos computacionalmente intensivos, es fundamental poder accederlas de *forma programática* (sin supervisión de un usuario). En este contexto, *GoMiner* ofrece adicionalmente la posibilidad de instalar un cliente

Java®, con la posibilidad de acceder al repositorio de Zeeberg et al. (2003) o incluso instalar las bases de datos de forma local.

Por su parte *DAVID*, inicialmente habilitó una interfaz programática mediante mensajes utilizando localizador uniforme de recursos (URL del inglés Uniform Resource Locator), limitando la consulta a una longitud de hasta 1024 caracteres. Interfaces vía servicios web fueron implementadas por Jiao et al. (2012), utilizando clientes para los lenguajes Java®, Perl®, Python® y Matlab®, dejando afuera uno de los lenguajes más difundidos en la comunidad bioinformática como R®.

En el caso de *GOstat*, Falcon y Gentleman (2007) escribieron el paquete *GOstats* en lenguaje R®, siguiendo la filosofía del desarrollo original, siendo uno de los que ofrecen una implementación programática para consultar las base de datos. No obstante, en la comunidad de biólogos esta herramienta no ha tenido repercusión en su uso, dado a que es necesario saber programar en lenguaje R® para explotar su potencialidad como lo haría un bioinformático. Adicionalmente, esta herramienta solo utiliza GO como base de datos de anotación.

1.3.2. Versiones y reproducibilidad de resultados

La información de la versión de base de datos que poseen instaladas las herramientas, es difícil de obtener. En caso de ofrecerlo como en *DAVID* y *GoMiner*, no es posible seleccionar una versión de la base de datos específica. Más aún, la versión web actual de *GoMiner* (`discover.nci.nih.gov/gominer/htgm.jsp`), utiliza el motor 328 y la última actualización de la base de datos es de enero de 2011. Adicionalmente, ya no es posible la instalación local de la base de datos del cliente Java®, dado que los esquemas de las bases han sido actualizados y el cliente no puede utilizarlas para realizar un análisis funcional, porque el software no es mantenido en el tiempo. Esta situación sesga el análisis funcional al incluir conocimiento parcial y no el estado actual del arte.

Por su parte, *DAVID*, posee un ciclo de mantenimiento bianual de la base de datos y el motor de cálculo (versión 6.7) data desde enero de 2010. En este sentido los resultados son solo reproducibles en tanto y en cuanto no se actualice la base de datos. Esta particularidad permite que diferentes miembros de un mismo equipo de investigación puedan obtener los mismos resultados, e incluso revisores de artículos

científicos o personas externas que deseen reproducir los resultados.

En el caso de *GOstat*, el sitio web (gostat.wehi.edu.au) no especifica la versión de anotación que utiliza, probablemente sea la de la fecha de la publicación (2004), siendo en todo caso la más obsoleta de todas las herramientas, con resultados siempre reproducibles. Por el contrario, en *GOstats* se debe especificar de forma explícita el paquete de anotación, el cual se pueden obtener de Bioconductor (www.bioconductor.org) y tienen una actualización cada seis meses. De manera que para reproducir los resultados, el usuario debe utilizar la misma versión usada originalmente. Cabe destacar que en las últimas dos herramientas, el usuario puede utilizar un archivo de anotación creado para satisfacer sus propias necesidades, como sucede cuando se desea realizar enriquecimiento funcional únicamente en GO sobre un organismo aún no anotado.

En la mayoría de los casos, la imposibilidad de reproducción de los resultados publicados en un artículo científico se debe a que el/los autores no especifican con qué versión de la herramienta han llevado a cabo el análisis. En este sentido es usual no poder utilizar la misma lista de genes candidatos publicada, dado que por ejemplo los símbolos de los genes han sido actualizados y por ende no son reconocidos por el sistema. Más aún, casi la totalidad de los artículos no reportan la lista de referencia utilizada.

1.3.3. Carga de datos

El ingreso de datos a cualquiera de estas herramientas requiere que el usuario suba dos listas de identificadores: una para los candidatos que se quiere analizar y la otra para establecer la referencia de comportamiento basal. No obstante, *GoMiner* utiliza como identificadores los símbolos (mnemónico corto) de los genes y un signo/sentido del valor de expresión asociado a cada gen candidato: 1/-1 cuando el gen esté sobre/sub-expresado en relación a la referencia, respectivamente. En el caso de *DAVID*, existe una gran versatilidad dado que admite 34 bases de datos de identificadores diferentes como por ejemplo: Affymetrix®, Agilent®, Illumina®, Ensembl, RefSeQ, UniProt, WormBase, EntreZ, etc. Por el contrario, *GOstats* sólo permite utilizar identificadores de paquetes de anotación de plataformas comerciales como Affymetrix® y Agilent®, o únicamente EntreZ en el caso de paquetes de or-

ganismos disponibles en Bioconductor (Gentleman et al., 2004). Por otro lado, tanto *DAVID* como *GOstat/s* no utilizan información del sentido de la expresión, es decir, sólo requieren del identificador.

En el caso que el usuario desee utilizar más de una herramienta, deberá tener un elevado dominio de ellas, dado que muchas veces tendrá que exportar la información de una a otra. Consecuentemente será necesario convertir los identificadores requeridos por una herramienta a alguno de los compatibles en la otra, situación que produce en ciertos casos frustración a los usuarios y dificultan la utilización conjunta de estas herramientas en un solo paso. A su vez, las ontologías se encuentran anotadas a nivel de genes. De manera que si se quiere aplicar sobre datos de proteínas, es necesario obtener los identificadores equivalentes a nivel de gen. Mas aún, en el estudio de secretoma, se deberá construir una referencia apropiada para el contexto experimental, utilizando en algunos casos una tecnología diferente a la empleada para identificar a los candidatos.

Por otra parte, todas las herramientas deben permitir seleccionar sobre qué ontología se realizará el análisis funcional. En este contexto, *GOMiner* y *GOstat/s* han sido desarrolladas para explotar al máximo GO (sección 1.1.1), como parte de su nombre lo indica. *DAVID*, por otra parte, permite no sólo el uso de GO sino también de una amplia variedad de repositorios biológicos y científicos agrupados en 10 grandes categorías: **enfermendades** (*Omin, Genetic_association_db,* etc.), **categorías funcionales** (*COG_ontology, PIR_seq_feature,* etc.), **GO** (*PB, FM* y *CC* a diferentes niveles), **anotaciones generales** (*EntreZ_gene_summary, Cytoband,* etc.), **literatura** (*PubMed, GeneRif_summary,* etc.), **acceso principal** (*Ensembl, EntreZ_gene_id*, etc.), **vías metabólicas** (*KEGG, Biocarta, Panther, Reactome*, etc.), **dominio de proteínas** (*PFam, Interpro,* etc.), **interacción de proteínas** (*Bind, Mint,* etc.) y **expresión te tejidos** (*UP_tissue*, etc.).

1.3.4. Análisis de enriquecimiento funcional

La mayoría de las herramientas de SEA (sección 1.2.1) sólo permiten analizar diseños experimentales simples (tipo caso-control), es decir, cómo se comporta una lista de genes candidatos con respecto a la referencia. Consecuentemente no es posible analizar, *a nivel funcional*, diseños de mayor complejidad, como tampoco incluir

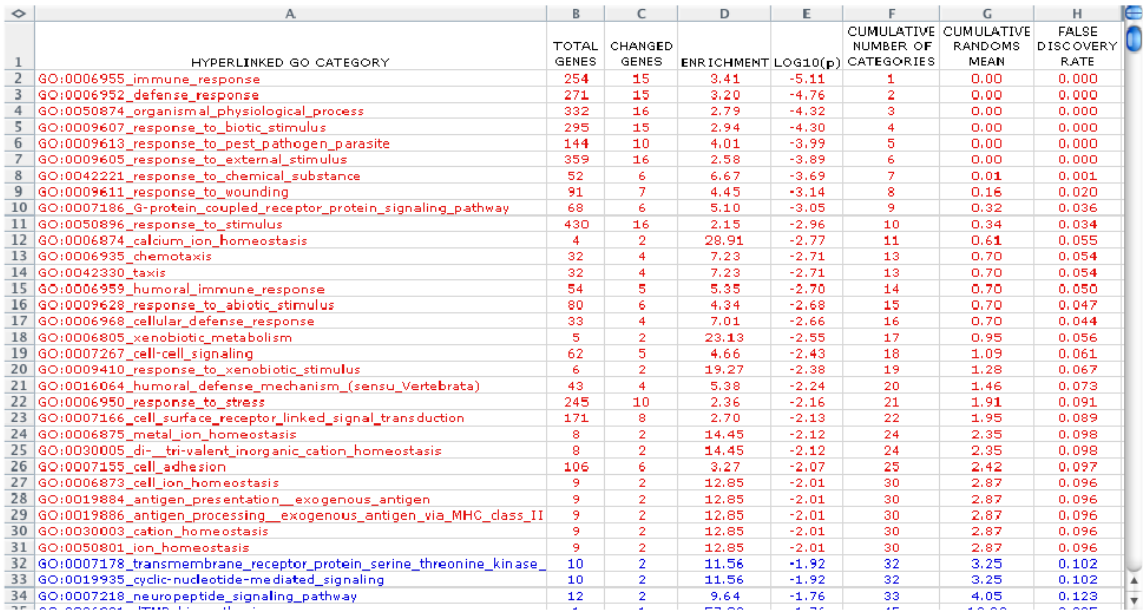
información temporal, clínica, etc. No obstante, en el caso que el usuario utilice diferentes listas de candidatos con diferentes factores experimentales, estas herramientas no permiten integrar/comparar los resultados obtenidos. Incluso para el caso simple, el usuario es el único responsable de integrar las salidas de las diferentes herramientas. Por otra parte, no existe un "patrón de oro" (del inglés *gold standard*) para validar los resultados en este tipo de metodologías, recurriendo de forma habitual a una validación mediante literatura científica, o una validación biológica por una técnica diferente.

GoMiner y *DAVID* adicionalmente permiten realizan un análisis de tipo MEA (sección 1.2.1). En el caso de *GoMiner* este procesamiento está solamente disponible desde su sitio web, donde se realiza un agrupamiento (*clustering* en inglés) utilizando los valores de intensidad o signos de expresión de los genes contra las anotaciones funcionales, representándolos mediante un mapa de calor (genes vs términos). No obstante, esta funcionalidad no está presente para el cliente Java®. Por otra parte, *DAVID*, puede realizar un agrupamiento como el descripto en la sección 1.2.1, *agrupando términos* utilizando la evidencia de anotación de los genes o, *agrupando genes* utilizando la evidencia de anotación. Cabe destacar que esta funcionalidad sólo está disponible en su sitio web (`david.abcc.ncifcrf.gov`).

1.3.5. Visualización de resultados

Las salidas de herramientas de SEA son por lo general **listas tabulares** extensas, en el orden de cientos a miles de filas por decenas de columnas, como se muestra en la figura 1.5 para *GOMiner*. En ella se pueden apreciar cada uno de los términos analizados de la ontología utilizada (e.g. GO) en filas, donde se puede encontrar la totalidad de genes pertenecientes al término, cuantos de ellos son de la lista de candidatos y estadísticos asociados al análisis (valor p, FDR, etc.).

Todas estas herramientas pueden generar también **reportes HTML** como los mostrados para *GOstat* (figura 1.6(a)) y *DAVID* (figura 1.6(b)). No obstante, en ambos casos se requiere de un gran esfuerzo para *exploración* debido a la extensión del reporte, e *interpretación/vinculación* de estos resultados funcionales con la información de expresión de los genes. En *GOstat* la inspección de los términos GO se debe hacer posicionándose arriba del código (e.g. GO:0004872) para obtener el nom-

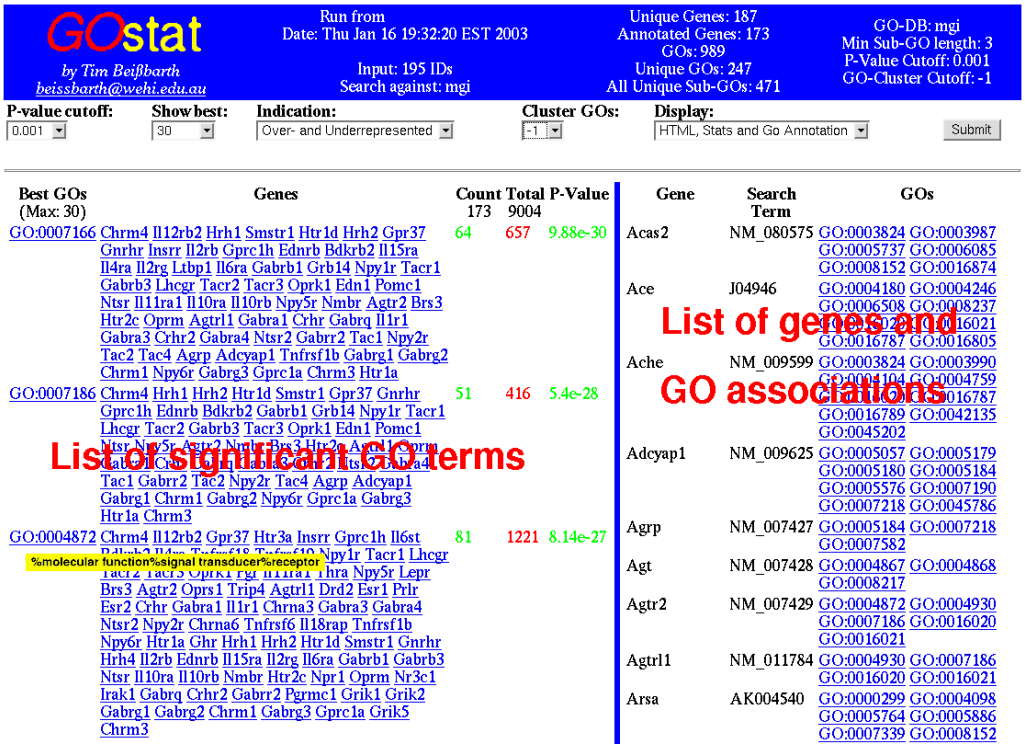


	A	B	C	D	E	F	G	H
1	HYPERLINKED GO CATEGORY	TOTAL GENES	CHANGED GENES	ENRICHMENT	LOG10(p)	CUMULATIVE NUMBER OF CATEGORIES	CUMULATIVE RANDOMS MEAN	FALSE DISCOVERY RATE
2	GO:0006955_immune_response	254	15	3.41	-5.11	1	0.00	0.000
3	GO:0006952_defense_response	271	15	3.20	-4.76	2	0.00	0.000
4	GO:0050874_organismal_physiological_process	332	16	2.79	-4.32	3	0.00	0.000
5	GO:0009607_response_to_biotic_stimulus	295	15	2.94	-4.30	4	0.00	0.000
6	GO:0009613_response_to_pest_pathogen_parasite	144	10	4.01	-3.99	5	0.00	0.000
7	GO:0009605_response_to_external_stimulus	359	16	2.58	-3.89	6	0.00	0.000
8	GO:0042221_response_to_chemical_substance	52	6	6.67	-3.69	7	0.01	0.001
9	GO:0009611_response_to_wounding	91	7	4.45	-3.14	8	0.16	0.020
10	GO:0007186_G-protein_coupled_receptor_protein_signaling_pathway	68	6	5.10	-3.05	9	0.32	0.036
11	GO:0050896_response_to_stimulus	430	16	2.15	-2.96	10	0.34	0.034
12	GO:0006874_calcium_ion_homeostasis	4	2	28.91	-2.77	11	0.61	0.055
13	GO:0006935_chemotaxis	32	4	7.23	-2.71	13	0.70	0.054
14	GO:0042330_taxis	32	4	7.23	-2.71	13	0.70	0.054
15	GO:0006959_humoral_immune_response	54	5	5.35	-2.70	14	0.70	0.050
16	GO:0009628_response_to_abiotic_stimulus	80	6	4.34	-2.68	15	0.70	0.047
17	GO:0006968_cellular_defense_response	33	4	7.01	-2.66	16	0.70	0.044
18	GO:0006805_xenobiotic_metabolism	5	2	23.13	-2.55	17	0.95	0.056
19	GO:0007267_cell-cell_signaling	62	5	4.66	-2.43	18	1.09	0.061
20	GO:0009410_response_to_xenobiotic_stimulus	6	2	19.27	-2.38	19	1.28	0.067
21	GO:0016064_humoral_defense_mechanism_(sensu_Vertebrata)	43	4	5.38	-2.24	20	1.46	0.073
22	GO:0006950_response_to_stress	245	10	2.36	-2.16	21	1.91	0.091
23	GO:0007166_cell_surface_receptor_linked_signal_transduction	171	8	2.70	-2.13	22	1.95	0.089
24	GO:0006875_metal_ion_homeostasis	8	2	14.45	-2.12	24	2.35	0.098
25	GO:0030005_di-__tri-valent_inorganic_cation_homeostasis	8	2	14.45	-2.12	24	2.35	0.098
26	GO:0007155_cell_adhesion	106	6	3.27	-2.07	25	2.42	0.097
27	GO:0006873_cell_ion_homeostasis	9	2	12.85	-2.01	30	2.87	0.096
28	GO:0019884_antigen_presentation__exogenous_antigen	9	2	12.85	-2.01	30	2.87	0.096
29	GO:0019886_antigen_processing__exogenous_antigen_via_MHC_class_II	9	2	12.85	-2.01	30	2.87	0.096
30	GO:0030003_cation_homeostasis	9	2	12.85	-2.01	30	2.87	0.096
31	GO:0050801_ion_homeostasis	9	2	12.85	-2.01	30	2.87	0.096
32	GO:0007178_transmembrane_receptor_protein_serine_threonine_kinase_	10	2	11.56	-1.92	32	3.25	0.102
33	GO:0019935_cyclic-nucleotide-mediated_signaling	10	2	11.56	-1.92	32	3.25	0.102
34	GO:0007218_neuropeptide_signaling_pathway	12	2	9.64	-1.76	33	4.05	0.123

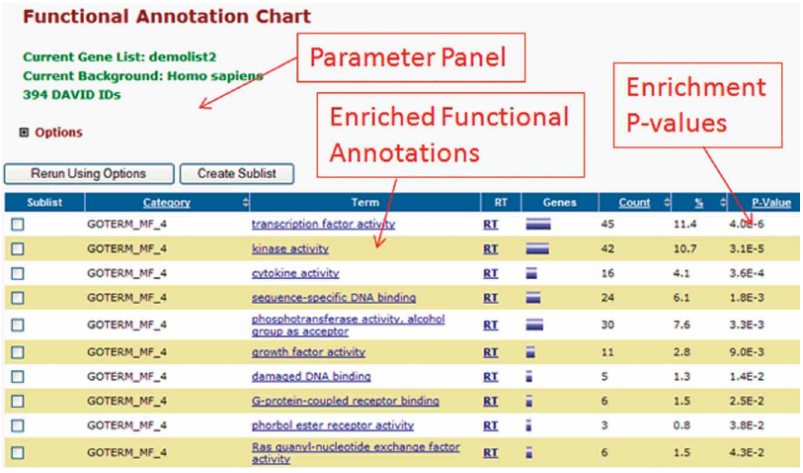
Figura 1.5: Captura de pantalla de una hoja de cálculo, para la salida tabular de SEA cuando se utiliza *GoMiner*. Fuente Zeeberg et al. (2003)

bre asociado en un recuadro de texto amarillo emergente (e.g. "*molecular function, signal transducer-receptor*"). Además se puede inspeccionar a los genes de cada término con vínculos a bases externas o viendo los códigos GO a los cuales se encuentra asociados. Por su parte, *DAVID* presenta una visión similar a la vista tabular (figura 1.6(b)), donde se puede navegar de forma interactiva. No obstante, cada hipervínculo genera una nueva ventana de Internet Explorer®, FireFox® o Chrome®, dificultando así la navegación. Adicionalmente, el usuario debe mantenerse conectado a internet y frente a una inactividad mayor a 5 minutos, se pierde la sesión. Si esto sucede, se deberá repetir el análisis desde la carga de datos.

Se han desarrollado algunas estrategias para mejorar la exploración de los resultados de enriquecimiento de GO. En particular, *GOstat* y el cliente de *GoMiner*, utilizan **árboles jerárquicos desplegables** para poder inspeccionar los diferentes niveles de la estructura de GO como se muestra en la figura 1.7. Si bien esta alternativa permite navegar los resultados, utilizando la propia estructura de GO, no resulta la forma más apropiada, ya que al menos se duplica la información, al romper el grafo dirigido acíclico (GDA). Por ejemplo, a un nodo del cuarto nivel del GDA

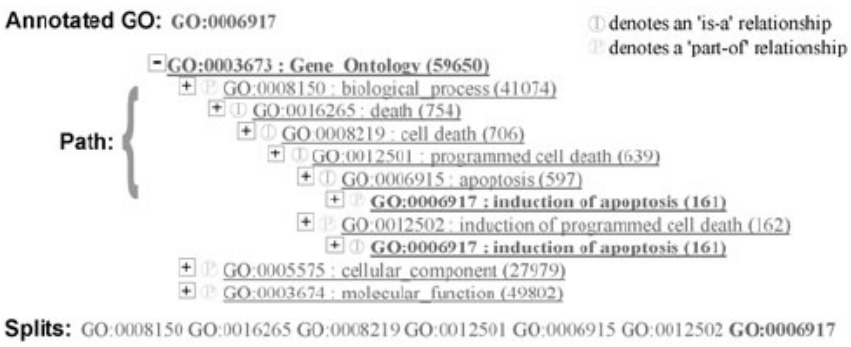


(a) *GOstat*: `gostat.wehi.edu.au`

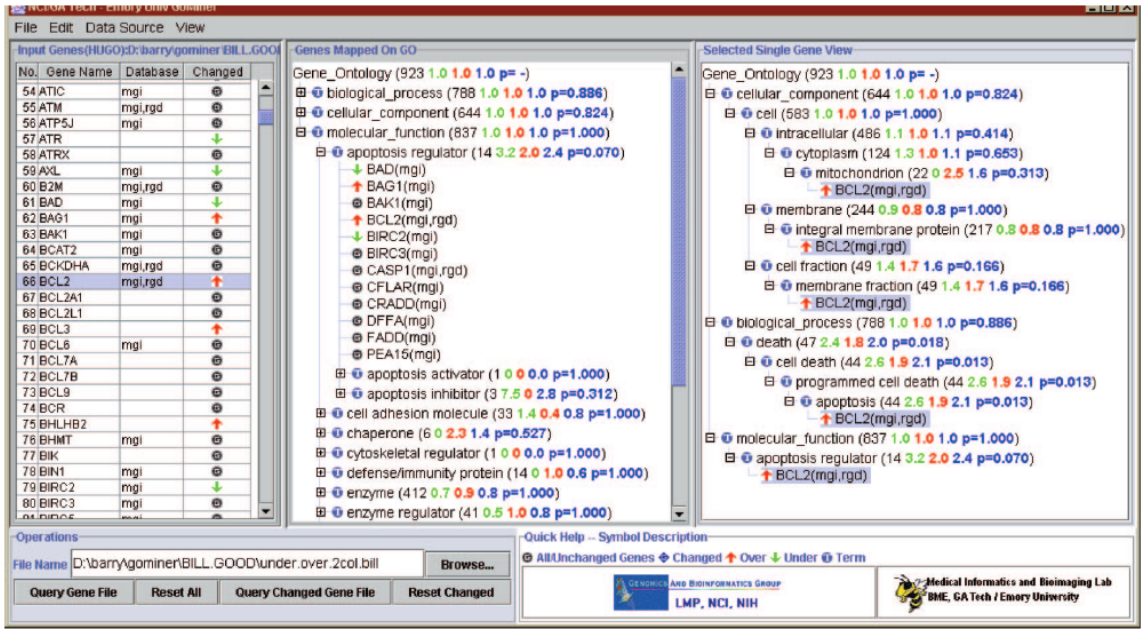


(b) *DAVID*: `david.abcc.ncifcrf.gov`

Figura 1.6: Ejemplos de reportes HTML de SEA para *GOstat* (a) y *DAVID* (b). Imágenes extraídas de Beissbarth y Speed (2004) y Huang et al. (2009b)



(a) Exploración web de *GOstat*: `gostat.wehi.edu.au`



(b) Cliente Java® de *GoMiner*

Figura 1.7: Alternativas de exploración utilizando árboles para representar a Gene Ontology (GO). Notesé que *GoMiner* agrega información de la expresión de un gen con flechas rojas (sobrexpresión), verdes (subexpresión) y circulo gris (sin cambio). Imágenes extraídas de Beissbarth y Speed (2004) y Zeeberg et al. (2003)

que pueda ser accedido por dos caminos (ramas) diferentes, aparecerá en dos ramas desplegables del árbol junto con todos sus descendientes. Esto no genera conflictos conceptuales dadas las relaciones entre nodos de GO, pero aumenta la cantidad de información a la hora de la exploración de los resultados. Cabe destacar que *GoMiner* (figura 1.7(b)), incorpora a cada término el nivel de expresión (sub, igual o

sobre-expresión) respecto de la referencia de cada gen utilizando flechas y colores verde, gris y rojo respectivamente. A su vez, presenta una vista de árbol donde se despliegan los términos en que participa un único gen, como se aprecia en el panel derecho de la figura 1.7(b) para el gen BCL2 en este ejemplo.

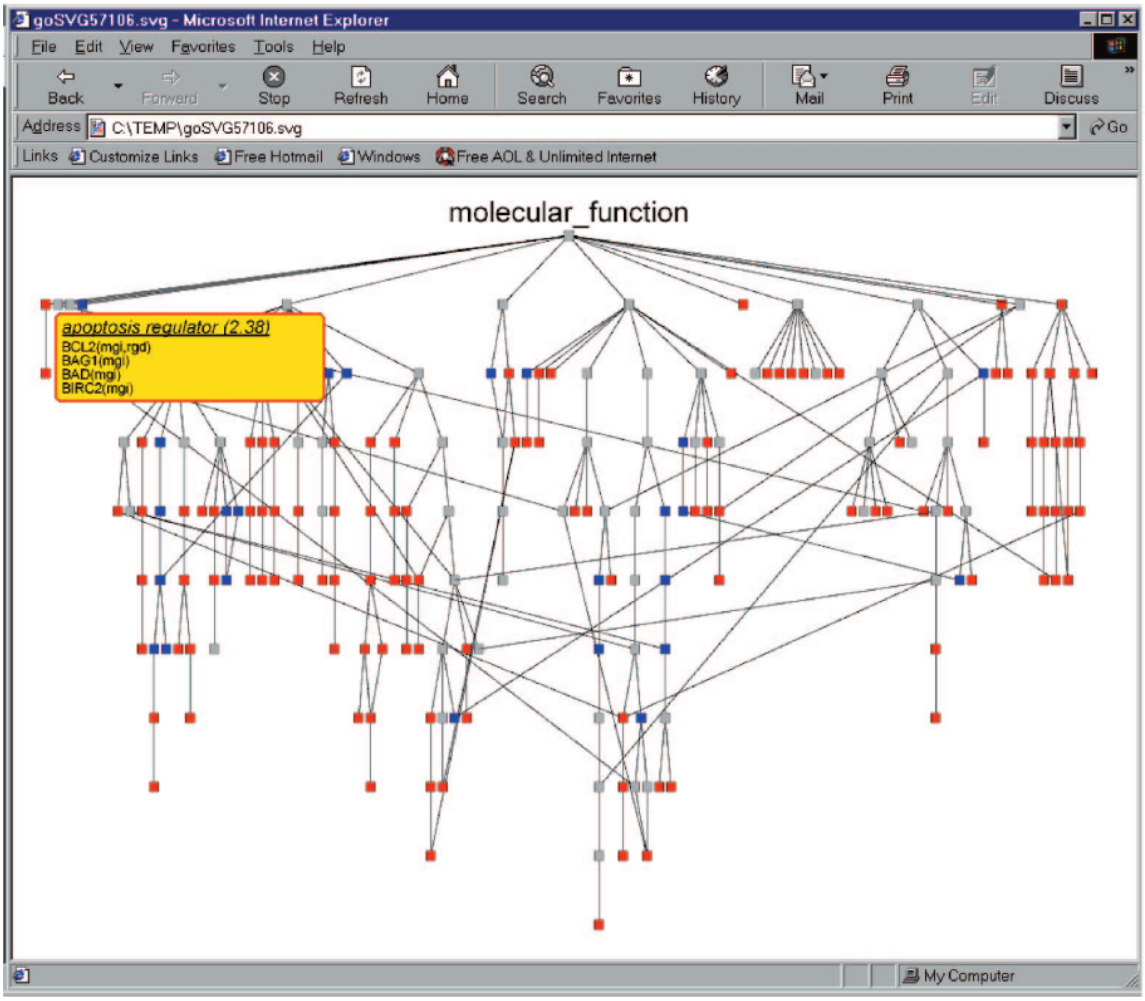


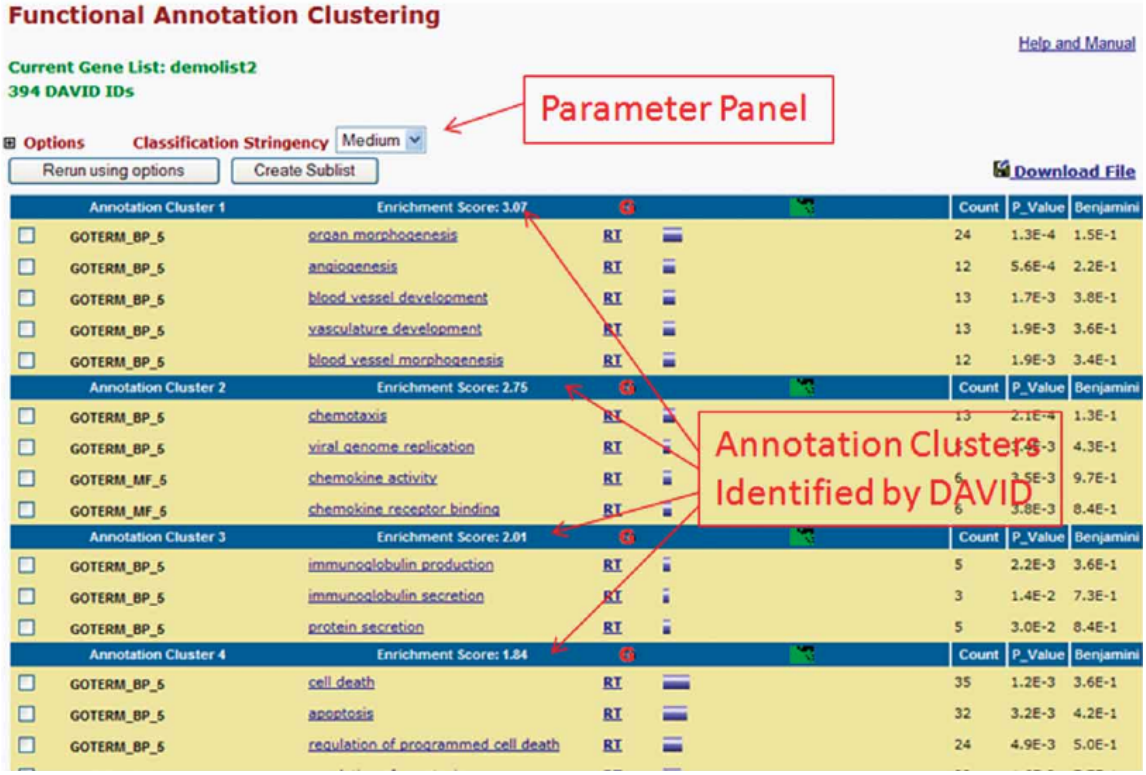
Figura 1.8: Ejemplo de grafo de enriquecimiento de Gene Ontology obtenido por *GoMiner*. Los nodos en color azul/gris/rojo representan términos sub/sin cambio/sobre-enriquecidos respectivamente. Note que al posicionarse sobre un término, un cuadro amarillo emergente indica el nombre del concepto biológico (*apoptosis regulator*) y los símbolos de los genes presentes (*BCL2, BAG1, etc*). Imagen extraída de Zeeberg et al. (2003).

Frente a la duplicación de información que produce este tipo de visualización, *GoMiner* presenta los resultados de enriquecimiento mediante los propios **grafos de GO**. Para ello utiliza gráficos vectoriales redimensionables (SVG, del inglés Scalable Vector Graphics) como se muestra a modo de ejemplo para funciones moleculares en la figura 1.8. No obstante, el cliente posee una limitación en el tamaño máximo de imagen que puede generar, situación que limita la cantidad de nodos que se puedan representar, no siendo posible utilizarla cuando por ejemplo en procesos biológicos se presentan muchos términos enriquecidos (más de 100 nodos). Por su parte, *GOstats* permite utilizar el paquete *Rgraphviz* (Gentry et al., 2013) para visualizar el grafo, pero de una forma muy primitiva. La imagen puede destacar los nodos enriquecidos, al igual que los nombres de términos, mas no posee capacidad adicional para explorar los resultados (inspeccionar los genes asociados a un término, etc). Otra diferencia es que el grafo se encuentra dispuesto de forma invertida, es decir, con el nodo raíz en la parte inferior de la figura.

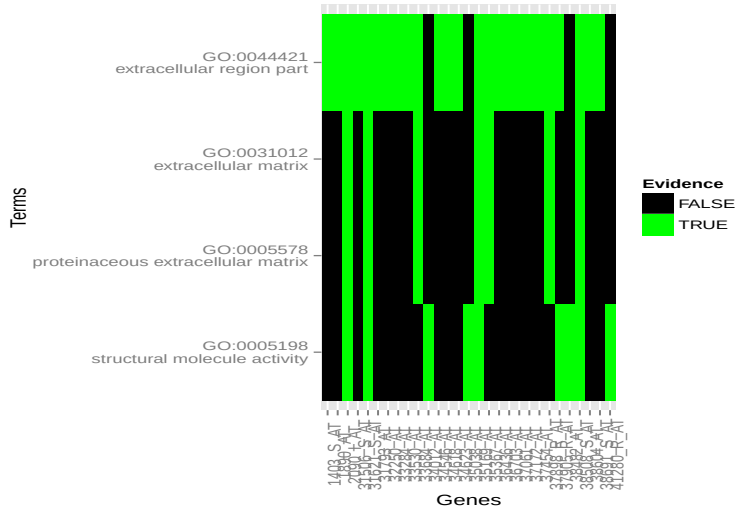
Se han desarrollado otras alternativas para mejorar la exploración de los resultados obtenidos por SEA, relacionados a la longitud de las salidas y visualización de los mismos. Por ejemplo, Al-Shahrour et al. (2004) sugieren recortar el GDA agrupando términos que comparten una cantidad similar de genes. En Zeeberg et al. (2005) proponen utilizar valores q mediante simulación sobre la lista de candidatos, para disminuir la cantidad de términos enriquecidos por azar, mientras que en Huang et al. (2009a) proponen realizar un filtrado más restrictivo por umbral de enriquecimiento (sobre los valores p) y fijar una cantidad mínima de genes candidatos. En cualquiera de los caso anteriores podría perderse información biológica valiosa, por el solo hecho de tratar de facilitar la exploración de los resultados.

En el caso de realizar un análisis del tipo MEA utilizando *DAVID*, los resultados se presentan mediante reportes HTML sobre los agrupamientos realizados sobre genes o términos, como se muestra en la figura 1.9(a). Básicamente es un reporte similar al obtenido al realizar SEA (figura 1.6(b)), donde se presenta un pequeño encabezado para cada agrupamiento. Adicionalmente, ofrece para cada agrupamiento matrices de pertenencia (figura 1.9(b)), donde se puede visualizar en una grilla la evidencia existente (o no) entre los genes y términos de la literatura (KEGG, GO, etc.).

En este contexto, Huang et al. (2009a) aseguran que el análisis de resultados de



(a) Agrupamiento funcional de términos enriquecidos



(b) Evidencia de muchos términos (filas) a muchos genes (columnas)

Figura 1.9: Exploración web de resultados de MEA (del inglés Modular Enrichment Analysis) utilizando *DAVID*. Imagen extraída de (Huang et al., 2009b)

SEA es un proceso exploratorio más que una mera visualización de los resultados estadísticos. No obstante, incluso para un análisis tipo caso-control, el usuario es el único responsable de integrar las extensas tablas o gráficas de salidas obtenidas de la aplicación de diferentes herramientas. De esta manera la propia complejidad de integración de resultados, al igual que la falta de técnicas de resumen visual de información que se pueda realizar sobre ellas, limita la capacidad de análisis. Los anteriores impactan negativamente en la extracción de patrones que pueda realizarse sobre la información disponible, donde la aplicación de técnicas de minería de datos es de gran provecho en este campo, como se muestra es la presente tesis.

Capítulo 2

Minería de datos

En este capítulo se describe brevemente el concepto de "**Minería de Datos**", como un subproceso del análisis de datos, dentro de un contexto mucho más general como el "**Descubrimiento de Información en Bases de Datos**" (DIBD), más comúnmente conocido en inglés como *Knowledge Discovery in Data bases* o **KDD**.

En el contexto del análisis ontológico funcional, el KDD proporciona un marco de referencia *ordenado* de trabajo, *aportando* herramientas y *dirigiendo* el trabajo hacia la *búsqueda* de información relevante. Éste comprende distintas etapas que van desde la conceptualización de los experimentos, la obtención de los datos de entrada, el control de calidad, la adecuación e integración de distintas fuentes de información, el análisis con las herramientas elegidas, hasta la presentación de los resultados mediante informes con visualizaciones apropiadas.

En este capítulo se desarrollan brevemente los conceptos a tener en cuenta en esta metodología y las etapas a seguir para el análisis de datos "ómicos": proteómicos y genómicos.

2.1. Generalidades

Las tecnologías de alto rendimiento permiten explorar proteomas y/o genomas de distintas especies de una sola vez, de manera que es posible medir la expresión de miles de proteínas y/o genes, en forma simultánea. Esto implica que la cantidad de información disponible y accesible hayan sobrepasado, largamente, los *métodos tradi-*

cionales de análisis de datos, los cuales se han vuelto impracticables. Éstos se basan en que un usuario humano (biólogo, investigador, etc.), *manipule directamente los datos*, extrayéndolos y/o realizando búsquedas guiadas por su experiencia o pericia. Si bien las tecnologías de bases de datos proporcionan un almacenamiento eficiente, e incluso un abanico de herramientas para su análisis, la interrelación o interconsulta entre las distintas fuentes de información biológica es compleja y engorrosa. Adicionalmente, las escasas capacidades para visualizar eficientemente el conocimiento encontrado, generan restricciones que limitan los posibles análisis (Huang et al., 2009a). Es por ello que considero que las técnicas de distintas disciplinas, que han dado origen al termino inglés "**Knowledge Discovery in Data bases**" (KDD), son muy apropiadas para abordar este problema.

El concepto KDD empieza a concebirse a finales de la década de los 80, para referirse a un amplio conjunto de procesos. El objetivo principal es *encontrar* o *extraer* información en datos, y enfatizar la *utilización* de un método particular de "**Minería de Datos**" (MD), del inglés Data Mining, en un contexto de más "alto nivel" (Fayyad et al., 1996).

El concepto de MD usualmente se aplica para referirse a un conjunto de técnicas que pueden utilizarse para encontrar *estructuras* y *relaciones* subyacentes, en un conjunto de datos. Las técnicas que se utilizan en MD provienen, en un principio, de campos como la Estadística, el aprendizaje maquinal (del inglés "*Machine-Learning*"), la visualización, la simulación, etc. En este sentido, la MD contiene todos aquellos conceptos que antes referían a *reconocimiento de patrones, clasificación, predicción, agrupamiento*, etc. Todos estos términos son utilizados por distintas disciplinas para describir categorías de problemas de *predicción* y *descripción*, que metodológicamente pueden resolverse de manera similar.

En lo que resta del capítulo se describe brevemente las distintas etapas involucradas en un proceso de KDD. Adicionalmente, se muestra cómo la utilización de los mismos pasos proporcionan una metodología ordenada de análisis, ayudando significativamente a la extracción de información útil y éxito del campo de aplicación: la *genómica* y *proteómica funcional*.

2.1.1. Objetivos

La MD se nutre de técnicas estadísticas, teoría de grafos, árboles de decisión, técnicas de aprendizaje maquinal del inglés "Machine Learning", etc., que son términos y técnicas de procesamiento que han sido desarrolladas en las últimas décadas (Han et al., 2011). Ellas han encontrado una gran variedad de aplicaciones en distintas áreas de la ciencia, la industria y el comercio, intentando resolver dos objetivos de "alto nivel" o generales: la *predicción* y la *descripción*. Estos objetivos pueden alcanzarse mediante la realización de alguna/s de la/s siguiente/s tarea/s:

Agrupamiento: El agrupamiento o "clustering" es una técnica muy utilizada como herramienta descriptiva. Consiste en encontrar/formar una cantidad finita de grupos/categorías que describan a los datos. Estas categorías pueden ser mutuamente excluyentes, o bien una representación jerárquica y solapada de las mismas (Gordon, 2010). Por ejemplo utilizar un mapa de calor (del inglés, *heatmap*) en un experimento tipo caso-control, para verificar si las muestras del mismo tipo se agrupan juntas (Wilkinson y Friendly, 2009).

Clasificación: Consiste en "aprender" o encontrar una función, que proyecta (clasifica) un dato en una o más clases predefinidas. Este puede ser el caso de utilizar firmas moleculares, por ejemplo para asignar subtipos intrínsecos de cáncer de mamas con la PAM50 (Parker et al., 2009).

Regresión: Consiste en "estimar" una función, a la que se le introduce un dato de entrada con el objeto de predecir un nuevo valor numérico. Por ejemplo, el valor de expresión de un gen para un tiempo futuro, en un experimento donde se cuenta con diferentes mediciones en el tiempo (Zuo et al., 2004).

Resumen: Incluye métodos que describen los datos en forma compacta. En este sentido puede mencionarse métodos de estadística descriptiva (valor medio, desviación estándar, etc.) y métodos de visualización como diagramas de cajas (*boxplots* en inglés), gráficos de agrupamiento, grafos, árboles, etc. (Walpole et al., 1999).

Modelos de dependencias: Consiste en encontrar un modelo que describa relaciones significativas entre las variables de análisis. Estos pueden ser algún tipo

de gráfico como los grafos dirigidos acíclicos (sección 1.1.1), o pueden ser tablas que agrupen información contextual relacionada (Peña, 2002).

Detección de cambios y desviaciones: Se centra en la detección de diferencias significativas en los datos, basándose en observaciones pasadas de los mismos. Se puede mencionar, por ejemplo, la detección de proteínas/genes diferenciales en un experimento tipo caso-control utilizando modelos lineales (Graybill, 2000).

En el contexto del *análisis funcional*, la **predicción** implica la utilización de algunas variables, como por ejemplo proteínas y/o genes, o campos de un conjunto de datos de base de datos (identificadores, funciones biológicas, etc.), para la predicción de funciones biológicas que puedan estar modificadas por las variables de interés (ver sección 1.2). Por ejemplo, encontrar una vía de KEGG (sección 1.1.2) donde participen los genes alterados en el experimento, que permitan describir la hipótesis biológica bajo estudio. Por otra parte, la **descripción** se focaliza en encontrar "patrones" o "relaciones" que proporcionen una explicación de los datos, que sea fácilmente interpretable por una persona. Por ejemplo, en un experimento con diferentes niveles para un tratamiento, utilizar un diagrama de Venn para describir los genes que se expresan de forma diferencial entre dichos niveles. También se pueden utilizar algunas medidas de resumen de estadística descriptiva respecto del nivel de expresión de los genes para las comparaciones anteriores.

2.1.2. Etapas

El KDD es un proceso *iterativo* e *interactivo*; consiste en una serie de etapas sucesivas donde, a través de la aplicación de algoritmos particulares, en el sentido de "revolver", "escarbar" sobre los datos en la búsqueda de conocimiento. El abordaje en sí consta de cinco etapas: i) Entendimiento del problema, ii) Entendimiento de datos, iii) Modelado, iv) Evaluación y v) Reportes.

Entendimiento del Problema

En esta etapa se realiza lo que se conoce como **entendimiento del negocio**. En ella el investigador se interioriza, involucra y relaciona en los aspectos del problema a

abordar. Esto permite comprender, vislumbrar la esencia y naturaleza del problema al cual se requiere dar una solución, siendo primordial comprender los conceptos y el vocabulario del **dominio del problema**. También se deben comprender los diferentes procesos que van a proporcionar o han proporcionado los datos, el equipamiento utilizado y los actores (recursos humanos) asociados a cada uno de ellos.

En esta etapa es donde se **definen los objetivos** del proceso de KDD (con sus hipótesis a comprobar), se identifican los actores y se planifican las tareas a realizar. Esto es básicamente lo expuesto en el capitulo 1, donde el dominio de aplicación es el *análisis ontológico funcional*, donde se buscan aquellos procesos/vías que se ven modificados o "*enriquecidos*" por el experimento.

Entendimiento de datos

Esta etapa consiste en **construir una base de datos** *ad hoc*, donde se seleccionan aquellos datos que se asumen puedan aportar información - en función de lo abordado en la etapa anterior. En general, no se trabaja sobre toda la base de datos disponible, ya que ésta puede ser una base de datos de producción y podría no tenerse acceso permanente a ella por razones de seguridad, o bien porque su tamaño hace dificultoso trabajar en tiempos razonables. Usualmente se deberá realizar un muestreo de la base de datos principal. Esto suele ser una tarea tediosa y lenta que implicará la comunicación permanente entre las diferentes partes interesadas, los diseñadores de la base de datos y los usuarios de la misma.

La propia creación de la base de datos posee como tarea inicial la **familiarización de los datos**. Esto implica una serie de actividades como revisar la base de datos, identificar el/los tipos de datos y sus atributos que se considere podrían aportar información del contexto del problema, revisar la integridad de la base de datos y consistencia de sus registros, aplicar transformaciones sobre los datos, etc. Esto permitirá identificar problemas de calidad y descubrir signos iniciales que direccionen las estrategias para detectar información oculta. Para ello se utilizan técnicas de visualización y resumen de datos, de manera tal de obtener una visión global de cómo se comportan y qué tipo de distribuciones tienen. Entre las diferentes tareas es posible particularizar:

- **Creación de un conjunto de datos**: Selección de un conjunto de ejem-

plos/individuos sobre los que se va a realizar el análisis, que se cree aportan información para responder a los objetivos propuestos, o bien son potencialmente útiles para el proceso de descubrimiento.

En el contexto del *análisis funcional*, tomaremos como punto de partida los datos generados a partir de la utilización de tecnologías de alto rendimiento y del conocimiento existente en otras bases de datos. Esto comprende la información correspondiente a los valores de expresión de proteínas o genes, e información de anotación provista por el fabricante (secuencia de oligonucleótidos, identificadores, etc.).

- **Consistencia de datos**: Este concepto se relaciona con que diferentes cosas, pueden estar representadas por el mismo nombre en diferentes sistemas o bien, que atributos que refieren al mismo tipo de observación estén representados con distintos nombres en diferentes sistemas. Esto es especialmente posible cuando se trabaja con diferentes fuentes de información tales como diferentes bases de datos de anotación ontológica (sección 1.1).

  Habitualmente se contextualiza en operaciones entre bases de datos como la unión ("*join*" o "*merge*" en inglés), o la transformación o "mapeo" de identificadores de proteínas o genes de una a otra base. Aquellos identificadores que pertenezcan a ambas bases, representan datos consistentes, es decir, están "mapeados". En caso contrario, no podrán ser utilizados en el análisis posterior.

- **Integridad de datos**: este concepto evalúa las *relaciones permitidas entre los atributos*. Por ejemplo si nuestro dato representa a una proteína o gen, podemos esperar que pueda tener un símbolo (mnemónico) o algunos sinónimos; pero seguramente no podemos esperar que un mismo símbolo se refiera a diferentes proteínas o genes.

  La integridad también está relacionada al *rango aceptable de valores* para un determinado atributo. Por ejemplo en el caso de los niveles de expresión de proteínas y/o genes, siempre se esperan valores positivos de intensidad como medida indirecta de su expresión. El conocimiento del rango de valores permite

evaluar potenciales valores extremos y su posible naturaleza. Los valores extremos deben identificarse siempre, ya que en general requieren un tratamiento especial y su impacto suele ser significativo en las distintas técnicas de MD.

- **Filtrado de los datos**: Operaciones básicas a los efectos de "limpiar" los datos, en el sentido de eliminar aquellas características no deseable: ruido en la señal, calidad insuficiente en los datos, identificación y eliminación de sesgos, gestión de la ausencia de datos y datos atípicos, selección de candidatos de interés, etc.

  En el contexto de tecnologías de alto rendimiento, es habitual utilizar las métricas de calidad de señal que ofrece el fabricante para filtrar los datos y utilizar solo aquellos datos con cierta medida de confiabilidad (Affymetrix (2004), Archer y Reese (2010) Hackstadt y Hess (2009), McClintick y Edenberg (2006) y Bourgon et al. (2010)). Adicionalmente, en análisis funcional, se aplica un filtro para seleccionar aquellas proteínas o genes candidatos que se expresan de forma diferencial entre dos condiciones, usualmente mediante modelos lineales (Graybill, 2000).

- **Reducción, proyección o integración de datos**: Búsqueda de características relevantes, que permitan una mejor representación de los datos para el objetivo propuesto. En este sentido, se reduce la dimensión de la base de datos (en cantidad de variables) para quedarnos con aquéllas que aportan más información o para encontrar invariantes. Un enfoque multivariado clásico es la utilización de técnicas como análisis de componentes principales (**PCA** del inglés Principal Component Analysis, Abdi y Williams (2010)) o regresión por mínimos cuadrados parciales (**PLSR** del inglés Partial Least Squares Regression, Geladi y Kowalski (1986)).

Modelado

A esta etapa también se la refiere en literatura como MD (Orallo et al., 2004). Consiste en la **aplicación de algoritmos** de aprendizaje automático y **uso de técnicas estadísticas** para encontrar patrones en el conjunto de datos previamente seleccionado. De esta manera, los patrones encontrados serán traducidos a *conoci-*

miento que permitan responder a las consignas planteadas en la etapa de "entendimiento del problema". Esta etapa comprende a las siguientes tareas:

- **Elección de la tarea de Minería de Datos**: Decidir qué tipo de tarea es la que vamos a utilizar para alcanzar los objetivos, es decir, si es un proceso de *predicción* o *descripción* (ver sección 2.1.1). Una vez definido, pueden existir una variedad de técnicas que puedan aplicarse, y cada una de ellas podrá tener requerimientos específicos que deberán satisfacerse. El éxito de esta etapa depende fuertemente de la realización adecuada de las etapas anteriores.

  En análisis funcional, la tarea justamente consiste en la predicción del enriquecimiento funcional sobre diferentes procesos y/o vías biológicas (ver sección 1.2).

- **Ejecución de la tarea de Minería de Datos**: Llevar a cabo el análisis propiamente dicho, es decir, aplicar los diferentes algoritmos y modelos computacionales, buscando aquellos patrones que caractericen los datos, etc. En esta tesis, la ejecución de la tarea dependerá de la herramienta seleccionada para enriquecimiento funcional: SEA, GSEA o MEA (ver sección 1.3).

Evaluación

En esta etapa es donde se resalta la **naturaleza iterativa** del KDD. Justamente, con el fin de conseguir la mejor solución posible se evalúan las salidas de los diferentes algoritmos y modelos computacionales propuestos, en función de los criterios u objetivos del problema planteado. Muchas veces es un paso incluido en la etapa anterior, dado que recurrentemente los modelos aplicados y/o propuestos para el descubrimiento de conocimiento deben ser evaluados en pos de encontrar el mejor modelo.

Esta etapa se basa fundamentalmente en técnicas estadísticas de **validación**, con la finalidad de determinar la validez de los patrones encontrados sobre la base de datos. Usualmente se realiza un remuestreo mediante "Bootstrap" (ver más adelante, Orallo et al. (2004)) y control de errores por comparaciones múltiples (FDR, del inglés *False Discorery Rate*, Benjamini y Hochberg (1995)). En esta etapa es fundamental determinar si han habido cuestiones que no hayan sido suficientemente consideradas,

como por ejemplo artefactos identificados *a posteriori* de realizado el análisis, dado que se debe decidir sobre el uso de los resultados obtenidos.

En biología, no son aplicables de forma directa las técnicas de validación utilizadas en modelos computacionales. No existe un "patrón de oro" (del inglés *gold standard*), con el cual se pueda validar el "modelo biológico". Por el contrario, es habitual utilizar una tecnología diferente a la empleada para la obtención de datos, para obtener resultados similares, a los efectos de "validar" lo encontrado. En tecnologías de alto rendimiento la comunidad científica acepta la utilización de la reacción en cadena de la polimerasa en tiempo real (R-T PCR del inglés *Real-Time Polymerase Chain Reaction*, Erlich (1989)), como el estándar. Por otra parte, también es habitual validar utilizando la evidencia existente en la literatura científica o haciendo experimentos complementarios.

Reporte

Una vez que se considera que la etapa de "modelado" representa con precisión al problema, es necesario presentar la solución a todas las partes involucradas: negocios/comercios, investigadores, o bien difundirlas en la comunidad científica que la requiera.

Las características de esta etapa pueden variar en función de los objetivos del proyecto. Estos pueden ir desde "informes de avances" al cumplimentar cada una de las etapas anteriores, que deben contener los resultados y salidas parciales obtenidos en cada una de las etapas, la presentación de gráficos, una asesoría técnica, un informe final, difusión en páginas web, hasta la implementación de una herramienta software. Dentro de este espectro de posibilidades debemos tener en cuenta dos aspectos fundamentales:

- **Visualización**: Diferentes técnicas de mostrar la información, permiten la exploración e interpretación de los resultados. Existen diversas alternativas de presentación de los resultados que comprenden desde informes en formatos tabular, tablas resumen, gráficos, páginas HTML, etc. como los presentados en la sección 1.3.5.

- **Consolidación del conocimiento adquirido**: Incorporar este conocimiento

dentro del sistema, o simplemente documentar y presentar lo realizado a las partes interesadas, mediante un informe o una publicación en alguna revista científica en el ámbito académico.

El hecho de finalizar alguna de las cinco etapas anteriores (Entendimiento del problema, de datos, Modelado, Evaluación y Reporte), no quiere decir que no debamos volver a ella para realizar alguna corrección o ajuste. Justamente, sobre estas etapas y/o tareas se puede iterar (repetir o volver hacia atrás) en cualquier momento que se considere oportuno, dado que la solución del problema no es un proceso lineal.

En el resto del capítulo se desarrollan, con mayor profundidad, cada una de las cinco etapas del KDD, dado que son importantes para el éxito de su aplicación en "Minería de Datos en Análisis Ontológicos-Funcionales".

2.2. Entendimiento del problema

Antes de empezar a trabajar con los datos, hay que definir qué tipo de problema se quiere resolver, es decir, si el problema es de *predicción* o de *descripción*. Es fundamental tener en claro esto, dado que ayudará a la elección de los datos a incluir y las necesidades sobre los datos de salida (si los hubiera). Definir el problema también permite ir acotando las posibles elecciones, sobre la/las técnica/s y herramienta/s de análisis que se van a utilizar.

En esta tesis, el problema "principal" puede enmarcarse dentro de la clase general de problemas de *predicción*. El objetivo primario es la identificación de categorías/términos que puedan estar modificados ("*enriquecidos*") por el experimento y que puedan aportar información biológica relevante (ver capítulo 1). Es un problema muy particular del campo de aplicación específico de las ciencias "ómicas" y requiere de la aplicación de diversas metodologías de MD.

2.3. Entendimiento de datos

Una vez definido el contexto del problema, es posible inferir qué tipo de información se va a necesitar, circunscritos estrictamente al dominio de la aplicación. En este sentido, se puede pensar sobre qué tipo de datos se va a trabajar (numéricos o de

otro tipo), haciendo referencia en este caso a una cuestión meramente metodológica o tecnológica.

También es necesario pensar o evaluar si se van a necesitar datos de salida, cuál es su relación con los datos de entrada, cuántos pasos intermedios hay y cuáles son las entradas-salidas de estos pasos intermedios, cuántos datos de salida vamos a necesitar, cuántos de entrada, si existe la posibilidad o facilidad de poder acceder a dichos datos, si se requerirá información extra, dónde está dicha información, etc.

Este tipo de aspectos son sumamente importantes en una aplicación médica o biológica, dado que puede haber diversos tipos de restricciones. Por ejemplo, si trabajamos con muestras humanas o animales, existen cuestiones éticas involucradas, costos operativos, etc. Por ello debemos ser muy cautos en el diseño de la estrategia y en el proceso de adquisición de los datos, haciendo énfasis en un protocolo que contemple:

I Período de adquisición y tipo de la población, etc.

II Tipo de error asociado a la metodología utilizada y por ende la expectativa de replicados necesarios, limitaciones de los datos (por ejemplo el rango dinámico de la metodología y manipulación de las muestras).

III Otras que puedan surgir por la particularidad de la investigación.

En particular para una investigación en el campo de la medicina, es preciso contar con un protocolo de investigación que tenga en cuenta:

IV Evaluación del problema y protocolo de investigación, si es necesario, evaluado por un comité de ética.

V Condición y disponibilidad de las muestras biológicas.

VI Consentimiento de los sujetos (en caso de ser muestras humanas), o de los dueños de las muestras.

VII Si el proceso de adquisición o extracción de los datos, es un proceso cruento o no, etc., o cuestiones técnicas que puedan alterar o impactar en el análisis.

En el ámbito de la biología experimental, los factores que pueden influir sobre el fenómeno que se desea estudiar pueden ser muy variados. Por lo tanto, es necesario plantear y estudiar concisamente: qué es lo que se desea analizar, cuál es el fenómeno que se quiere ver y cuáles son las variables controladas y a controlar.

En el caso concreto de la proteómica y la genómica, los niveles de expresión de las proteínas/genes pueden estar alteradas por la manipulación de la muestra siguiendo el protocolo del laboratorio húmedo, el proceso de escaneado, los lotes de los reactivos o chips, etc. En este sentido es muy importante acotar algunos parámetros, para que nos permita estandarizar las muestras, el entorno de la variable a observar (en este caso los niveles de expresión), etc. Todas estas decisiones deben ser analizadas e implementadas en el momento de realizar el diseño del protocolo de investigación (cual excede el alcance de esta tesis), los procedimientos operativos de los laboratorios de biología molecular, etc. Sin embargo, es de mucha utilidad evaluar o vislumbrar la existencia de fuentes de variación que puedan provenir de problemas de laboratorio. Usualmente el software propietario utilizado para la obtención de datos ofrece medidas de control de calidad sobre la partida de datos. Estos reportes permiten en primera instancia, comparar los resultados con valores de desempeño esperable por el fabricante. Dicha situación permite la detección temprana de algún artefacto no contemplado en el protocolo de laboratorio.

2.3.1. Creación de un conjunto de datos

En el contexto de las tecnologías de alto rendimiento, la creación de un conjunto de datos se encuentra ligado a las tecnologías de alto rendimiento utilizadas. En particular, abordaremos las dos utilizadas en esta tesis: **electroforesis bidimensional diferencial** (*2D-DIGE*) y **microarreglos de ADN**.

Diferencia en geles de electroforesis bidimensional

Actualmente existen diversas técnicas para el estudio del proteoma. Una de ellas es la utilización de *diferencias en geles de electroforesis bidimensional* para medir diferencias de expresión en proteínas. Esta tecnología es conocida como *2D-DIGE* del inglés "*bidimensional Difference In Gel Electrophoresis*" (Marouga et al., 2005).

La técnica radica en separar las proteínas existentes en la muestra en dos dimensiones de acuerdo con su potencial isoeléctrico y peso molecular. La particularidad de 2D-DIGE es que permite colocar sobre el mismo gel hasta tres muestras distintas de proteínas marcadas con fluoróforos diferentes. Luego en el mismo gel, es posible comparar la abundancia (nivel de expresión) de las proteínas de cada muestra. No obstante, dependiendo del diseño experimental y la técnica estadística utilizada en el análisis es común utilizar más de un gel.

En un experimento tipo caso-control, por ejemplo, se puede realizar una manipulación genética de una línea celular activando o desactivando genes según la/s hipótesis biológica/s que se desea/n investigar. Posteriormente se realiza un cultivo de ellas y se extraen las proteínas de interés: aquellas secretadas, o un extracto intracelular. Las diferentes muestras se etiquetan con diferentes fluoróforos (Cy#) según su procedencia (figura 2.1):

Cy5 - Tratadas: muestras que provienen de una línea celular manipulada.

Cy3 - Control: muestras de la misma línea celular sin manipulación.

Cy2 - Estándar: un preparado que contiene una mezcla de todas las réplicas biológicas tanto *Tratadas* como de *Control*, para poder estandarizar los geles y comparar los niveles de expresión entre los diferentes geles.

Una vez marcadas las muestras (Cy3, Cy5 y Cy2) se mezclan y se colocan en una tira de gradiente de pH inmovilizado, sobre la que se aplica una diferencia de potencial a fin de separar las proteínas de acuerdo a su potencial isoeléctrico (primera dimensión). Luego cada cinta es colocada a lo largo del extremo superior en una pieza de gel de poliacrilamida, usualmente de forma rectangular, donde por el efecto de la gravedad las proteínas se separan de acuerdo a su peso molecular (segunda dimensión). Este proceso se repite para cada gel dependiendo del diseño experimental. Siguiendo con el experimento tipo caso-control, usualmente se emplean al menos cuatro geles (GE, 2008).

Una vez terminada la migración bidimensional de las proteínas, cada gel es escaneado a tres longitudes de onda (una por cada fluoróforo), obteniendo así tres imágenes por cada gel (figura 2.1). Las imágenes contienen manchas (spots), cuya intensidad representa la concentración de proteínas según la ubicación en el gel,

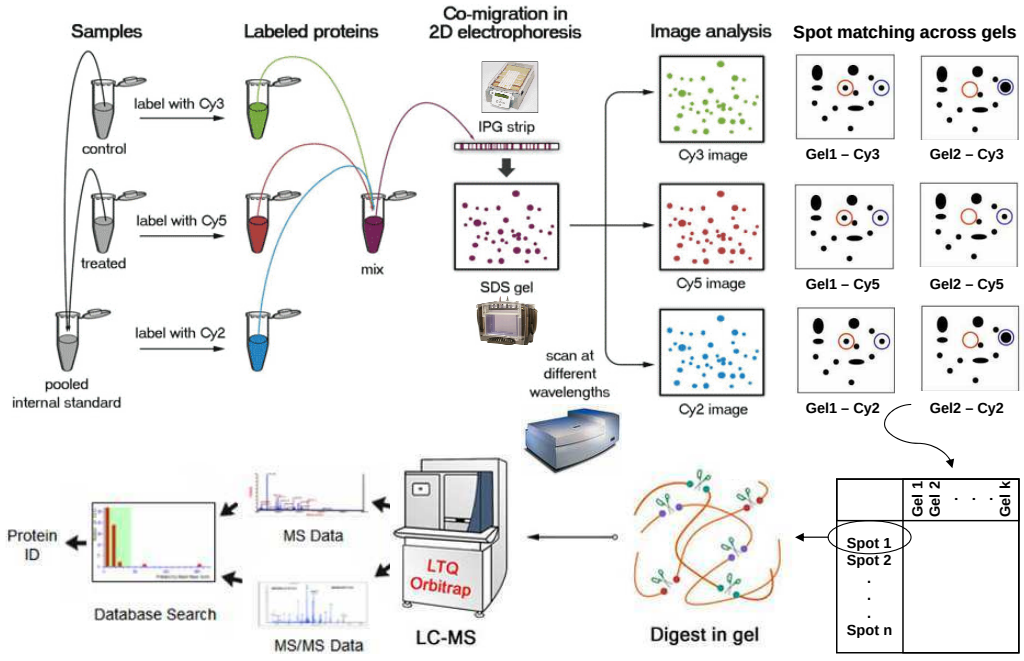


Figura 2.1: Esquema del flujo de trabajo para el análisis de expresión diferencial de proteínas en geles de electroforesis bidimensional (2D-DIGE). Imagen adaptada de www.ciq17.com/yp/web/show.php?userid-39/category-jishufuwu/id-1067.html, www.protein.iastate.edu/q-star.html y GE (2008).

codificada por su potencial isoeléctrico y su peso molecular. En este sentido, la concentración de la/s proteína/s de cada mancha se cuantifica teniendo en cuenta el tamaño de la mancha (en pixeles) y la intensidad de la misma. Así, una mancha más oscura representa mayor concentración (expresión).

Usualmente el fabricante del escáner provee de software para la detección de las manchas, como por ejemplo DeCyder® (GE, 2008). Estos programas realizan una segmentación de la imagen, a los efectos de detectar y cuantificar la abundancia de cada spot en cada gel. Utilizando la información del estándar (Cy2) se aplican diferentes transformaciones sobre las imágenes a los efectos de emparejar (normalizar) las manchas entre los diferentes geles. Una vez finalizado este proceso, se obtiene una matriz de los niveles de expresión (abundancia) de cada spot, en las diferentes combinaciones de tratamientos. Potencialmente cada spot representaría una proteína

que *a priori* se desconoce.

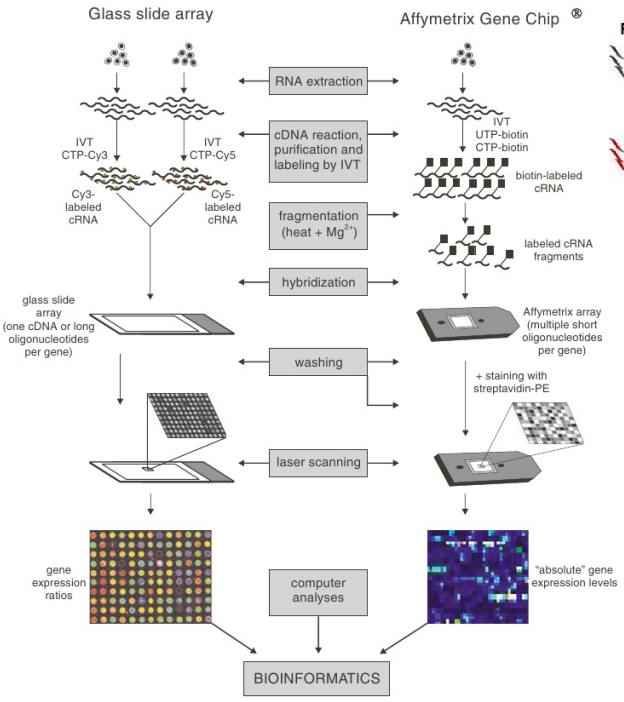
La identificación de las proteínas implica recortar cada uno de los spots, en un gel teñido con una tinción que sea visible a simple vista. Cada spot se corta (digiere) utilizando una enzima que corta la estructura lineal de la/s proteínas presente/s, cada vez que se encuentra una secuencia determinada de aminoácidos (figura 2.1). Estos fragmentos son luego introducidos en un espectrómetro de masa (usualmente para geles Maldi ToF ToF u Orbitrap) y los espectros obtenidos (MS y/o MS/MS) se comparan con resultados teóricos de la digestión de todas los proteínas conocidas utilizando la misma enzima. Así, es finalmente posible obtener el/los identificador/es de la/s proteína/s (ID) presentes en cada spot. De esta manera, se cuenta tanto con la matriz de expresión como de la información de anotación de las proteínas. Esta información es el punto de partida del KDD, al utilizar este tipo de tecnología de alto rendimiento.

Microarreglos de ADN

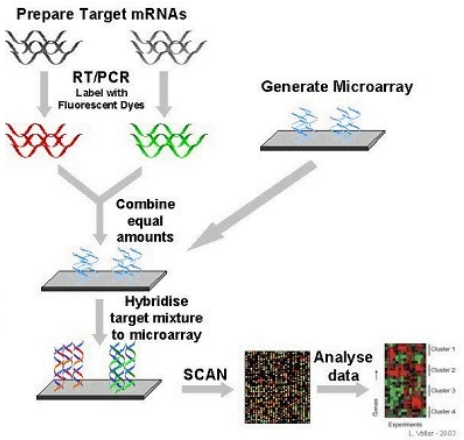
La información contenida en un organismo se encuentra almacenada en el genoma en forma de moléculas en ADN. No obstante, el ADN debe ser transcripto a ARN mensajero (ARNm, *transcriptoma*), el cual es traducido a proteínas (*proteoma*) siendo éstas últimas las efectoras de las diferentes funciones biológicas que sostienen el funcionamiento.

Los microarreglos (del inglés *microarrays*) de ADN, son usados para obtener el *perfil de expresión* de una célula a nivel del transcriptoma. Esto permite indagar en diferentes contextos experimentales, los mecanismos de regulación, vías metabólicas y funciones celulares asociadas. Para ello se monitorea la expresión de miles de transcriptos simultáneamente (López et al., 2005).

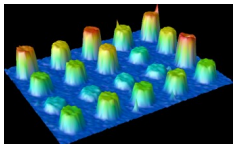
Básicamente, un microarreglo es una colección de oligonucleótidos o fragmentos de ADNc conocidos (sintetizados a partir de ARNm), dispuestos sobre una superficie sólida (chip) en forma de grilla o arreglo (Tarca et al., 2006). Estos fragmentos se denominan *sondas* y se emplean para identificar secuencias complementarias a ellas, provenientes de la muestra (figura 2.2). Tecnológicamente para chips Affymetrix®, para medir la expresión de un transcripto es necesario utilizar la información de un conjunto de sondas, dado que no es posible sintetizar la secuencia completa de ADNc



(a) Comparación de tecnologías



(b) Hibridización con dos colores



(c) Intensidades

Figura 2.2: Esquema de la obtención de datos en tecnologías de microarreglos. A) Comparación de tecnologías de uno y dos colores. B) Esquema de hibridización para el caso de dos colores. C) Esquema idealizado de la representación de la intensidad en una imagen. Imágenes obtenidas de Staal et al. (2003) y `www.microarray.lu/en/MICROARRAY_Overview.shtml`.

en el chip. A su vez, existen dos tipos de tecnologías, dependiendo de la cantidad de fluoróforos diferentes que se puede aplicar sobre un mismo chip: los de dos colores y los de un color. En la figura 2.2(a) se comparan las diferencias existentes en la preparación de las muestras para la obtención de resultados.

En cualquiera de las tecnologías, el principio de funcionamiento se basa en la complementariedad de las secuencias. Es decir, las secuencias de la muestra se unen (hibridizan), con aquellas inmovilizadas en el chip que complementan a sus bases (ver figura 2.2(b)). Una vez terminado este proceso, los chips son leídos por un escáner utilizando la longitud de onda de los fluoróforos. Así, la señal emitida por el fluoróforo es proporcional a la cantidad del marcador presente, lo que permite realizar una cuantificación indirecta a partir de la/s imagen/es escaneada/s. En el

caso de dos colores, se cuantifica la expresión de cada fluoróforo y es habitual utilizar la intensidad relativa, mientras que en los de un color, se usa la expresión absoluta. Este proceso da como resultado una grilla de intensidades como los de la figura 2.2(c). Luego los datos de cada chip se disponen en columnas en una matriz de expresión "cruda", donde en este caso por columnas tendremos chips (tratamientos) y por filas genes. No obstante dependiendo de la plataforma tecnológica, como por ejemplo Affymetrix®, la información de expresión de los transcriptos se obtiene mediante un procesamiento que consiste en la corrección del ruido de fondo y resumen de información de oligonucleótidos de un mismo transcripto (Gentleman et al., 2005).

La corrección del ruido de fondo permite disminuir la variabilidad que se produce debido a la hibridación no específica y producto del sistema de detección óptico. A su vez, como un transcripto es representado en el chip por un conjunto de sondas, dicha información es resumida en un único valor de expresión para cada trasncripto. Incluso existen chips que poseen sondas repetidas del mismo transcripto para aumentar la confiabilidad del nivel de expresión. De esta manera, se cuenta con la matriz de expresión "resumida", similar a la obtenida en geles de proteínas (sección 2.3.1), que junto con la información de anotación, conforman los datos obtenidos por esta tecnología.

2.3.2. Consistencia e integridad de información

Las tecnologías de alto rendimiento generan bases de datos de elevada dimensión y estructura. Sin embargo, los datos de estas salidas se encuentran lejos de transformarse en información útil para el investigador. Para ello, es necesario poder *relacionar* dichos datos con conocimiento previamente adquirido por la comunidad científica. Justamente aquí es donde entra en juego el concepto de **consistencia e integridad** sobre la información contenida en **bases de datos de anotación**.

Bases de datos de anotación

La información de anotación almacenada en estas bases de datos comprende el conocimiento existente. Diferentes organizaciones y consorcios que mantienen bases de datos biológicas con diversa intencionalidad. Justamente, en función de ello será el tipo de dato que se almacena. Algunas son específicas para un organismo como

por ejemplo **FlyBase** para *Drosophila* (moscas, FlyBase Consortium (1994)) o la base de datos para el proteoma de levaduras (**YPD**, *Yeast Proteome Database*) para *Saccharomyces cerevisiae* (Hodges et al., 1999) o para compuestos y reacciones químicas como **PubChem** (Bolton et al., 2008). Incluso existen algunas que pretenden incorporar información de diferentes organismos a nivel de genes como en **Entrez Gene** (Maglott et al., 2011), proteínas en **UniProt** (Apweiler et al., 2004) y **PIR** (Wu et al., 2002), incluso vías metabólicas como en **Reactome** (Joshi-Tope et al., 2005) y **KEGG** (ver, sección 1.1.2), o vocabulario controlado como en **GO** (ver, sección 1.1.1).

Cada uno de estos repositorios almacena la información mediante bases de datos relacionales con un esquema propietario. En el mejor de los casos, estas bases de datos son de libre acceso, tanto al esquema como a los datos. No obstante, el desafío es poder relacionar todo ese conocimiento para cada secuencia, proteína o gen que brinda la tecnología de alto rendimiento utilizada.

En el caso de *2D-DIGE*, la identificación es un proceso *a posterior* del análisis de los datos, como vimos en la sección 2.3. Mientras que al utilizar *microarreglos* el fabricante usualmente provee un archivo de anotación *a priori*, dado que sabe qué secuencia de ADNc u oligonucleótidos se encuentra en cada posición de la grilla. A su vez, los fabricantes ofrecen diferentes archivos de anotación, donde utilizan identificadores propietarios para relacionarlos con las diferentes bases de datos. Si bien las diferentes tecnologías proveen de software propietario para analizar los datos, existe una serie de problemas de consistencia e integridad cuando el investigador pretende realizar análisis acorde a sus necesidades.

Inconvenientes en la utilización de la anotación

En el contexto de análisis de datos de alto rendimiento, es habitual utilizar simultáneamente diferentes herramientas bioinformáticas. Cada una de ellas fueron desarrolladas para un propósito específico, presentando diferentes características/fortalezas. En algunos casos, funcionan de forma cerrada (cajas negras), no permitiendo acceder a los resultados de etapas intermedias o extender sus funcionalidades. Este suele ser el caso de los software propietarios asociado al equipo que obtiene los datos. Por otra parte, es usual que las diferentes herramientas se encuen-

tren ligadas a un tipo de identificador (ID) particular. Aquí nos encontramos frente a un problema de consistencia, donde es necesario realizar una **conversión de ID** para poder utilizar dichas herramientas.

En sí misma la conversión no es un proceso trivial, dado que depende de muchos factores como las estructuras de almacenamiento, cómo se estable la relación entre bases de datos, cuáles son las versiones compatibles, etc. Desafortunadamente la mayoría de las herramientas disponibles no documentan cómo realizan dicho proceso. Si bien existen conversores que dicen soportar muchos tipos de IDs, ello no implica que establezcan de forma adecuada la relación entre la base de datos origen y la de destino. Todas estas particularidades hay que tener en cuenta a la hora de realizar la conversión de IDs. En este contexto Huang et al. (2009a) recomiendan utilizar Onto-Translate (Draghici et al., 2003), MatchMiner (Bussey et al., 2003), IDConverter (Alibés et al., 2007) y DAVIDIDConverter (Huang et al., 2007). Lamentablemente los IDs que no logren ser emparejados se pierden, lo cual introduce sesgo sobre la información biológica disponible para el análisis. Consecuentemente, algunas proteínas/genes pueden no participar del análisis, a pesar de que pueden tener un rol crucial en el contexto del experimento.

Otro problema de consistencia se encuentra directamente asociado al **tipo** de ID utilizado en el análisis (Zeeberg et al., 2004). El **símbolo** es uno de los más utilizados y consiste de un mnemónico corto, empleado para referirse a la proteína o gen en cuestión. Por ejemplo, en *Homo sapiens* para el gen "septin 9", se utiliza el símbolo "SEPT9". Si bien "SEPT9" resulta más inteligible para el investigador que su contra parte en Entrez Gene ID, "10801", cuando es utilizado en programas tipo Microsoft Excel® es transformado a un dato de tipo fecha "Sep-09". Adicionalmente, los símbolos no siguen una convención unificada para su denominación, como por ejemplo, utilizar parte de la descripción del gen. De hecho, en muchos casos son acrónimos del apellido del investigador que reportó su aparición. Incluso pueden ser referido en publicaciones contemporáneas con distintos símbolos, dando lugar a los "**alias**" o "**sinónimos**" conocidos del gen. Siguiendo con el ejemplo de "SEPT9", a la fecha cuenta con los siguientes sinónimos: MSF, MSF1, NAPB, SINT1, PNUTL4, SeptD1 y AF17q25. En otros casos, más de un gen diferente puede compartir el mismo símbolo. Por ejemplo, el símbolo "ANXA8" refiere a tres genes cuyos símbolos

oficiales a la fecha son ANXA8, ANXA8L1 y ANXA8L2, siendo que son tres entidades diferentes. Justamente, debido a todos estos problemas de consistencia e integridad de datos, el comité para la nomenclatura de genes humanos (**HGNC**, de sus siglas en inglés, Povey et al. (2001)) realiza un gran esfuerzo por normalizar la nomenclatura de los símbolos. Esta situación da lugar a un nuevo problema, la **estabilidad** de los IDs.

La estabilidad usualmente es un problema subestimado, siendo que no es uno menor. La utilización de los símbolos es, sin lugar a dudas, el caso más crítico (Zeeberg et al., 2004), dado que tanto el HGNC como otros consorcios los modifican en forma recurrente. Esto impacta en la imposibilidad de **reproducción de los resultados** publicados en un artículo científico. Sumado a lo anterior, las herramientas disponibles usualmente no publican la **versión de base de datos** que poseen instalada, siendo común la imposibilidad de su utilización por diferencias en la versión utilizada. Por otra parte, el mayor problema es la **falta de trazabilidad** de los IDs dentro de una misma base de datos. En este aspecto, algunos consorcios solo mantienen el registro actual de las proteínas/genes, no permitiendo acceder a las versiones anteriores; o incluso desde la versión anterior no es posible acceder al registro actual, perdiendo así su trazabilidad. De esta manera, el investigador no sabe si: i) existe evidencia de que dicho ID en realidad no tiene una función biológica, ii) se unificó con otro identificador o iii) ha sido actualizado por uno nuevo.

Todos los problemas anteriores de consistencia e integridad hacen que la anotación posea en sí misma gran complejidad. A ellos hay que sumarles el agravante de que, dependiendo de las bases de datos de anotación utilizada, la conversión o la propia trazabilidad de un único identificador debe realizarse de **forma manual**. Esta situación es impracticable si se considera el caudal de datos que generan las tecnologías de alto rendimiento. Es decir, no existe un acceso programático (sin supervisión del usuario) para realizar las diferentes tareas requeridas sobre las decenas a centenas de miles de IDs involucrados en un experimento.

2.3.3. Filtrado de datos

Una vez que se cuenta con los datos para el estudio, se procede a **preparar los datos**. El filtrado de datos es fundamental en el desarrollo de un proceso de MD,

siendo una etapa significativa y a veces crítica para el éxito de la aplicación (Orallo et al., 2004). El objetivo de esta etapa es simplificar el proceso de análisis (modelado o reconocimiento de patrones), enfatizando la información relevante del sistema bajo estudio. Para ello se trabaja en la reducción del ruido y/o eliminación de datos inconsistentes, dado que ambos casos pueden oscurecer la información subyacente en los datos, causando confusión. Adicionalmente, se hace uso del conocimiento previo aportado por los expertos o por el analista, poniendo especial cuidado en que este conocimiento no produzca desviaciones en el análisis, de manera que los resultados no resulten sesgados por los datos elegidos y no reflejen la información contenida.

Anotación para microarreglos

Únicamente para el caso de microarreglos, los archivos de anotación del fabricante poseen información adicional para cada sonda, como es el caso de utilizar chips de Affymetrix®. Dentro de la diversidad de información, el fabricante especifica características adicionales para cada sonda como *tipo de sonda* y *características de la secuencia* (Affymetrix, 2004). Esta información es de utilidad a la hora eliminar datos que puedan oscurecer la etapa de modelado (aumenta variabilidad, no cumplen supuestos del modelo, etc).

En lo que respecta al **tipo de sonda**, este campo define la intencionalidad de la misma. En este contexto, el fabricante especifica cuáles sondas son:

- **Principales**: diseñadas específicamente para el organismo bajo estudio. Estas sondas se encuentran identificadas como "*main*" (principal en inglés).
- **Controles**: sirven para diferentes tipos de control. Entre ellos se encuentran los controles propios del fabricante (denominados "*control->affx*"), aquellas específicas del microarreglo ("*control->chip*"), las utilizadas para cuantificar el ruido de fondo del mismo organismo ("*control->bgp->genomic*") o de uno diferente ("*control->bgp->antigenomic*")
- **Normalizadores**: son utilizadas para homogeneizar la intensidad de la imagen, a nivel de exones ("*normgene->exon*") o intrones ("*normgene->intron*").
- **De rescate**: sondas de micro ARN que no alinean con el genoma del organismo, o lo hacen de forma muy limitada ("*rescue->FLmRNA->unmapped*").

En este contexto, deben eliminarse aquellas sondas que no codifican las características principales, es decir, todas aquellas utilizadas por el fabricante (controles, normalizadoras y de rescate). A su vez, las sondas principales poseen en su ID una **terminación que codifica** diferentes características de la secuencia, siendo las más comunes (dependiendo del chip):

- **_at**: sondas que alinean con un transcripto conocido.
- **_a_at**: sondas que alinean con un transcripto alternativo para el mismo gen.
- **_s_at**: aquellas sondas que alinean con múltiples transcriptos de diferentes genes (hibridación cruzada).
- **_x_at**: sondas donde no fue posible seleccionar una única secuencia o un conjunto de ellas con idénticas secuencias entre múltiples transcriptos.

De esta manera, valores de expresión proveniente únicamente de sondas del tipo "_at" y/o "_a_at" representan datos no ambiguos, es decir, provienen de un único transcripto, razón por la cual son los utilizados en esta tesis. Si bien los datos de las sondas "_s_at" son de utilidad, requieren un tratamiento especial para su posterior validación, dado que se debe diseñar un experimento adicional para determinar a cuál/es de todos los trasncriptos diferentes pertenece su nivel de expresión. Es por ello que no se incluyen en el análisis, al igual que las sondas terminadas en "_x_at".

Este filtro utiliza el conocimiento previo, aportado por los expertos, en el sentido de que se conoce *a priori* la posición y la secuencia específica de cada sonda. Justamente, elimina aquellas sondas que puedan introducir desviaciones en el análisis al utilizarlas en un contexto diferente para el cual fueron diseñadas. Sin embargo, también es necesario enfatizar la propia información obtenida del sistema bajo estudio. Para ello se recurre a un filtrado utilizando información del nivel de expresión (señal), adquirido por la tecnología de alto rendimiento.

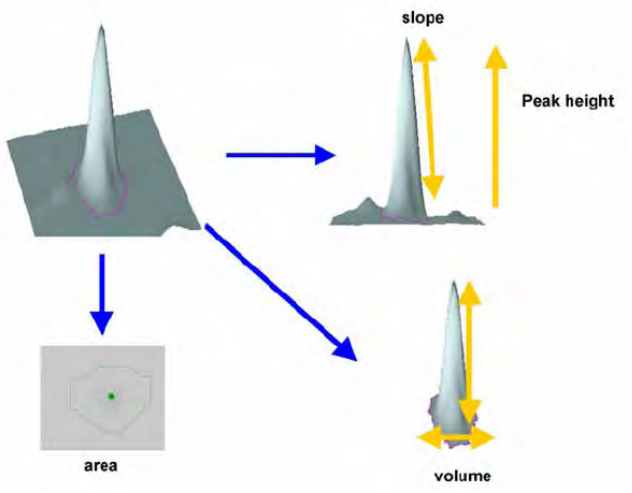
Control de calidad de la señal

Este filtro es necesario y de vital importancia, dado que los datos provenientes del mundo real no son ideales. Especialmente al trabajar con datos biológicos, resulta

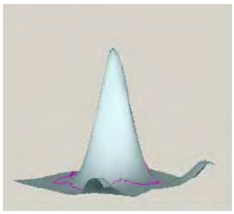
muy común encontrar datos incompatibles que dificultan la integración de distintas fuentes de datos. Por lo tanto, revisar la calidad de la señal es un paso obligatorio y conveniente (Batista y Monard, 2003; Zhang et al., 2003).

En esta etapa se realiza un proceso de revisión y limpieza de los niveles de expresión de proteínas/genes donde se preparan y seleccionan los datos para su posterior análisis. En este contexto se verifican diferentes aspectos como: que los datos sean correctos, la ausencia de datos (del inglés *missing values*), sus características distribucionales, la presencia de datos con una magnitud no esperada (anómalos o del inglés *outliers*), la existencia de datos de distinta naturaleza (numéricos, booleanos y/o tipo carácter), etc.

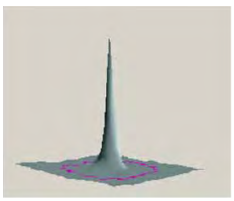
En el caso de utilizar microarreglos de ADN, los fabricantes proveen de ciertas **medidas de detectabilidad**, sobre los niveles de expresión como presencia/ausencia, referidos como "calls" en tecnología Affymetrixs®, o "banderas de calidad" en Agilent®, Heebo®, o cualquier otra técnicas de dos colores. Estas medidas por lo general ofrecen un rango de *valor p*, para el cual se definen regiones de confiabilidad donde el valor de la señal es: aceptable (presente), marginal o ausente (Affymetrix, 2004).



(a) Parámetros de una mancha



(b) Proteína



(c) Partícula de polvo

Figura 2.3: Caracterización de una mancha (a) y ejemplos de: una proteína (b) y una partícula de polvo (c). Imágenes obtenidas de GE (2008).

Se ha demostrado que el filtrado de los genes utilizando este tipo de mediciones, no modifica la distribución marginal de los genes que se expresan de forma diferencial. Por el contrario, el no filtrado afecta fuertemente a la normalización e incrementa la tasa de falsos positivos (Affymetrix, 2004; Archer y Reese, 2010; Bourgon et al., 2010; Hackstadt y Hess, 2009; McClintick y Edenberg, 2006), razón por la cual en esta tesis se adoptó considerar aquellos valores que no poseen confiabilidad en la medición, como valores ausentes. Es decir, aquellos que se encuentran en el nivel del ruido. En este sentido, los investigadores deben utilizar sólo aquellos genes que "sistemáticamente" están presentes en el estudio.

Por el contrario, al trabajar con niveles de abundancia de proteínas al utilizar 2D-DIGE, el panorama es diferente. En esta tecnología, los software provistos por los fabricantes para la cuantificación de las manchas, usualmente, no entregan información de detectabilidad de la forma que lo hacen para el caso de los microarreglos. No se conoce *a priori* cuantos spots hay, ni su ubicación, o incluso cuál/es proteínas son. Sin embargo, las manchas se encuentran caracterizadas por los parámetros que se muestran en la figura 2.3(a). Esta información es de utilidad para **discriminar proteínas** (figura 2.3(b)) de manchas que puedan ser producto de una partícula de polvo (figura 2.3(c)), o aquellas que han saturado el rango dinámico del escáner. Para ello, GE (2008) recomienda utilizar un filtro basado en los parámetros de la tabla 2.1 para eliminar las manchas no fiables.

Tabla 2.1: Parámetros recomendados para filtrar manchas con artefactos

Propiedad	**Valor sugerido**	**Unidad**
Pendiente	>1.1	Intensidad / píxel
Área	<100	Cantidad de píxeles
Volumen	<10000	Cant. píxeles x intensidad
Altura de pico	$<80 \wedge >65000$	Intensidad

Valores por defecto utilizados en DeCyder®. Fuente GE (2008).

La aplicación de este tipo de filtro, permite eliminar artefactos y valores no confiables. En este contexto, la no aplicación del mismo, puede no remover diferentes fuentes de variabilidad, que pueden ser perjudicial en la etapa de modelado.

Normalización

El proceso de normalización es otro aspecto importante en el entendimiento de datos. Básicamente implica llevar a todas las variables de entrada a un mismo rango de trabajo. De esta manera, se intenta evitar que alguna dimensión particular (variable) domine sobre las otras. En general la normalización es un proceso de escalado y en muchos algoritmos de MD mejoran sustancialmente su rendimiento, eficiencia y/o interpretación (Orallo et al., 2004).

En el caso particular de datos de expresión de proteínas y/o genes, existen diferentes fuentes de variabilidad que pueden influir sobre la medición de los niveles de expresión. En este contexto, la normalización no sólo se aplica en el sentido habitual, es decir, llevar las variables a **escalas comparables**, sino que también para **eliminar efectos técnicos**, que no tienen que ver con la biología que se está estudiando. Estas fuentes de variación se conocen como *variabilidad tecnológica*.

La experiencia ha demostrado que existe una cantidad sustancial de fuentes de variabilidad tecnológica en datos de microarreglos y geles de proteínas. La calidad de los datos puede estar influenciada por todos y cada uno de los pasos que preceden al análisis, desde la extracción y marcación de la muestra, condiciones de hibridización, adquisición de la imagen, e incluso pequeñas imperfecciones de fabricación en los chips o geles utilizados para la obtención de datos. Por lo tanto se deben inspeccionar los datos por posibles inter e intra artefactos en los microarreglos o geles de proteínas. Algunas de estas fuentes de variación pueden provocar desviaciones sistemáticas en la medición, variaciones que pueden ser estimadas y corregidas mediante técnicas de normalización (Quackenbush, 2002). Si no son eliminadas, estas variaciones sistemáticas aumentan la incidencia de falsos positivos durante el análisis (Nadon y Shoemaker, 2002).

Uno de los supuestos de este tipo de experimentos es que la gran mayoría de las proteínas/genes presentes en el modelo biológico, no se van a ver afectados por el tratamiento. Es decir, que el nivel de expresión esperable por la mayoría de ellos debe ser basal (a nivel de una referencia) y solo algunos de ellos tendrán un valor de expresión influenciado por el experimento (valores mayores o menores que la referencia). No obstante, al considerar la distribución de las intensidades registradas por el medio óptico (valores positivos), poseen valores muy pequeños y morfología

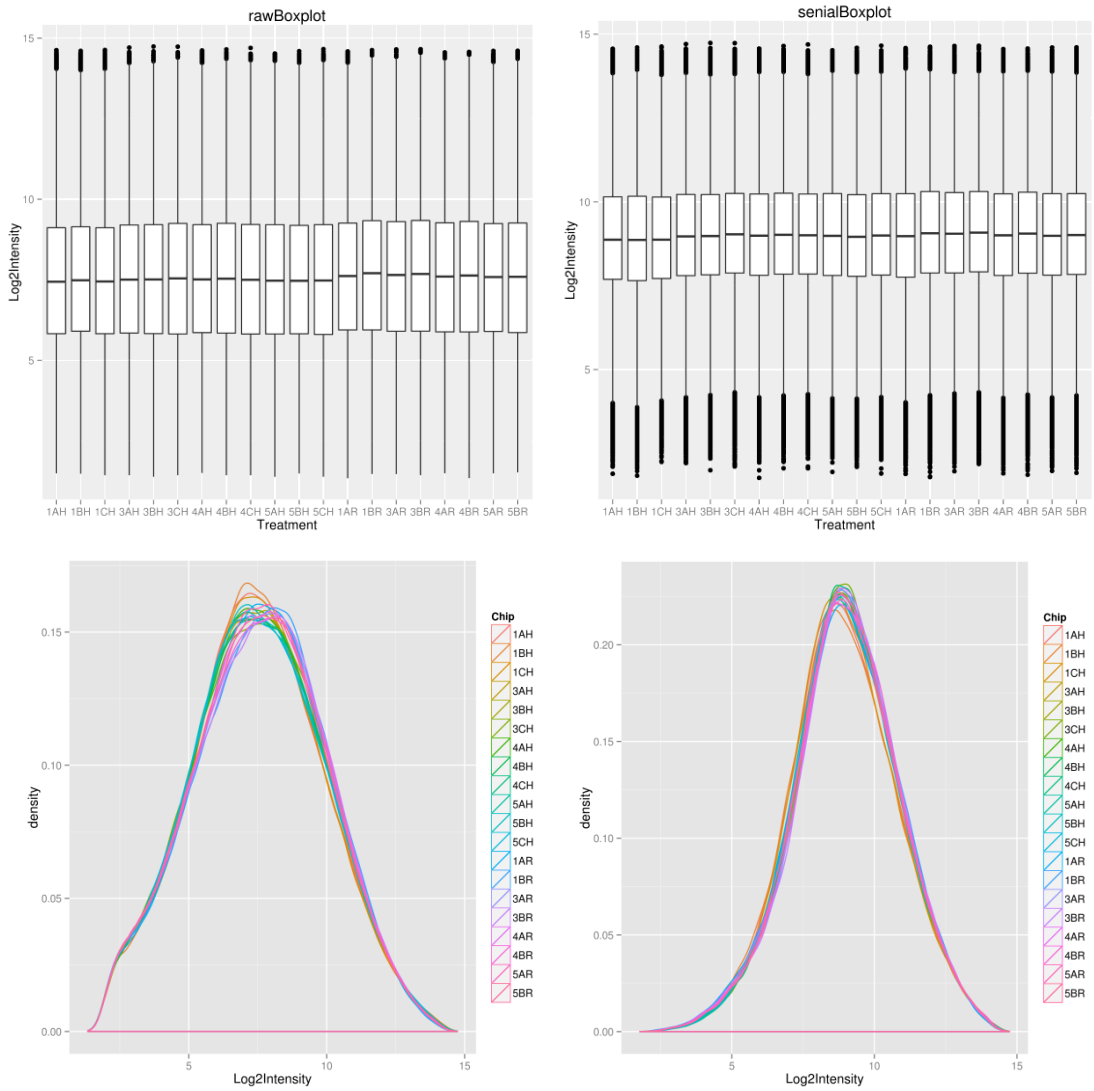


Figura 2.4: Normalización de muestras de microarreglos de Affymetrix®. Nivel de expresión medido en escala logarítima antes y después de normalizar por cuantiles, en el panel izquierdo y derecho respectivamente. En el panel superior se muestran los diagramas de caja y en el inferior, la función de densidad de cada muestra.

muy asimétrica, razón por la cual es usual aplicarles una transformación logarítmica a cada muestra, para cambiar el rango de la variable y corregir en parte la asimetría, como se aprecia en los diagramas de caja y densidad de la figura 2.4. Sin embargo, existen casos como el presentado en el panel izquierdo, donde se aprecia que existen pequeñas diferencias en las funciones de densidades, cuando no son esperables desde la biología y son producidas por la variabilidad tecnológica. Es decir, que si comparamos individuos entre las distintas muestras podemos encontrar diferencias en el nivel de expresión sólo por variación en la técnica. En estos casos es necesario "*normalizar*" los datos.

En base al supuesto biológico, inicialmente los datos de las distintas muestras (intensidades en escala logarítmica) eran normalizados por "**escala**". Simplemente a todos los valores de cada muestra se le restaba su propia mediana y dividía por la desviación estándar (Smyth y Speed, 2003). De esta manera, se corregían las posibles diferencias entre las escalas de las diferentes muestras. Existen otras alternativa de normalización, en donde por ejemplo se realiza un escalamiento, para que todas las muestras tengan la misma desviación absoluta respecto a la mediana (**"MAD"** del inglés Meadian Absolute Deviation, Yang et al. (2002)). También se pueden estandarizar por "**cuantiles**", para que todos tengan la misma distribución empírica (Bolstad et al., 2003), como se muestra en la figura 2.4, o realizar un promedio robusto entre los diferentes microarreglos ("**RMA**", del inglés Robust Multi-array Average Irizarry et al. (2003)). Con estas diferentes normalizaciones, se logra que los datos provenientes de diferentes muestras tengan distribuciones similares, removiendo así artefactos propios de la técnica.

2.3.4. Reducción, proyección o integración de datos

El desempeño de los algoritmos de MD para la búsqueda de patrones, es dependiente del tamaño del espacio de entrada. Este espacio consiste en todas las posibles entradas a nuestro sistema, que en el contexto de tecnologías de alto rendimiento, son los niveles de expresión de las proteínas y/o genes. No obstante, la cantidad de ellas suele ser del orden de cientos a decenas de miles, superando ampliamente a la cantidad de ejemplos/casos/sujetos/muestras biológicas. Particularmente en proteómica y genómica, disminuir la cantidad de variables de entrada favorece significativamente

el proceso de análisis ontológico funcional (sección 1.2).

En general, la reducción o selección de las variables de entrada se realiza mediante el uso de modelos estadísticos. Estos comprenden desde una simple prueba t (Walpole et al., 1999), hasta modelos lineales (Graybill, 2000) o incluso lineales mixtos (Pinheiro y Bates, 2009). Estos modelos permiten encontrar aquellas proteínas/genes que presentan diferencias en los niveles de expresión, a un nivel de significancia dado, entre las diferentes condiciones experimentales dependiendo de la hipótesis biológica del investigador (control vs tratamiento 1 o tratamiento 2, etc.).

Selección de proteínas/genes

En el contexto de las tecnologías de alto rendimiento, el punto de partida es la **matriz de expresión**. En ella cada fila representa un individuo correspondiente a un *spot* (proteína) o *probeset* (gen) y cada columna representa una combinación de tratamientos, para una replica biológica dada. En este contexto cada individuo es modelado de forma independiente.

En el **caso más elemental** de un experimento tipo caso-control con **solo una réplica** para cada condición, tendremos una matriz de N filas (proteínas/genes) en el orden de decenas de miles, por dos columnas (control y tratamiento). En esta configuración, es imposible aplicar un modelo estadístico. No obstante, es habitual obtener la diferencia de expresión entre las dos condiciones y a partir de su valor absoluto utilizar alguno de los siguientes criterios de selección:

- Establecer un **umbral** para seleccionar aquellos individuos que lo superen.
- Ordenar las filas de la matriz y seleccionar una **cantidad** establecida.
- Obtener la función de densidad empírica de la diferencia de expresión y establecer un **percentil** para la selección, a una o dos colas.

En el caso de **diseños de mayor complejidad**, es usual ajustar un *modelo lineal* para cada uno de los individuos de forma independiente (Graybill, 2000). Este modelo dependerá del diseño experimental bajo estudio. Por ejemplo, en un experimento de microarreglos con tres niveles de tratamiento (A, B y C) y r repeticiones, es posible utilizar un modelo de clasificación unifactorial (2.1) para cada uno de los genes:

$$y_{ij} = \beta_0 + \beta_1 \tau_B(i) + \beta_2 \tau_C(i) + \varepsilon_{ij} \quad i = 1, \ldots, N \quad j = 1, \ldots, r \tag{2.1}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2) \quad \wedge \quad Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0 \quad \forall\, i \neq k \wedge j \neq l \tag{2.2}$$

donde:

y_{ij} es el valor observado de expresión del i-ésimo gen, en la j-ésima repetición,

$\beta_0, \beta_1, \beta_2$ y σ^2 son parámetros desconocidos a estimar,

ε_{ij} es el error aleatorio no observable, sujeto a los supuestos de (2.2)

$\tau_B(i)$ y $\tau_C(i)$ son variables binarias para indicar la pertenencia o no (1 o 0), del i-ésimo gen al tratamiento B o C respectivamente.

El modelo (2.1) es ajustado mediante **Mínimos Cuadrados Ordinarios** (Walpole et al. (1999) y Graybill (2000)). Luego, el investigador puede plantear la/s diferentes **pruebas de hipótesis** particular/es, a partir de una combinación lineal de las medias de tratamientos, lo que se conoce como **contraste** (Walpole et al. (1999) y Graybill (2000)). Por ejemplo, puede seleccionar aquellos genes que se expresen de forma diferente entre dos pares de tratamientos (A vs B), es decir, seleccionar los genes para los cuales existe evidencia para rechazar la hipótesis nula (poseen igual expresión en ambos tratamientos).

En esta tesis se utiliza la implementación de modelos lineales del paquete **limma** (Smyth, 2004) del lenguaje R (R Core Team, 2013). Este paquete adicionalmente permite realizar una corrección empírica de Bayes, para reducir las varianzas de cada modelo (proteína/gen) hacia un valor común y aumentar los grados de libertad de las varianzas individuales. De esta manera, se tiene una prueba de hipótesis con mayor potencia estadística. Por otra parte, también hay que tener en cuenta que debido a la cantidad de modelos que se deben ajustar, es necesario realizar una corrección por comparaciones múltiples de los valores p, obtenidos para cada prueba de hipótesis. Esto permite reducir la cantidad de falsos positivos, por ejemplo utilizando FDR (del inglés False Discovery Rate, Benjamini y Hochberg (1995)). Luego de fijar un valor de significancia, por ejemplo $\alpha = 0{,}05$, se determinan aquellos **candidatos** (proteínas/genes) que se expresan de forma diferencial. Estos candidatos serán comparados con aquellos disponibles en el experimento, que harán de lista de **referencia** para el análisis funcional.

Así, a través de los modelos lineales, el investigador tiene la flexibilidad de definir el contraste específico para su problema biológico (ver capítulo 1). Justamente, la salida de un contraste permite reducir la cantidad de proteínas/genes presentes en la totalidad (referencia) de datos, a una lista de candidatos más pequeña para realizar el análisis funcional.

Integración de perfil de expresión de proteínas/genes

Una manera de comprobar si las proteínas/genes candidatos han sido seleccionados de forma apropiada, es a través de la "visualización" del perfil de expresión. Para ello se utilizan "**mapas de calor**" (del inglés heatmap), donde se comprueba que los candidatos elegidos, distingan adecuadamente las condiciones/tratamientos involucrados en cada uno de los contrastes de interés (Wilkinson y Friendly, 2009).

En el contexto del análisis ontológico-funcional, el mapa de calor representa los valores de la matriz de expresión, utilizando una paleta de colores. Usualmente se utiliza una escala continua de color *rojo* y *verde* para representar aquellos candidatos *subexpresados* y *sobrexpresados* respecto a un tratamiento, respectivamente. En esta gráfica, el orden tanto de las filas y columnas se modifica de forma tal que se puedan apreciar la existencia (o no) de **asociaciones** entre candidatos y tratamientos (figura 2.5). A tales efectos, se realiza un **doble agrupamiento** sobre los valores de expresión de dichos candidatos. En las filas (geles/microarreglos) se observa si las réplicas biológicas de cada condición se comportan de manera similar, es decir, pertenecen al mismo agrupamiento. En columnas (proteínas/genes) se evalúa si la expresión de un mismo candidato es influenciada por la condición experimental, es decir, pasa de sobrexpresado a subexpresado o viceversa.

Este comportamiento se muestra en la figura 2.5, para el ejemplo del contraste propuesto entre los tratamientos A y B de la sección 2.3.4. En la figura se aprecia que las dos réplicas de cada tratamiento pertenecen al mismo agrupamiento, es decir, para el caso de A (A1 y A2) y para B (B1 y B2). Por otra parte, en columnas se observan dos grupos de genes sobre y subexpresados, que invierten su nivel de expresión cuando se los mide en el otro tratamiento. De esta manera se puede comprobar visualmente que el comportamiento de los candidatos seleccionados es el esperado para el diseño experimental propuesto, como es el caso de la figura 2.5.

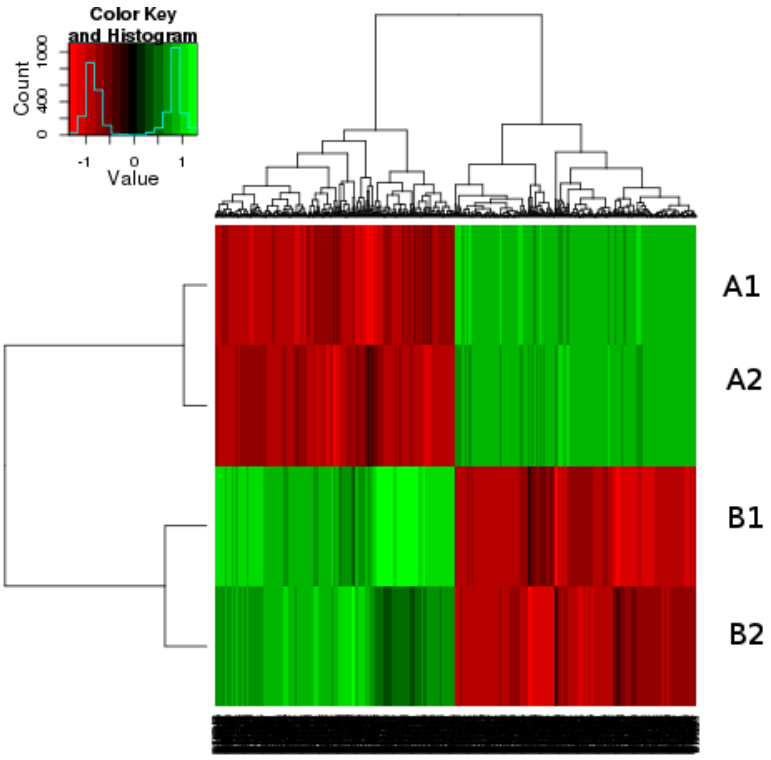


Figura 2.5: Mapa de calor de los genes seleccionados con expresión diferencial entre los tratamientos A y B. Note que las réplicas biológicas de cada tratamiento (en filas) se encuentras agrupadas como se aprecia en el dendograma de la izquierda, induciendo un agrupamiento de genes (en columnas) que cambian su nivel de expresión en cada tratamiento.

2.4. Modelado

Una vez realizado el **entendimiento del problema** (sección 2.2) y **entendimiento de datos** (sección 2.3), quedando así *definido el problema* y *preparados los datos*, es tiempo de revelar la información que estos contienen. Para ello, el siguiente paso es seleccionar una metodología o un algoritmo de MD para realizar la etapa de **modelado**.

En la presente tesis, la metodología de MD seleccionada como motor de cálculo para el *análisis funcional* es **SEA** (del inglés Set Enrichment Analysis). A través de ella, es posible evaluar qué procesos y/o funciones biológicas, donde participan una lista de candidatos en su conjunto, se encuentran modificados (enriquecidas) por el experimento cuando son comparados con una lista (basal) de referencia, como se introdujo en la sección 1.2. Para ello se utilizó como punto de partida, una lista de proteínas/genes **candidatas** (reducida), obtenidas mediante modelos lineales (sección 2.3.4) y como lista de **referencia**, la totalidad de proteínas/genes que han transitado con éxito por los diferentes pasos del entendimiento de datos (consistencia, integridad, filtrados, normalización, etc.). El análisis se realizó sobre la información biológica contenida en las ontologías de **GO** y **KEGG**, la cual ha sido presentada en la sección 1.1.

2.5. Evaluación

No siempre es posible contar con una cantidad suficientemente grande de datos, como sería deseable, dificultando la tarea de generar tres (o al menos dos) conjuntos de datos para entrenar, validar y evaluar el/los modelo/s utilizado/s en la etapa de modelado (sección 2.4). En estas circunstancias es donde las técnicas de **validación** juegan un papel preponderante, proporcionando estadísticas más fiables sobre los resultados, aún cuando el número de datos sea reducido. En este caso, es usual utilizar la técnica de "**Bootstrap**" como estrategia de validación (Efron, 1979).

La idea básica de *Bootstrap* es que la inferencia sobre los datos de una población pueden ser obtenida a partir de una muestra representativa. No obstante, por diferentes tipos de *restricciones*, la muestra termina siendo pequeña y/o no se pueden obtener muestras adicionales del fenómeno bajo estudio. Entonces, el comportamien-

to de la población se modela a través de nuevas muestras con reposición sobre los datos originales, lo que se conoce como "*remuestreo*", es decir, obtener una nueva muestra del tamaño original, permitiendo que existan valores repetidos. Este tipo de metodología es usualmente aplicada cuando:

- La **distribución teórica** del estadístico de interés es *desconocida*.
- La **distribución teórica** del estadístico *no es fácil de calcular*.
- El **tamaño de la muestra es insuficiente** para *estimar* el estadístico.
- Es necesario *estimar* la **potencia** del modelo utilizado y sólo se dispone de una *muestra piloto pequeña*.

A través de esta metodología, potencialmente se puede contar con una cantidad de "**muestras bootstrap**" suficientemente grande (cientos a miles), para estimar el comportamiento de la población. En caso contrario, resulta imposible poder abordar cualquiera de las aplicaciones anteriores.

En el contexto del *análisis ontológico funcional*, no es usual utilizar técnicas de "**validación mediante simulación**". Justamente, la aplicación de este tipo de metodologías permite aumentar la fiabilidad sobre los resultados en términos de **potencia estadística**, en el sentido de detectar enriquecimiento, cuando el efecto "verdaderamente existe". Es decir, reducir la posibilidad de enriquecimiento "espurio", producto de artefactos que puedan sesgar los resultados funcionales obtenidos.

En biología no existe un "patrón de oro" (del inglés *gold standard*) con el cual se pueda validar el "modelo biológico", razón por la cual es habitual utilizar una tecnología diferente a la empleada para la obtención de datos, para obtener resultados similares. Esto se conoce como "**validación biológica**". En tecnologías de alto rendimiento, la comunidad científica acepta la utilización de la reacción en cadena de la polimerasa en tiempo real (R-T PCR del inglés *Real-Time Polymerase Chain Reaction*, Erlich (1989)) como el estándar biológico. Por otra parte, cuando no es posible realizar una validación biológica, es habitual referirse a la evidencia existente en la literatura científica para "**validar por bibliografía**". Cabe destacar que en esta tesis se excluye la validación biológica como estrategia de la etapa de evaluación.

2.6. Reporte

En esta etapa se presentan los resultados obtenidos del *análisis ontológico funcional*, a través de las diferentes etapas del KDD. En este contexto, las herramientas bioinformáticas presentadas en la sección 1.3 utilizan diversas estrategias de **visualización**, entre las cuales es posible encontrar (sección 1.3.5):

- **Listas tabulares**: extensas tablas de texto plano, en el orden de cientos a miles de filas por decenas de columnas, con la diferente información funcional (proteínas, genes, funciones/términos, valores p, etc.).
- **Reportes basados en tecnologías web**: usualmente son páginas estáticas, o con capacidades dinámicas limitadas, donde se pueden explorar los reportes tabulares.
- **Árboles jerárquicos desplegables**: estructuran (agrupan) la información de forma jerárquica, permitiendo explorar los resultados. En el caso de GO, este tipo de visualización produce una duplicación de información.
- **Imágenes prediseñadas**: en esta categoría existe un abanico de posibilidades, entre las siguientes:
  - **Grafos de GO**: utilizan la propia estructura de GO, como estrategia de resumen y visualización, de los resultados ontológico-funcionales.
  - **Vías metabólicas de KEGG**: representan las relaciones existentes entre los diferentes compuestos, enzimas, etc. presentes en la vía.
  - **Anotación**: permiten visualizar la evidencia existente (o no), entre proteínas/genes y diferentes categorías/términos de interés, utilizando una tabla de doble entrada.

Si bien las anteriores son alternativas válidas para **visualizar** y **explorar** los resultados, con las limitaciones descriptas en la sección 1.3, éstas no permiten integrar resultados funcionales obtenidos de diferentes análisis. En este sentido, en esta tesis se aborda esta problemática mediante metodologías del tipo **MEA** para integrar/explorar este tipo de resultados, como estrategia de **consolidación del conocimiento**.

2.6.1. Comentarios finales

En este capítulo, se ha utilizado el KDD como un marco ordenado de trabajo, aportando herramientas de MD y dirigiendo el trabajo hacia la búsqueda de información relevante en el contexto del *análisis ontológico-funcional*. En este sentido, el procesamiento de los datos es extenso y es necesario comprender tanto la génesis de los datos como los algoritmos involucrados en las diferentes herramientas de MD, para que ellas brinden información clara e interpretable en cada una de las etapas del análisis.

Capítulo 3

Aportes realizados al análisis ontológico-funcional desde la MD

En este capítulo se muestran las **diferentes metodologías desarrolladas**, a los efectos de proporcionar *herramientas de MD* que permitan un *análisis más estructurado y completo* de la información biológica existente y la *visualización* de nuevas relaciones inferidas de la *integración/contraste* sobre diseños experimentales de mayor complejidad.

Basados en la perspectiva del KDD del capítulo 2, se presenta el **flujo de trabajo** completo del *análisis ontológico-funcional*. De esta manera el lector tendrá una visión global, donde podrá particularizar la aplicación de los conceptos introducidos en el capítulo 1. Más aún, podrá rápidamente comprender dónde se encuentra el foco de los aportes realizados por esta tesis, que abordan problemas concretos presentados en los dos capítulos anteriores en las diferentes etapas del KDD.

En lo que respecta al *entendimiento de datos*, se propone una estrategia modular y extensible, que permite incorporar diferentes tecnologías de alto rendimiento (2D-DIGE, microarrays, secuenciamiento, etc.), para abordar la "**consistencia e integridad de identificadores**". A partir de este aporte, es posible disminuir el sesgo de anotación mediante la integración de diferentes bases de datos, para incorporar proteínas/genes que son descartados desde el comienzo del análisis por una incorrecta manipulación como se describió en la sección 2.3.2. Por otra parte, se propone una estrategia para la "**exploración multivariada y control de calidad**" de

los valores de expresión de las proteínas/genes utilizando la información del diseño experimental. Esto permite comprobar la existencias de fuentes de variabilidad tecnológica no tenidas en cuenta, al igual que explorar la variabilidad que introducen los diferentes niveles de tratamientos controlados en el experimento.

En la etapa de *modelado* se codificó una herramienta que permite **conectividad al portal DAVID** (**RDAVIDWebService**) para realizar el análisis funcional propiamente dicho (Fresno y Fernández, 2013b). De esta manera, el investigador puede acceder de forma programática a uno de los portales de exploración funcional de mayor impacto en la comunidad científica en los últimos años. Esto posibilita realizar diferentes análisis no disponibles para la **integración y contrastes de múltiples referencias** (Fresno et al., 2012), como también automatizar procesos de consulta a DAVID sin intervención manual. Los dos aportes anteriores, permiten en su conjunto una estrategia de *evaluación* (validación) sobre la robustez del enriquecimiento obtenido mediante técnicas de remuestreo (bootstrap). Por último, se introduce una interfaz donde se pueden "**visualizar y explorar los resultados**" de forma interactiva, incorporando la información de expresión (Fresno et al., 2011). De esta manera, el investigador puede extraer mayor información de los *reportes* y focalizar su atención en la exploración biológica, frente a lo tedioso que resulta este tipo de exploración en la actualidad.

3.1. Flujo de trabajo

El procesamiento de los datos en el contexto del *análisis ontológico-funcional* es extenso, aún bajo un marco de trabajo ordenado como el KDD. En este sentido, es necesario comprender tanto la génesis de los datos, al igual que los algoritmos de MD involucrados en cada uno de las etapas como se muestra en las figuras 3.1 y 3.2.

Como se presentó en la sección 2.3, en la etapa de **entendimiento de datos** se obtienen los *valores de expresión* de proteínas o genes (figura 3.1), dependiendo de la plataforma tecnológica utilizada. A estos datos se le adiciona la información de *anotación* obtenida de la identificación de las proteínas (2D-DIGE) o la provista por el fabricante (microarreglos). En la sección 2.3.2 se presentaron los diferentes inconvenientes relacionados a la *consistencia e integridad* de anotación en lo que respecta

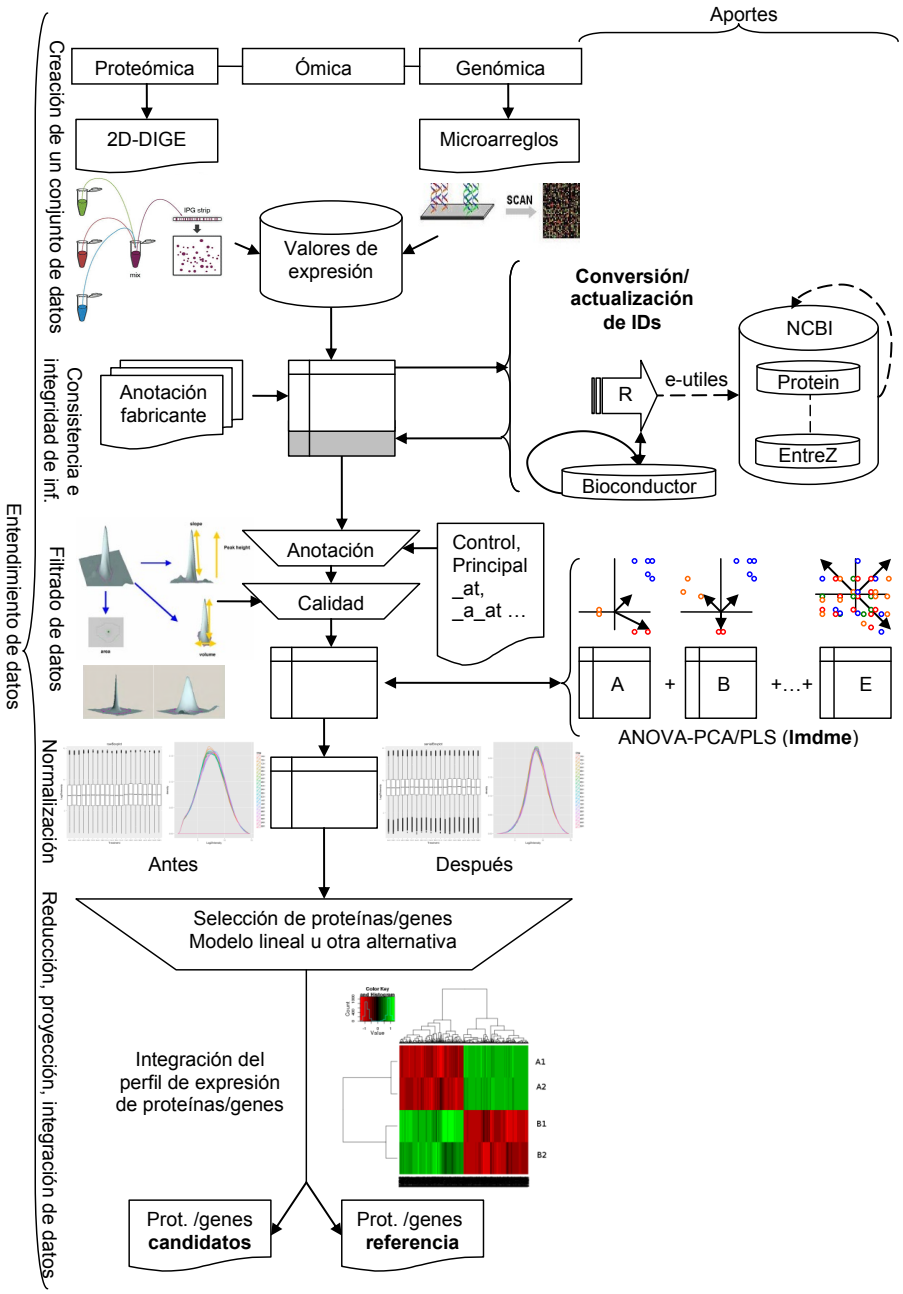


Figura 3.1: Diagrama de flujo de las diferentes etapas del "*entendimiento de datos*" involucradas en el *análisis ontológico-funcional* según se describe en detalle en el capítulo 2. Adicionalmente, se incorporan las diferentes contribuciones realizadas a lo largo del desarrollo de doctorado.

a la información de los IDs (estabilidad, versiones, trazabilidad, etc.). Justamente, el desconocimiento de los anteriores y la falta de acciones correctivas, impacta negativamente en el propio proceso de MD sobre la búsqueda de patrones. En este sentido, el análisis presenta un sesgo dado que es imposible extraer patrones sobre datos que se omiten *a priori*, por una exclusión de proteínas/genes que no fueron directamente reconocidos por la herramienta bioinformática utilizada. No obstante, estas proteínas/genes pueden ser incluidas en el análisis siempre y cuando se acceda a la información que se encuentra disponible en otros repositorios de anotación. Este es el primer desafío que se abordó en esta tesis, donde se priorizó indagar en forma automática la **conversión y actualización de los IDs**. Para ello, se codificaron diferentes módulos dependientes de la tecnología en lenguaje **R** (R Core Team, 2013), donde se conecta de forma local y programática a los paquetes de anotación disponibles en el repositorio de **Bioconductor** (Gentleman et al., 2005). En caso de no ser exitosa la conversión/actualización, se accede a las bases de datos del **NCBI** como *Protein* y *EntreZ*, entre otras (Maglott et al., 2011), mediante la interfaz de **e-utiles** (National Center for Biotechnology Information, 2010). De esta manera, el investigador puede tener trazabilidad y conocer el estado actual de cada proteína/gen, accediendo a información adicional (proteínas/genes en el recuadro gris de la figura 3.1), que en caso contrario se excluye en el análisis, con la consecuente pérdida de datos biológicos potencialmente útiles.

Una vez finalizada la etapa de consistencia e integridad de identificadores, se cuenta con una tabla con datos de expresión y anotación. A esta tabla se le aplican diversos *filtros* de *anotación* y *calidad de señal* como se presentó en la sección 2.3.3. Sin embargo, los abordajes clásicos no consideran la naturaleza multivariada del diseño experimental, razón por la cual se implementó una metodología (disponible en la librería **lmdme**, Fresno et al. (2014); Fresno y Fernández (2013a)) para disgregar las fuentes de variabilidad de los diferentes factores mediante una descomposición **ANOVA** a través de modelos lineales (sección 2.3.4). De esta manera, la matriz de expresión puede ser interpretada como la suma de la contribución de los diferentes factores, esquematizadas en los aportes de la figura 3.1 por las matrices $\mathbf{A}$, $\mathbf{B}$, ..., $\mathbf{E}$. Sobre estas matrices se puede realizar un análisis de componentes principales (**PCA**), conocido como ASCA/APCA (De Haan et al. (2007) y Smilde et al. (2005)), o

regresión de mínimos cuadrados parciales (**PLS**, Shawe-Taylor y Cristianini (2004)), la cual es novedosa en este tipo de análisis de descomposición ANOVA. Así, es posible explorar de forma multivariada la existencia de patrones de correlación existentes en los datos que puedan deberse a efectos no esperados, de manera de evaluar la *calidad* del experimento o buscar patrones de proteínas/genes relacionados al diseño experimental planteado, mediante gráficos conocido como **biplots** (Peña, 2002).

Seguidamente debe siempre estudiarse la necesidad de normalizar la matriz de valores de expresión, dado que es común que la misma se vea afectada por variabilidad técnica. Tanto para reducir dichas fuentes, como para cumplir con los supuestos biológicos deberá aplicarse un proceso de normalización como se describió en la sección 2.3.3. Esta transformación también es necesaria para la etapa de *reducción, proyección e integración de datos*, en especial para **seleccionar proteínas/genes candidatas** mediante *modelos lineales* u *otra alternativa* de las propuestas en la sección 2.3.4. Luego, se puede **explorar/analizar los perfiles de expresión** mediante *mapas de calor*, para comprobar si existe un agrupamiento dada la selección de proteínas/genes realizada. Así, la salida del *entendimiento de datos* culmina en dos listas de proteínas/genes: una con los **candidatos** y otra de **referencia**, que son necesarios para el *análisis de enriquecimiento funcional*, como se describió en la sección 1.2.

En la sección izquierda de la figura 3.2 se presenta el flujo que hasta el desarrollo de esta tesis era habitual de aplicar. En la etapa de *modelado*, se utiliza la información funcional contenida en las ontologías de interés como por ejemplo GO y KEGG (sección 1.1), para realizar una análisis del tipo **SEA** como se describió en la sección 1.2.1, utilizando las dos listas obtenidas como salida del *entendimiento de datos*. Dependiendo de la/s herramienta/s utilizada/s, se debe realizar un procesamiento con intervención del usuario, usualmente a través de un portal web (sección 1.3), obteniendo como salida la totalidad de **categorías ontológicas** bajo análisis. Estas categorías son luego *evaluadas* con algún método de corrección por comparación múltiple (e.g. **FDR**) para la selección de términos enriquecidos, como se describió en la sección 2.1.2. Estos resultados son presentados mediante alguna de las posibilidades de *reportes*, dependiendo de la herramienta utilizada: **tablas o páginas HTML**, **imágenes prediseñadas** (e.g. grafos de GO). Así, el investigador puede explorar

e indagar sobre las relaciones inferidas producto del experimento, con las limitaciones ya mencionadas en la sección 1.3.5 como por ejemplo: la extensión del reporte, duplicación de información, imposibilidad de integración de datos de expresión o resultados funcionales de otros experimentos, etc.

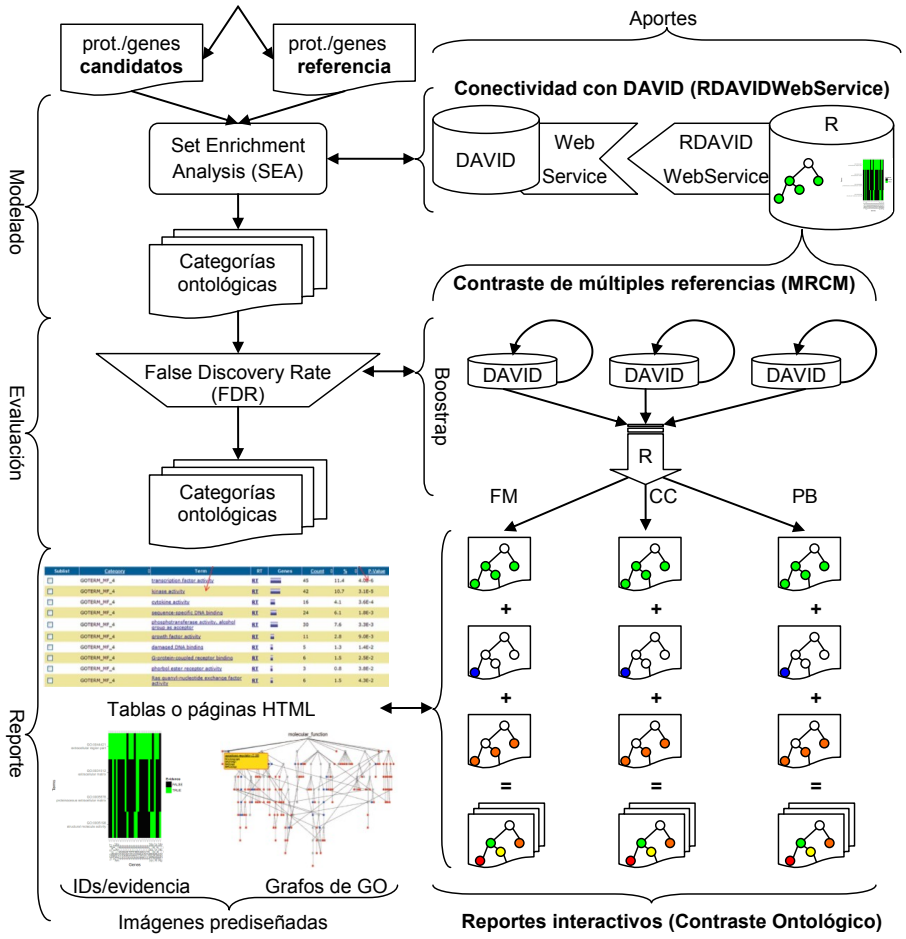


Figura 3.2: Diagrama de flujo de las diferentes etapas del KDD donde se hace énfasis en el "*modelado*", "*evaluación*" y "*reportes*", involucradas en el *análisis ontológico-funcional* según se describe en detalle en el capítulo 2. Adicionalmente, se incorporan las diferentes contribuciones realizadas a lo largo del desarrollo de doctorado.

Frente a los diferentes inconvenientes presentados en los capítulos 1 y 2, en la sección derecha de la figura 3.2 se muestran las mejoras propuestas en esta tesis para facilitar el análisis ontológico-funcional. Mediante los desarrollos de esta tesis,

la etapa de *modelado* se puede realizar de forma automática a través de la librería **RDAVIDWebService** (Fresno y Fernández (2013b,c)). Ella tiene implementado un módulo que permite **conectividad al portal DAVID**, `http://david.abcc.ncifcrf.gov/`, a través de la interfaz de servicios web que ésta provee (DWS, Jiao et al. (2012)). Esto permite obtener resultados de tipo SEA/MEA de forma programática desde el lenguaje R. Por otra parte, las mismas visualizaciones disponibles desde la interfaz web de DAVID, están ahora todas disponibles en R. Estas visualizaciones no son provistas por la interfaz DWS que proporciona DAVID (sección 1.3.5). Adicionalmente RDAVIDWebService permite generar grafos de enriquecimiento de GO incorporando funcionalidades de diferentes librerías de Bioconductor (Gentleman et al., 2005), que no están disponible en el sitio web de DAVID. De esta manera, es posible contextualizar los resultados utilizando la estructura de GO para su exploración, a los efectos de tener una rápida visión funcional de los resultados experimentales.

Otra característica distintiva de RDAVIDWebService, es que permite incorporar resultados obtenidos desde el portal web DAVID, es decir, incluso de aquellos almacenados en análisis anteriores. Esta facilidad permite trabajar sin conectividad a internet, lo que es muy útil cuando es necesario reanalizar los datos, al igual que cuando se trabaja en colaboración entre distintos grupos de investigación que acceden a DAVID por cualquier medio disponible, para obtener las diversas visualizaciones de los resultados que la librería brinda.

Una vez finalizada la etapa de *modelado*, se continúa con la etapa de *evaluación* del conocimiento adquirido. Para dicho fin, en esta tesis se implementó la metodología de **contraste de múltiples referencias** (MRCM, Fresno et al. (2012)). Mediante la utilización de RDAVIDWebService es posible implementar una estrategia de validación de términos enriquecidos mediante remuestreo **bootstrap** (sección 2.5) sobre la lista de referencia. Ésta no sería posible sin RDAVIDWebService ya que de otra manera debería de realizarse manualmente y, dado que la metodología requiere al menos 100 o más remuestreos, se torna una tarea tediosa. Esta validación permite obtener una medida de "*potencia estadística*" sobre la "robustez" de los términos enriquecidos frente a la referencia utilizada (sección 1.2.2). Es decir, a partir del remuestreo obtenemos un valor adicional al **FDR**, el cual nos permite conocer

la proporción de veces que se encuentra enriquecido de la totalidad de simulaciones realizadas. Esta es una característica no disponible en las herramientas funcionales actuales, la cual asiste al investigador en la detección de categorías ontológicas que no presenten enriquecimiento espurio (por azar), para su posterior "validación biológica".

En cuanto a los "*reportes*", se propone como alternativa utilizar la metodología desarrollada llamada "**Contraste Ontológico**" (Fresno et al., 2011). Esta metodología permite integrar resultados de diferentes análisis funcionales, por ejemplo de las simulaciones bootstrap del MRCM, diseños experimentales de mayor complejidad o incluso de resultados de diferentes análisis. Estos resultados son integrados en un único **reporte interactivo**, el cual puede ser explorado utilizando el navegador web de preferencia del investigador, sin necesidad de conectividad a internet. A su vez, estos reportes incluyen la **información de expresión** de las proteínas/genes permitiendo una rápida integración visual en un entorno unificado para su exploración. Adicionalmente, es posible continuar el análisis de las proteínas/genes de una categoría de interés, utilizando los hipervínculos hacia las bases de datos del **NCBI**.

Cada uno de los aportes realizados en esta tesis, aborda aspectos específicos de la problemática relacionada al análisis ontológico-funcional. En este contexto, en las diferentes etapas del KDD se proporciona una *herramienta de MD*, que permite un *análisis más estructurado y completo*, brindando *información clara e interpretable*.

3.2. Consistencia e integridad de anotación

En esta tesis se siguió la recomendación propuesta por Zeeberg et al. (2004), para abordar la problemática de consistencia e integridad de anotación. Esta consiste en transformar lo más temprano posible los datos de anotación, a un tipo de identificador que sea *estable* y a su vez la base de datos posea *trazabilidad*. En este sentido, se optó por utilizar identificadores del tipo **Entrez Gene ID** presentado en la sección 2.3.2, dado que posee las dos características deseadas y adicionalmente es posible conocer el **estado actual** del ID: "*no codifica*", es "*obsoleto*" o el "*vigente*" a la fecha. En caso que fuere necesario, es posible acceder a la información asociada a dicha proteína/gen, como por ejemplo su símbolo, alias conocidos, descripción, publicaciones, etc. A su

vez, este tipo de ID es el utilizado en la ontología de GO (sección 1.1.1) y diversas herramientas bioinformáticas como por ejemplo DAVID y GOstats (sección 1.3).

En este contexto, se propone una metodología de **conversión/actualización de IDs** basada en la obtención temprana de los identificadores equivalente de **Entrez Gene ID**, es decir, lo más próximo a la generación de los datos, teniendo en cuenta las particularidades de cada tecnología de alto rendimiento utilizada. El resultado de dicha **conversión** atraviesa un proceso iterativo, donde se utilizan diversos repositorios de anotación para establecer el "estado" del ID. En el caso de que el estado no fuere el actual, es necesario **actualizar** la información al último registro disponible. De esta manera, es posible conocer de forma transparente la historia de cada proteína, gen, secuencia, etc. a lo largo de los diferentes módulos de anotación, sin que ello implique pérdida de información. Ahora, el investigador puede optar por diferentes abordajes dependiendo de los resultados de cada módulo: utilizar solamente aquellos que se han convertido y actualizado con éxito, volver a la identificación de las proteínas/genes cuyos GIs/IDs de fabricantes no han sido convertidos, etc.

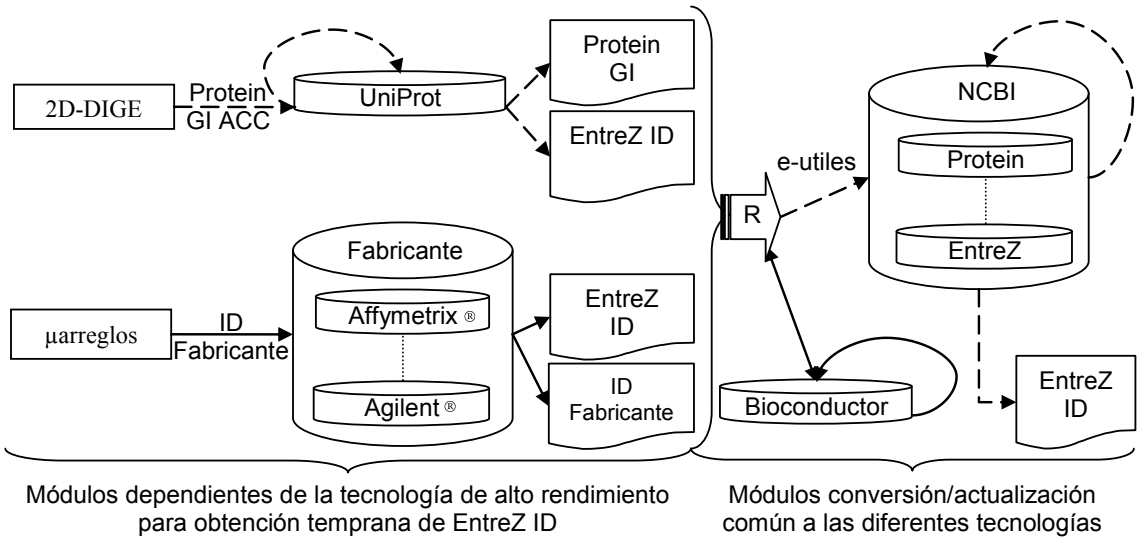


Figura 3.3: Diagrama de flujo para la conversión/actualización de anotación en experimentos realizados con tecnología 2D-DIGE o microarreglos de ADN. Note que el tipo de trazo de línea representa el acceso a los datos: *discontinuo*, requiere conectividad a internet, y *continuo* representa un acceso local.

En la figura 3.3 se muestra el diagrama de flujo de trabajo implementado para las tecnologías de alto rendimiento utilizadas en esta tesis: 2D-DIGE y microarreglos de ADN (sección 2.3.1). Adicionalmente, se muestra cómo la salida de estos módulos es incorporada a través del lenguaje R (R Core Team, 2013) para realizar las diferentes consultas a repositorios de anotación, de manera de convertir/actualizar la anotación.

3.2.1. Módulo de proteómica

En el caso de geles de proteínas, 2D-DIGE (sección 2.3.1), usualmente la identificación de proteínas termina con ID del tipo **PROTEIN_GI_ACCESSION** (GI), como se aprecia en la figura 3.3. Este tipo de identificador se encuentra disponible en una de las bases de datos de proteínas de mayor difusión, como es el caso de **UniProt** (Apweiler et al., 2004). Este consorcio posee un sitio web en el cual se pueden realizar conversiones de identificadores de forma *manual*, accediendo a la página `www.uniprot.org/?tab=mapping`. También proporciona una interfaz para acceder de forma *programática* a través de consultas mediante URLs (del inglés, Uniform Resource Locator), representada por la línea de trazo discontinuo de la figura 3.3. Estos URLs son similares a los que se generan en la barra de dirección del explorador de internet (FireFox®, Chrome®, etc.), cuando el usuario hace clicks en la página web. La ventaja de la interfaz programática es que la información se obtiene utilizando el mismo protocolo y puerto que el usuario utilizaría para navegar por internet. Es decir, no requiere ninguna configuración de firewall y/o proxies adicional para su utilización. A su vez, los resultados de la consulta pueden ser depurados creando el URL correspondiente, desde la pestaña de conversión del sitio web de **Uniprot**.

En esta tesis se desarrolló un conjunto de rutinas escritas en lenguaje R (R Core Team, 2013), que utilizan la librería **RCurl** (Lang, 2013a) para acceder de forma programática a la interfaz de **UniProt**. Estas rutinas se encuentran disponibles en el anexo digital `uniprot.R` de la sección A.1.1. Para utilizarlo, es necesario cargar en memoria el módulo con el comando `source`.

```
>source("uniprot.R")
>names(uniprot)
[1] "Tool" "Base" "Format"  "Columns"  "Compress"  "Mapping"  "Query"
```

Este módulo cuenta con el objeto `uniprot`, que posee diferentes atributos como por ejemplo: cuál es la dirección web `Base` de la interfaz (www.uniprot.org), qué herramientas (`Tool`) están disponibles (anotación, consulta, convertidor), cual es el formato (`Format`) de los reportes (.txt, .tab, etc.), las columnas (`Columns`) seleccionadas por defecto, etc. A través del objeto `uniprot`, es posible generar consultas desde R para luego obtener los resultados de una búsqueda (`Query`) o de la conversión de IDs (`Mapping`).

Por ejemplo, es posible definir dos `protein_gi_ids` como `"119577981"` y `"29462"`, para convertirlos a identificadores del tipo `UniProtKB AC`, consultando al objeto `uniprot$Mapping$Mapper`:

```
>protein_gi_ids<-c("119577981", "29462")
>out<-uniprot$Mapping$Mapper(ids=protein_gi_ids,
      from=uniprot$Mapping$From["GI number*"],
      to=uniprot$Mapping$To["UniProtKB AC"])
>out$tab
       From     To
1 119577981 B2ZZ90
2     29462 P09486
```

La salida de la consulta se almacena en el objeto `out`, donde el texto separado por tabulaciones de la conversión se encuentra en el campo `$tab`. En esta consulta fue posible convertir los dos identificadores, solicitados en la columna `From`, obteniendo la respuesta en la columna `To`.

Adicionalmente, es posible acceder a la información de anotación asociada a los IDs de las proteínas convertidas en las bases de conocimiento de **Uniprot** curadas (**Swiss-Prot**) y no curadas (**TrEMBL**), utilizando los campos establecidos por defecto: `Entry`, `Entry name`, `Status`, `Protein names`, `Gene names`, `Organism` y `Length`. En este ejemplo en particular se obtiene la siguiente tabla:

```
> rbind(out$uniprotReviewedYes, out$uniprotReviewedNo)
   Entry   Entry name     Status
1 P09486   SPRC_HUMAN   reviewed
2 B2ZZ90 B2ZZ90_HUMAN unreviewed
```

```
                                Protein names
1 Secreted protein acidic and rich in cysteine
2          Acetyl-Coenzyme A carboxylase alpha
       Gene names             Organism Length
1        SPARC ON Homo sapiens (Human)    303
2 ACACA hCG_30204 Homo sapiens (Human)   2346
```

A su vez, este tipo de salida también puede ser ajustada (formato de salida, campos del reporte, etc.) dependiendo de las necesidades del usuario, como se describe en la ayuda del anexo digital A.1.1.

Una vez determinada la identidad de los GI IDs, en términos de identificadores de **Uniprot**, es posible realizar una nueva conversión pero ahora utilizando los últimos como punto de partida, a los efectos de obtener los `Entrez Gene (GeneID)` buscados por la metodología propuesta, como se muestra a continuación:

```
>out<-uniprot$Mapping$Mapper(ids=out$tab$To,
      from=uniprot$Mapping$From["UniProtKB AC/ID"],
      to=uniprot$Mapping$To["Entrez Gene (GeneID)"])
>out$tab
    From   To
1 B2ZZ90   31
2 P09486 6678
```

En este ejemplo la conversión utilizando la herramienta propuesta fue *exitosa*, como se aprecia en la salida de `out$tab`. Bajo este esquema de trabajo, es posible convertir la totalidad de proteínas candidatas de forma programática, evitando que el investigador acceda de forma manual, maximizando así el tiempo de procesamiento e información de anotación disponible.

En el caso que *no se hayan convertido* los IDs provistos, es posible realizar una consulta (`Query`) utilizando diferentes criterios de búsqueda dependiendo de la información que se tenga disponible: símbolo, descripción, etc. Continuando con el ejemplo anterior, en el caso de disponer del símbolo "**SPARC**", y conociendo que se está trabajando con proteínas humanas, con ayuda de la interfaz web, el usuario puede

explorar las diferentes alternativas de búsqueda avanzada que cumplan sus necesidades. En el ejemplo en cuestión puede seleccionar los campos "`gene`" y "`organism`" para especificar la siguiente consulta: `gene:SPARC AND organism:"Homo sapiens [9606]"`. Esta consulta es convertida automáticamente en formato compatible de URL, es decir, transforma los caracteres especiales: dos puntos por %3A, espacios por "+" y comillas dobles por %22, como se muestra a continuación (`$Query`):

```
>query<-"gene:SPARC AND organism:\"Homo sapiens [9606]\""
>uniprot$Query(query=query)
$Query
  "query=gene%3ASPARC+AND+organism%3A%22Homo+sapiens+[9606]%22"
$Data
   Entry   Entry name      Gene names     Status
1 P09486   SPRC_HUMAN        SPARC ON   reviewed
2 D3DQH8 D3DQH8_HUMAN SPARC hCG_39149 unreviewed
3 E5RK62 E5RK62_HUMAN           SPARC unreviewed
4 F5GY03 F5GY03_HUMAN           SPARC unreviewed
5 E5RJA5 E5RJA5_HUMAN           SPARC unreviewed
6 F5H4E2 F5H4E2_HUMAN           SPARC unreviewed
7 Q6QE20 Q6QE20_HUMAN           SPARC unreviewed
                                                                Protein names
1                          Secreted protein acidic and rich in cysteine
2 Secreted protein, acidic, cysteine-rich (Osteonectin), isoform CRA_a
3                                                        SPARC (Fragment)
4                                                        SPARC (Fragment)
5                                                        SPARC (Fragment)
6                                                        SPARC (Fragment)
7                                       Cysteine-rich protein (Fragment)
```

donde se aprecia que en la primera fila de `$Data` se encuentra la información de la misma proteína obtenida con anterioridad. Así, el investigador puede personalizar la consulta dependiendo de la precedencia de sus datos y reducir los resultados posibles. Si fuera el caso, es posible especificar valores en campos adicionales como por ejemplo elegir sólo las proteínas curadas (`Status=reviewed`).

Las funcionalidades de `Mapper` y `Query` implementadas en `uniprot.R`, permiten acceder a la interfaz de **Uniprot** y realizar diferentes operaciones para la *conversión* y *consulta* de proteínas, dependiendo de los datos que el usuario posea. Este módulo facilita la consulta programática a la interfaz web de **Uniprot**, automatizando las diferentes operaciones que pueda realizar el investigador. La salida de este módulo consiste en una tabla de anotación con los **EntreZ Gene ID** y en una con **Protein GI**. En el mejor de los casos, todos los identificadores fueron exitosamente convertidos.

3.2.2. Módulo de microarreglos

En el caso de microarreglos de ADN (figura 3.3), el punto de partida son los IDs del **fabricante** para cada plataforma utilizada (Affymetrix®, Agilent®, etc.). Utilizando R es posible acceder de forma local, a los diferentes archivos de anotación provisto por los fabricantes en sus respectivos sitios web, y obtener su equivalente **EntreZ Gene ID**. A diferencia del módulo de geles de proteínas (2D-DIGE), la conversión es más "*directa*" dado que se accede a un conjunto de campos y archivos definidos donde se conoce la secuencia de la sonda, como se describió en la sección 2.3.1.

En el caso de utilizar la plataforma de Affymetrix® para el chip HG-U133A 2.0, es posible encontrar el correspondiente archivo de anotación siguiendo el enlace `www.affymetrix.com/Auth/analysis/downloads/na33/ivt/HG-U133A_2.na33.annot.csv.zip`. Por lo general, dentro de este archivo se contiene uno de resumen (README.txt), donde se encuentran los descriptores de los campos y el segundo es la anotación propiamente dicha, usualmente en formato de campos separados por coma (HG-U133A_2.na33.annot.csv). Este último, por lo general posee un pequeño encabezado de una cantidad de líneas determinado, donde se especifica la fecha de las diferentes bases de datos utilizadas como por ejemplo *2012-05-07* para EntreZ Gene ID. Estas líneas deben ser ignoradas (`skip=25`), a los efectos de una correcta lectura del archivo (`read.csv`) como se muestra a continuación:

```
> anotacion<-read.csv(file="HG-U133A_2.na33.annot.csv",skip=25)
> dim(anotacion)
```

```
[1] 22277    41
> names(anotacion)
 [1] "Probe.Set.ID"                      "GeneChip.Array"
 [3] "Species.Scientific.Name"           "Annotation.Date"
 [5] "Sequence.Type"                     "Sequence.Source"
 [7] "Transcript.ID.Array.Design."       "Target.Description"
 [9] "Representative.Public.ID"          "Archival.UniGene.Cluster"
[11] "UniGene.ID"                        "Genome.Version"
[13] "Alignments"                        "Gene.Title"
[15] "Gene.Symbol"                       "Chromosomal.Location"
[17] "Unigene.Cluster.Type"              "Ensembl"
[19] "Entrez.Gene"                       "SwissProt"
[21] "EC"                                "OMIM"
[23] "RefSeq.Protein.ID"                 "RefSeq.Transcript.ID"
[25] "FlyBase"                           "AGI"
[27] "WormBase"                          "MGI.Name"
[29] "RGD.Name"                          "SGD.accession.number"
[31] "Gene.Ontology.Biological.Process" "Gene.Ontology.Cellular.Comp"
[33] "Gene.Ontology.Molecular.Function" "Pathway"
[35] "InterPro"                          "Trans.Membrane"
[37] "QTL"                               "Annotation.Description"
[39] "Annotation.Transcript.Cluster"     "Transcript.Assignments"
[41] "Annotation.Notes"
> conversion<-anotacion[,c("Probe.Set.ID","Entrez.Gene")]
> head(conversion)
  Probe.Set.ID       Entrez.Gene
1    1007_s_at 100616237 /// 780
2      1053_at              5982
3       117_at              3310
4       121_at              7849
5    1255_g_at              2978
6      1294_at              7318
```

En este ejemplo, la dimensión del archivo de anotación es de 22277 filas/sondas por 41 columnas/descriptores como describe la salida de `names(anotacion)`. En este contexto, es posible utilizar la columna "`Probe.Set.ID`" donde se especifica el ID del fabricante y la columna "`Entrez.Gene`" para obtener el tipo de ID requerido por la metodología propuesta. De esta manera, es posible construir una tabla de `conversion` de identificadores, para la cual se muestra su cabecera (`head(conversion)`). Note que para la primera sonda **"`1007_s_at`"**, existen dos IDs anotados en EntreZ ("100616237" y "780"), los cuales se encuentran delimitados por el carácter "`///`". Esta es otra particularidad que hay que tener en cuenta para el análisis funcional, es decir, la misma sonda debe manipularse con todos los EntreZ asociados. No obstante, las sondas terminadas en "_s_at" son removidas del análisis, como se describió en la sección 2.3.3.

Una particularidad que posee el módulo de microarreglos es que depende de la información de anotación provista por cada fabricante, razón por la cual deberá ser adaptado para cada una de las plataformas utilizadas. Así, bajo el esquema de la metodología propuesta, es posible convertir de forma automática los IDs, posterior a la adquisición de los valores de expresión. Consecuentemente, la salida de este módulo es una lista de EntreZ Gene ID y otra con los IDs del fabricante que no han logrado ser convertidos, como se muestra en la figura 3.3. En el mejor de los casos, todos los identificadores han sido exitosamente convertidos.

3.2.3. Módulo de conversión/actualización

Este módulo es el núcleo de la metodología propuesta, dado que es común a todas las tecnologías de alto rendimiento y es donde la conversión/actualización de IDs toma lugar. En este contexto, la entrada son las listas de IDs obtenidos en los módulos de proteómica y genómica (secciones 3.2.1 y 3.2.2). Las listas de IDs de entrada son procesadas utilizando el lenguaje R, permitiendo conversión/actualización de forma **local** y mediante consultas a **internet** como se muestra en el panel derecho de la figura 3.3.

Módulo de acceso local

El módulo de acceso **local** utiliza como estrategia de conversión alguno de los paquetes de anotación pertenecientes al repositorio **Bioconductor** (Gentleman et al., 2004). En el repositorio se encuentran dos grandes tipos de paquetes de anotación: aquellos que se corresponden con **microarreglos** de fabricantes, y los correspondientes a cada **organismo**.

Los paquetes de microarreglos son específicos para un chip y tienen una actualización de forma bianual. Una vez que los paquetes que sean necesarios son instalados en R desde internet, pueden consultarse de ahí en adelante de forma local utilizando el ID del fabricante. Continuando con el ejemplo de la sección 3.2.2, el chip HG-U133A 2.0 de la plataforma Affymetrix®, posee el paquete de anotación llamado "**hgu133a2.db**". La versión actual es la 2.9.0, y utiliza la definición de la base de datos de **EntreZ** de la fecha 2013-03-05. Para comenzar la conversión es necesario cargar la librería de la siguiente forma:

```
> library("hgu133a2.db")
> show(hgu133a2ENTREZID)
ENTREZID map for chip hgu133a2 (object of class "ProbeAnnDbBimap")
```

En este tipo de librerías, la información de anotación se encuentran estructurada en una serie de objetos cuyo nombre responde a la nomenclatura "`nombre_libreria_XXX`" donde XXX toma los valores: `ENTREZID`, `SYMBOL`, `GENENAME`, etc. Estos objetos son representados como grafos *bipartitos*, es decir, dos agrupaciones de nodos (izquierdos y derechos) conectados por arcos que permiten relacionar los IDs del fabricante (nodos izquierdos) con información del tipo XXX (nodos derechos). Teniendo esta idea en mente, es posible construir una tabla de conversión. Por ejemplo, los tres últimos IDs del ejemplo de la sección 3.2.2 definidos en el objeto `x` pueden ser convertidos siguiendo el siguiente código:

```
> x<-c("121_at", "1255_g_at", "1294_at")
> conversion<-unlist(mget(x=x, envir=hgu133a2ENTREZID,ifnotfound=NA))
> conversion<-data.frame(AffyID=names(conversion), EntreZ=conversion)
> row.names(conversion)<-1:nrow(conversion)
```

```
> conversion$Symbol<-unlist(mget(x=x, envir=hgu133a2SYMBOL, NA))
> conversion$Description<-unlist(mget(x=x,envir=hgu133a2GENENAME,NA))
> conversion
     AffyID EntreZ Symbol                                      Description
1    121_at   7849   PAX8                                    paired box 8
2 1255_g_at   2978 GUCA1A     guanylate cyclase activator 1A (retina)
3   1294_at   7318   UBA7 ubiquitin-like modifier activating enzyme 7
```

Los valores asociados a los IDs de `x` se obtienen utilizando la función `mget` del paquete **base** del motor de R. Adicionalmente a esta función se le debe especificar sobre qué objeto de anotación debe trabajar (`envir=hgu133a2ENTREZID`) y cuál es el valor por defecto (NA, del inglés not available, no disponible), en caso de que no exista un arco asociado a dicho ID. El resultado de esta búsqueda se puede almacenar en una estructura del tipo `data.frame` de R, a la cual se le puede incluir columnas con información de los símbolos (`Symbol`), nombre del gen (`Description`), etc., como se muestra en el ejemplo.

En el contexto de conversión/actualización de **anotación de microarreglos**, los paquetes presentes en Bioconductor son una alternativa válida para la conversión de IDs del módulo de microarreglos (sección 3.2.2). En otros casos puede ser utilizada de forma complementaria para obtener datos ausentes en la fuente de anotación del fabricante, como por ejemplo su símbolo, descripción, etc.

Por otra parte, los paquetes de anotación del **organismo** bajo estudio presentan otra alternativa para obtener la información de anotación de un tipo de ID en particular. Usualmente estos paquetes son denominados siguiendo la nomenclatura `org.YY.ZZ.db`, donde `YY` son dos letras que representan el organismo, y `ZZ` dos letras que determinan el tipo de ID principal de acceso. En estos paquetes los datos también se almacenan utilizando grafos bipartitos, de la misma manera que en los paquetes de microarreglos. Así, para el ejemplo anterior es posible utilizar la librería `org.Hs.eg.db`, la cual posee la información de anotación de humanos (`Hs`) y utiliza como identificador principal los **EntreZ Gene ID** (`eg`). De manera que es posible obtener la misma información de anotación realizando los cambios en los nombres de los objetos correspondientes, como se muestra a continuación:

```
> library("org.Hs.eg.db")
> x<-c("7849", "2978", "7318")
> anotacion<-unlist(mget(x=x, envir=org.Hs.egSYMBOL, NA))
> anotacion<-data.frame(EntreZID=names(anotacion),Symbol=anotacion)
> row.names(anotacion)<-1:nrow(anotacion)
> anotacion$Description<-unlist(mget(x=x,envir=org.Hs.egGENENAME,NA))
> anotacion
  EntreZID Symbol                                      Description
1     7849   PAX8                                    paired box 8
2     2978 GUCA1A     guanylate cyclase activator 1A (retina)
3     7318   UBA7 ubiquitin-like modifier activating enzyme 7
```

Utilizando los diferentes paquetes de anotación de organismos, el usuario se independiza de la tecnología de alto rendimiento utilizada (2D-DIGE, microarreglos, etc.), obteniendo así los símbolos, descripción, etc. asociados al ID principal de forma local. Esta metodología junto con los paquetes de anotación de los fabricantes de microarreglos, presentan una alternativa complementaria a los módulos de conversión específicos de cada tecnología (secciones 3.2.1 y 3.2.2). En este contexto, la manipulación de información de anotación se puede realizar de forma programática y local, mediante las diferentes alternativas que ofrece el repositorio de Bioconductor. Justamente, esto evita que el usuario deba instalar y mantener al día (actualización periódica) su propio repositorio local de anotación. Incluso en el caso que el organismo no se encuentre anotado, el usuario puede seguir los lineamientos descriptos en Bioconductor (`www.bioconductor.org`) para construir el paquete correspondiente, y subirlo al repositorio para que otros investigadores hagan uso de él.

La metodología propuesta, ha mostrado ser de gran utilidad en diferentes experimentos (Fresno et al. (2012), Loreti et al. (2013), Denninghoff et al. (2014)). Sin embargo, puede que aún con ella no sea posible convertir todos los IDs. Más aún, como se describió en la sección 2.3.2, es común que la información asociada a IDs se encuentre desactualizada. Esto se puede atribuir a que los propios archivos de anotación de los fabricantes o los paquetes de Bioconductor son viejos, desde años a meses respectivamente. De manera que es imprescindible obtener el "estado" de cada ID y **actualizar** su valor en caso de que sea necesario, mediante una consulta

a repositorios de internet, donde sea posible acceder a la última información vigente a la fecha del análisis.

Módulo de acceso a internet

Este módulo permite acceder, mediante una conexión a **internet**, a uno de los repositorios de mayor impacto en la comunidad científica, el **NCBI** (de las siglas en inglés, National Center for Biotechnology Information, Wheeler et al. (2007)). La gran ventaja de este repositorio es que representa un punto de acceso centralizado a diferentes bases de datos de *genes* (**EntreZ**, **UniGene**, etc.), *proteínas* (**Nucleotide**, **EST**, etc.), *publicaciones* (**PubMed**, **NLM Catalog**, etc.), etc. Adicionalmente, la actualización de las diferentes bases de datos se realiza de forma transparente para el usuario. Es decir, siempre se accede a la última versión disponible, sin que ello implique una instalación local de las diferentes bases de datos de anotación.

El sitio web de NCBI, `www.ncbi.nlm.nih.gov`, posee una interfaz de servicios web para su acceso de forma programática llamado **E-utiles** (National Center for Biotechnology Information, 2010). Este servicio web permite realizar prácticamente, todas las acciones que puede realizar un usuario desde la página `www.ncbi.nlm.nih.gov/sites/gquery?itool=toolbar`, las cuales pueden agruparse en tres categorías:

ELink: permite **vincular/convertir** los IDs presentes en una base de datos a otra.

ESummary: permite obtener la **información de resumen** (nombre, símbolo, estado actual, etc.) de los IDs solicitados, para una base de datos en particular.

ESearch: permite realizar una **búsqueda**, utilizando diferentes campos (nombre, símbolos, organismo, etc.) sobre las diferentes bases de datos, dependiendo de la información que posea el usuario.

Así, a través del acceso programático a **ELink** es posible convertir los IDs entre las diferentes bases de datos de forma no supervisada por el usuario, obteniendo el resultado del emparejamiento en cuestión de segundos, sin necesidad de hacer decenas de clicks en el sitio web para cada ID que se desea convertir. Además, **ESummary** permite obtener diferentes campos de anotación específicos para cada ID (nombre,

símbolo, etc.) y el estado actual de los registros ("vigente", "obsoleto" y "no codifica"). Esta es una característica distintiva frente a otros portales centralizados, dado que permite conocer el estado real de cada ID y utilizar la trazabilidad que ésta ofrece, para recorrer el historial de los diferentes IDs, a los efectos de obtener la última versión de ellos. Por otra parte, es posible realizar una búsqueda/consulta mediante **ESearch** sobre diferentes bases de datos de anotación, de manera similar a la realizada en **Uniprot** a través de `uniprot$Query` (sección 3.2.1).

En esta tesis se desarrolló un conjunto de rutinas escritas en lenguaje R (R Core Team, 2013), que utilizan la librería **RCurl** (Lang, 2013a) para acceder de forma programática a la interfaz de **E-utiles**. Las consultas se realizan utilizando algunas de las tres funcionalidades disponibles, es decir, **ELink**, **ESummary** y **ESearch**. No obstante, los resultados de cada una de estas funciones se encuentran en formato XML (de las siglas en inglés, eXtensible Markup Language), los cuales son adaptados utilizando la librería **XML** (Lang, 2013b), para su posterior utilización en R. Las diferentes rutinas desarrolladas se encuentran disponibles en el anexo digital "`eutiles.R`" de la sección A.1.2 y requieren de una cuenta académica solicitada a `eutilities@ncbi.nlm.nih.gov`.

Para utilizar el módulo de **eutiles**, es necesario cargarlo en memoria con el comando `source`. Por ejemplo, es posible invocar a `ELink` de una manera similar a la especificada en el módulo de proteómica de la sección 3.2.1, para la conversión de identificadores como se muestra a continuación:

```
> source("eutiles.R")
> email<-"user@institution.org"
> sal<-ELink(dbfrom="protein", db="gene", id="29462", email=email)
> data.frame(sal$Data)
  proteinID   dbFrom  dbTo  geneID
1     29462  protein  gene    6678
```

El "`id=29462`" de la base de datos **PROTEIN_GI_ACCESSION** especificada como fuente de origen (`dbfrom="protein"`), es convertido a un ID del tipo **EntreZ Gene ID** especificado por `dbfrom="genes"`. La salida de la conversión se almacena en el objeto `sal$Data`, el cual puede trasformarse en un objeto del tipo `data.frame`

para su fácil manipulación. En este ejemplo, la conversión fue realizada con éxito obteniendo el "`geneID=6678`".

El usuario también puede utilizar **ESummary** para obtener la información de anotación para un ID en particular. No obstante, es recomendable emplear un mecanismo del tipo *HTTP POST* cuando la consulta se realiza para una serie de `ids`, de manera que tanto la lista de `ids` como la respuesta se envíen utilizando el cuerpo del HTML, como se muestra en el ejemplo de invocación a `ESummaryHttpPost`:

```
> ids<-c("6678", "4444", "6963", "414100", "433190", "74103","70458")
> out<-ESummaryHttpPost(id=ids, db="gene", email=email)
> names(out$Data)
 [1] "GeneID"             "Name"               "Description"
 [4] "Orgname"            "Status"             "CurrentID"
 [7] "Chromosome"         "GeneticSource"      "MapLocation"
[10] "OtherAliases"       "OtherDesignations"  "NomenclatureSymbol"
[13] "NomenclatureName"   "NomenclatureStatus" "TaxID"
[16] "Mim"                "GenomicInfo"        "GeneWeight"
[19] "Summary"            "ChrSort"            "ChrStart"
```

La información de anotación se almacena en el objeto `out$Data`, que contiene 21 campos entre los cuales es posible nombrar: el `GeneID`, el símbolo (`Name`) y sus alias (`OtherAliases`), el nombre (`Description`), el estado actual (`Status`) y el ID vigente (`CurrentID`), entre otros. Por simplicidad, se muestran a continuación sólo cinco columnas para la consulta realizada:

```
> out$Data[, c("GeneID", "Name", "Status", "CurrentID", "Description")]
  GeneID          Name Status CurrentID
1   6678         SPARC      0         0
2   4444          MSK4      2         0
3   6963      TRBV/OR9      2         0
4 414100 D830029A09Rik      1     74103
5 433190     LOC433190      1     70458
6  74103          Nebl      0         0
7  70458 2610318N02Rik      0         0
```

```
                                                Description
1 secreted protein, acidic, cysteine-rich (osteonectin)
2   antigen identified by monoclonal antibody A123/A127
3 T cell receptor beta variable orphans on chromosome 9
4                             RIKEN cDNA D830029A09 gene
5                       similar to RIKEN cDNA 2610318N02
6                                             nebulette
7                             RIKEN cDNA 2610318N02 gene
```

En la primera fila del objeto `out$Data`, es posible ver la información de anotación del gen "`GeneID=6678`" convertido previamente con **ELink**. Lo novedoso en esta salida es la columna "`Status`". En esta columna se codifica el estado de los genes en: **0**) si es el **vigente**, **1**) si es **obsoleto** y **2**) si **no codifica** para ninguna proteína. En este ejemplo, el ID correspondiente a la primera fila se encuentra vigente, mientras que los dos siguientes han sidos removidos dado que no codifican. Por otra parte, el cuarto y quinto ID (`414100` y `433190`) son obsoletos (`Status=1`) por lo que en la columna `CurrentID` se presentan los IDs que los remplazan (`74103` y `70458` respectivamente). Por último, en las dos filas siguientes se muestra que dichos registros son de hecho los vigentes para los respectivos genes. De manera que es posible utilizar las columnas de `Status` y `CurrentID` de forma iterativa y programática, para recorrer el historial de los genes para obtener la información del último estado conocido. Así, a través de la metodología propuesta, es posible abordar la problemática de consistencia e integridad de IDs descripta en la sección 2.3.2. Más aún, el usuario tiene conocimiento adicional del estado de los identificadores, información no provista por otras herramientas de conversión, situación que le permite tener trazabilidad de las proteínas/genes de interés en todo momento.

Por otra parte, también es posible realizar una búsqueda mediante `ESearch` de la misma manera que se realizó para el caso de Uniprot en el módulo de proteómica (sección 3.2.1), con las correspondientes modificaciones como se muestra a continuación:

```
> query<-"SPARC[gene] AND \"Homo sapiens\"[Organism]"
> sal<- ESearch(ref="Sparc",db="gene",term=query)
> sal
```

```
$Query
  "SPARC[gene]+AND+%22Homo+sapiens%22[Organism]"
$Data
    ref   Count      Id
"Sparc"     "1"  "6678"
```

donde la búsqueda (`term=query`) realizada con `ESearch`, debe ser especificada sobre una base de datos específica (`db="gene"`). Adicionalmente, se incluye una referencia opcional (`ref="Sparc"`), la cual es de utilidad cuando se realiza una búsqueda de más de una proteína/gen de forma simultánea. La salida de la consulta se almacena en el objeto `sal`, que es una lista de dos elementos. El primer elemento, `$Query`, posee la conversión de `query` en formato URL compatible como se describió en la sección 3.2.1. El segundo elemento de `sal`, `$Data`, almacena el resultado de la consulta propiamente dicho. En ella se aprecia que existe una sola coincidencia (`Count=1`) para la búsqueda realizada, la cual apunta al `Id=6678`. Este ID efectivamente coincide con el criterio de búsqueda mostrado en la primera fila de los resultados de `ESummaryHttpPost` y los correspondientes al módulo de proteómica de la sección 3.2.1.

3.2.4. Comentarios finales

Contar con una metodología de trabajo que permita indagar en forma automática la conversión y estado actual de los ID, como la propuesta en esta tesis, es de gran ayuda para la comunidad científica. Esto impacta positivamente en el propio proceso de MD sobre la búsqueda de patrones y en las herramientas que se puedan utilizar sobre éstos para extraer conocimiento biológico. Consecuentemente, el investigador puede acceder a toda la información disponible. De esta manera se cuenta con trazabilidad y se conoce el estado actual de cada proteína/gen, permitiendo adoptar diferentes estrategias para continuar el análisis (de forma independiente) para aquellos IDs que no puedan continuar el flujo de trabajo habitual. Por ejemplo, permite utilizar un paquete de anotación de microarreglos para los IDs del fabricante incompletos en su propio archivo de anotación. La gran ventaja es que ahora se conoce de forma fehaciente hasta qué punto del flujo de trabajo han sido utilizados, mientras que las herramientas actualmente disponibles no reportan la pérdida de información.

La limitación que posee esta metodología se restringe a la anotación que se encuentre disponible. Ésta a su vez no se encuentra a cargo del usuario, dado que se accede a la fuente más actual disponible en internet mediante *E-utiles*. Esta particularidad es una enorme ventaja, ya que evita que el usuario tenga la responsabilidad de bajar cientos de gigabytes de información de bases de datos e instalar motores de bases de datos locales. Si aún así no fuese posible encontrar el Entrez Gene ID, es decir, falle la conversión/actualización, es posible recuperar la anotación. Esto último no esta implementado en esta tesis, pero si fuese necesario el usuario puede utilizar la información de secuencia de origen (proteínas u oligonucleótidos) y realizar una nueva identificación en tándem desde `blast.ncbi.nlm.nih.gov/Blast.cgi`, a través de un alineamiento de secuencias con **BLAST** (McGinnis y Madden, 2004).

3.3. Exploración multivariada y control de calidad

Una vez finalizada la etapa de consistencia e integridad de identificadores, se cuenta con una tabla con datos de expresión y anotación. A esta tabla se le aplican diversos filtros de anotación y calidad de señal como se presentó en la sección 2.3.3. Sin embargo, los abordajes clásicos no consideran que los experimentos de las diferentes ciencias "ómicas" como por ejemplo la *proteómica*, *transcriptómica*, *metabolómica* o *genómica*, tienen una **naturaleza multivariada**. Justamente, las tecnologías modernas nos permiten explorar una gran parte del proteoma o incluso todo el genoma, en donde cada proteína/gen es en esencia una variable explorada, para dilucidar su relación con algún resultado. Estos experimentos cada vez están incluyendo un número mayor de factores experimentales en el diseño (tiempo, dosis, etc.), o incluso información sujeto específica tales como la edad, sexo, linaje, etc.

En este contexto, la búsqueda de patrones relacionados con el diseño experimental, desde una perspectiva de la MD, debe realizarse mediante algún enfoque multivariado. Los enfoques más comunes son a través de análisis de componentes principales (**PCA**, Abdi y Williams (2010)) y regresión de mínimos cuadrados parciales (**PLS**, Geladi y Kowalski (1986)). Sin embargo, es reconocido en la comunidad científica que trabajar directamente con la matriz de expresión, puede enmascarar información de interés. Consecuentemente, la descomposición basada en análisis

de la varianza (**ANOVA**, Walpole et al. (1999)), se está volviendo popular para dividir las diferentes fuentes de variabilidad, antes de aplicar tales enfoques multivariados. Trabajos seminales en genómica fueron los de De Haan et al. (2007) en ANOVA-PCA (**APCA**) y Smilde et al. (2005) en modelos ANOVA-SCA (del inglés, ANOVA-simultaneous component analysis o **ASCA**). Sin embargo, la implementación de APCA en lenguaje R sólo está disponible para datos de espectros (*Spectra*) en el paquete `ChemoSpec` (Hanson, 2012). En cuanto a ASCA, no existe paquete en R para este modelo y sólo se encuentra disponible como una colección de funciones a partir de la traducción del código de MATLAB® de Nueda et al. (2007). Más aún, ASCA sólo acepta hasta tres matrices de diseño binarias, lo que limita su uso y hace que sea difícil su aplicación. Por otra parte, las estimaciones de coeficientes no ofrecen inferencia estadística sobre ellos, dado que se basan en el cálculo de medias utilizando las matrices de diseño.

En esta tesis se desarrolló un paquete R llamado `lmdme` (del inglés linear model decomposition for designed multivariate experiments, Fresno y Fernández (2013a)), para la descomposición ANOVA basada en modelos lineales (sección 2.3.4). Una amplia gama de modelos pueden ser especificados, de acuerdo con el diseño experimental, mediante una interfaz flexible para especificar la `formula` correspondiente. Debido a que los coeficientes se estiman por medio de *máxima verosimilitud* (Graybill, 2000), la significación estadística se ofrece de forma natural. A través de la metodología propuesta, es posible explorar de forma multivariada la existencia de patrones de correlación existentes en los datos que puedan deberse a efectos no esperados, de manera de evaluar la calidad del experimento o buscar patrones de proteínas/genes relacionados al diseño experimental planteado, mediante un análisis de PCA y/o PLS sobre los resultados de la descomposición ANOVA. Para ello, se provee de diferentes representaciones gráficas como *biplots*, *screeplots*, etc. (Peña, 2002).

En las siguientes secciones se presenta el **modelo**, al igual que su **aplicación** para la *exploración multivariada* de patrones y de evaluación de *control de calidad* en experimentos basados en tecnología de microarreglos. No obstante, esta implementación es adecuada para el análisis sobre las matrices de expresión obtenidas en experimentos de alto rendimiento, tales como geles *2D-DIGE*, *RNA-seq*, etc.

3.3.1. El modelo

Una explicación detallada de la descomposición ANOVA y análisis multivariado puede encontrarse en Smilde et al. (2005) y Zwanenburg et al. (2011). Sin pérdida de generalidad, se considera un experimento de *microarreglos* donde la expresión de los $(G_1,\ G_2,\ \ldots\ ,\ G_g)$ genes es medida, bajo un diseño experimental con dos factores principales: A, con a niveles $(A_1,\ A_2,\ \ldots\ ,\ A_i,\ \ldots\ ,\ A_a)$ y B, con b niveles $(B_1,\ B_2,\ \ldots\ ,\ B_j,\ \ldots\ ,\ B_b)$, con $R_1,\ R_2,\ \ldots\ ,\ R_k,\ \ldots\ ,\ R_r$ replicas para cada combinación de niveles $A \times B$.

Luego de preprocesar los datos como se describe en el capítulo 2 y sección 3.1, cada microarreglo/chip puede ser representado por un vector columna de mediciones de niveles de expresión de dimensión $g \times 1$. Consecuentemente, la totalidad de los datos del experimento puede ser expresado en una matriz de expresión (X) de dimensión $g \times n$, donde $n = a \times b \times r$ es la cantidad de microarreglos. Bajo este esquema de datos, en cada fila de la matriz X se encuentra la expresión de un único gen a través de las diferentes combinaciones de tratamientos $(A_i \times B_j)$, como se ilustra en la figura 3.4.

Independientemente de la generación de los datos, el modelo ANOVA aplicado a cada gen (fila) de X puede ser expresado como (3.1):

$$x_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i \times \beta_j + \varepsilon_{ijk} \tag{3.1}$$

donde x_{ijk} es la medición de expresión para "algún" gen, bajo la combinación "ij" de los factores A y B de la k-ésima réplica; μ es la media global; α, β y $\alpha \times \beta$ son los efectos principales y de interacción respectivamente; siendo el término de error $\varepsilon_{ijk} \sim N(0, \sigma^2)$. A su vez, (3.1) también puede ser expresada de forma matricial para todos los genes según (3.2):

$$X = X_\mu + X_\alpha + X_\beta + X_{\alpha\beta} + E = \sum_{l \in \{\mu,\alpha,\beta,\alpha\beta\}} X_l + E \tag{3.2}$$

donde X_l, E son matrices de dimensión $g \times n$ y contienen las medias correspondientes al l-ésimo término y el error aleatorio respectivamente. Sin embargo, en el contexto de modelos lineales, X_l puede ser reescrita como una combinación lineal mediante la

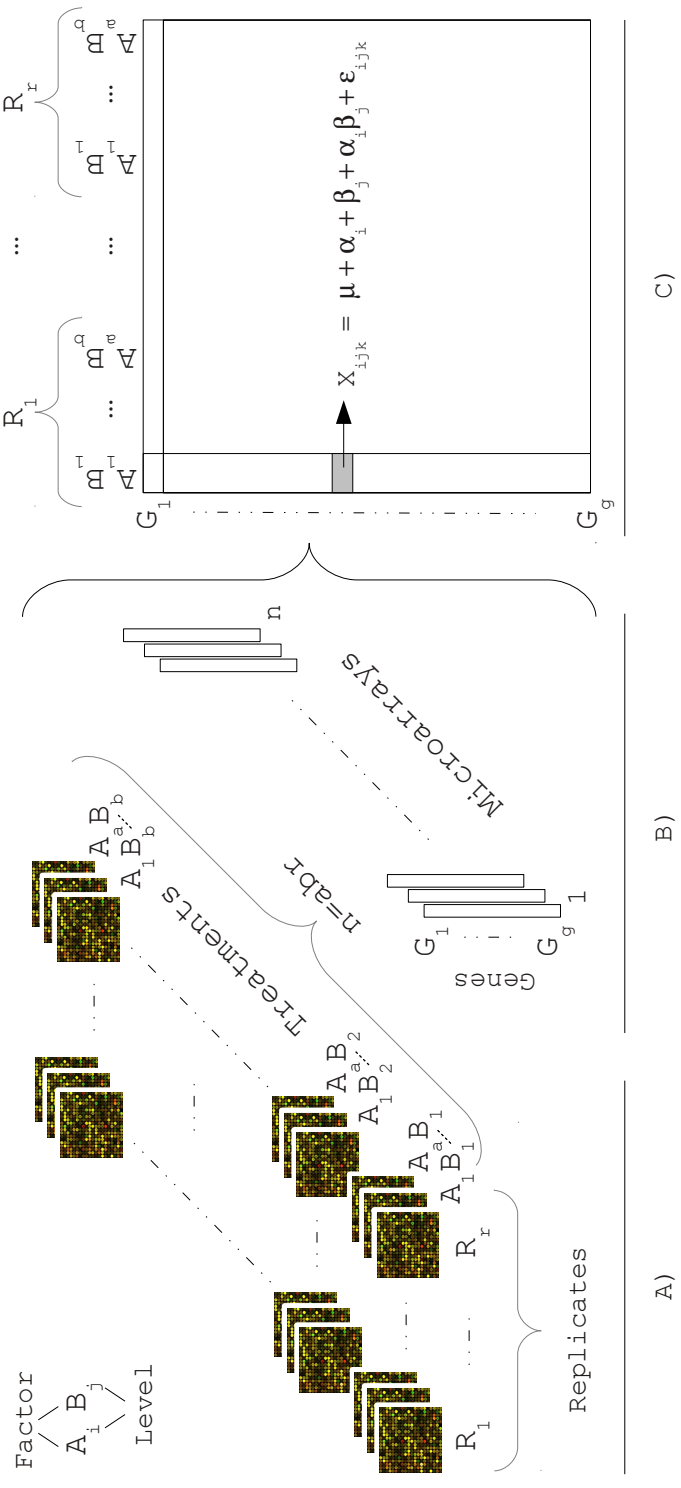


Figura 3.4: Representación de los datos de expresión de microarreglos. A) Esquema de los niveles de expresión de cada gen, para cada combinación de tratamientos A_iB_j y sus réplicas R_k, dando un total de $n = a \times b \times r$ chips. B) Los valores de expresión de cada chip (microarreglo), son representados como un vector columna. C) Los vectores columnas son agrupados dando lugar a la matriz de expresión X. Así, en una fila de la matriz X se almacenan los valores de expresión de un gen dado, para la combinaciones de todos los tratamientos. Entonces, las mediciones de una fila se someten al modelo ANOVA (3.1). Imagen extraída de Fresno et al. (2014).

multiplicación de dos matrices como se expresa en (3.3):

$$X = \sum_{l \in \{\mu,\alpha,\beta,\alpha\beta\}} X_l + E = \sum_{l \in \{\mu,\alpha,\beta,\alpha\beta\}} B_l Z_l^T + E = B_\mu Z_\mu^T + \ldots + B_{\alpha\beta} Z_{\alpha\beta}^T + E = \mu 1^\top + B_\alpha Z_\alpha^T + \ldots + B_{\alpha\beta} Z_{\alpha\beta}^T + E \quad (3.3)$$

donde B_l y Z_l son conocidas en la literatura como las matrices de *coeficientes* y de *modelo* con dimensión $g \times m_{(l)}$ y $n \times m_{(l)}$ respectivamente, con $m_{(l)}$, el número de niveles del factor l. Usualmente, el primer término es llamado *intercepto*, con $B_\mu = \mu$ y $Z_\mu = \mathbf{1}$ de dimensión $g \times 1$ y $n \times 1$ respectivamente. En este ejemplo, todas las matrices Z_l son binarias, permitiendo identificar si una medición pertenece ("1") o no ("0") al factor correspondiente.

En la implementación de Smilde et al. (2005) y Nueda et al. (2007), la estimación de la matriz de coeficientes se basa en cálculos de *promedios*, utilizando hasta tres matrices de diseño $Z_{\alpha,\beta,\alpha\beta}$, identificando los valores a promediar, para descomponer por completo la matriz original como se muestra en (3.1). Por el contrario, en la implementación de esta tesis, la estimación de los coeficientes del modelo se realiza de forma iterativa, utilizando un abordaje por *máxima verosimilitud*, mediante la función `lmFit` disponible en el paquete `limma` (Smyth et al., 2011). Consecuentemente, tres características no presentes hasta la fecha son incorporadas:

- Potencialmente cualquier modelo puede ser especificado utilizando una **interfaz flexible** para definir la `formula` del modelo correspondiente. El usuario sólo necesita proveer: i) la matriz de expresión X, ii) un `data.frame` (`design`) con la estructura de tratamientos del diseño experimental, y iii) especificar el modelo a través de un objeto de tipo `formula`, de la misma manera que habitualmente lo hace, mediante la función `lm` provista en R. Internamente, una invocación a la función `model.matrix`, automáticamente construirá las matrices Z apropiadas. Esto permite superar la restricción en la cantidad de factores en el diseño experimental, al igual que la tediosa definición de dichas matrices.

- **Pruebas de hipótesis** sobre las matrices de coeficientes $\hat{B}_l$. Una prueba T se realiza automáticamente para el s-ésimo gen, a los efectos de comprobar si el o-ésimo coeficiente es igual a cero o no, es decir, $H_0 : b_{so} = 0$ vs $H_1 : b_{so} \neq 0$.

Adicionalmente, una prueba F se realiza para determinar si en forma conjunta, todos los coeficiente b_{so} son o no iguales a cero, es decir, $H_0 : b_{s1} = b_{s2} = \ldots = b_{so} = 0$ vs H_1 : algún $b_{so} \neq 0$.

- **Corrección empírica de Bayes**. También es posible utilizar la función `eBayes` del paquete `limma`, para reducir las varianzas de las muestras en cada gen/fila hacia un valor común y permitir aumentar los grados de libertad para la estimación de las varianzas individuales, como se describe en Smyth (2004).

De Haan et al. (2007) estimaron los efectos principales y de interacción mediante restas sobre la media global como en un ANOVA tradicional (Walpole et al., 1999). Consecuentemente, los genes deben ser tratados como un factor adicional, mientras que, en las implementaciones de Smilde et al. (2005) y Nueda et al. (2007), la estimación es realizada gen a gen como en (3.1). De manera que, en un experimento de dos factores, como por ejemplo *tiempo* × *oxígeno*, en el modelo de De Haan et al. se incluyen dos interacciones dobles y una triple, dado que los genes son tratados como un factor, a diferencia de los modelos de Smilde et al. (2005) y Nueda et al. (2007)

Descomposición ANOVA

El modelo ANOVA (3.2) se descompone de forma iterativa utilizando (3.3), donde para cada paso se estiman las l-ésimas matrices $\hat{B}_l$, $\hat{E}_l$ y el vector de varianzas $\hat{\sigma}_l^2$. Luego, la matriz de contribución de un término particular $\hat{X}_l = \hat{B}_l Z_l^\top$ se sustrae de los residuos precedentes para formar la matriz a descomponer en el próximo paso, como se muestra en (3.4):

$$
\begin{aligned}
X &= X_\mu + X_\alpha + X_\beta + X_{\alpha\beta} + E = \sum_{l\in\{\mu,\alpha,\beta,\alpha\beta\}} X_l + E \\
\textit{paso } \mu: \quad X &= X_\mu + E_\mu \Rightarrow X = \hat{B}_\mu Z_\mu^\top + \hat{E}_\mu \Rightarrow \hat{E}_\mu = X - \hat{B}_\mu Z_\mu^\top \\
\textit{paso } \alpha: \quad E_\mu &= X_\alpha + E_\alpha \Rightarrow \hat{E}_\mu = \hat{B}_\alpha Z_\alpha^\top + \hat{E}_\alpha \Rightarrow \hat{E}_\alpha = \hat{E}_\mu - \hat{B}_\alpha Z_\alpha^\top \\
&\vdots \quad \vdots \\
\textit{paso } l: E_{l-1} &= X_l + E_l \Rightarrow \hat{E}_{l-1} = \hat{B}_l Z_l^\top + \hat{E}_l \Rightarrow \hat{E}_l = \hat{E}_{l-1} - \hat{B}_l Z_l^\top \\
&\vdots \quad \vdots \\
\textit{paso } \alpha\beta: \quad E_\beta &= X_{\alpha\beta} + E \Rightarrow \hat{E}_\beta = \hat{B}_{\alpha\beta} Z_{\alpha\beta}^\top + \hat{E} \Rightarrow \hat{E} = \hat{E}_\beta - \hat{B}_{\alpha\beta} Z_{\alpha\beta}^\top
\end{aligned}
\tag{3.4}
$$

donde el sombrero ("^") denota coeficientes/residuos estimados. En esta implementación, el primer paso siempre estima el término del *intercepto*, es decir, `formula=~1` en código R, con $\hat{B}_\mu = \hat{\mu}$ y $Z_\mu = 1$. Los modelos siguientes sólo incluirán el l-ésimo factor sin el intercepto, es decir, `formula=~lth_term-1`, donde `lth_term` refiere a α, β o $\alpha\beta$ en este ejemplo. Este procedimiento es bastante similar al propuesto por Harrington et al. (2005).

Análisis multivariado: PCA y PLS

Estos métodos explican la estructura de varianza/covarianza de un conjunto de observaciones (proteínas, genes, etc.) a través de una cantidad reducida de combinaciones lineales de las variables, por ejemplo, condiciones experimentales. Ambos métodos pueden ser aplicados sobre el l-ésimo paso de descomposición ANOVA de (3.4), abordando diferentes aspectos:

- **PCA** modeliza la estructura de *varianza* de una única matriz, usualmente con el objetivo principal de reducción e interpretación de los datos. Dependiendo de la matriz a la cual se aplica, da lugar a dos posibles métodos: **ASCA**, cuando PCA es aplicado a la matriz de *coeficientes*, $\hat{B}_l$ (Smilde et al., 2005); y **APCA** cuando PCA es calculado sobre los *residuos*, $\hat{E}_{l-1}$. El último es conceptualmente un ASCA y usualmente es aplicado a, $X_l + E$, es decir, las medias de la matriz de factor X_l, sumado al error del modelo totalmente descompuesto E en (3.1),

como en De Haan et al. (2007).

- **PLS** no solo generaliza, sino que también combina características de PCA y regresión para explorar la estructura de *covarianza* entre una matriz de entrada y una de salida, como se describe por Abdi y Williams (2010) y Shawe-Taylor y Cristianini (2004). PLS es particularmente útil cuando una o varias variables dependientes (salidas - O) deben ser predichas a partir de una gran cantidad de variables independientes (entradas) potencialmente con elevada correlación. En esta implementación, las entradas pueden ser la matriz de *coeficientes* $\hat{B}_l$ o los *residuos* $\hat{E}_{l-1}$. Dependiendo de la elección, la matriz de salida será una matriz diagonal `O=diag(nrow(`$\hat{B}_l$`))` o la matriz de diseño $O = Z_l$ respectivamente. Adicionalmente, el usuario puede especificar su propia matriz de salida, O, para verificar una hipótesis particular. Por ejemplo, en genómica funcional puede ser la matriz de clases de GO (sección 1.1.1) como se utiliza en GSEA (sección 1.2.1) por Subramanian et al. (2005).

Smilde et al. (2005) sugiere que se debe tener en cuenta el número esperado de los componentes en X, es decir, el rango de la matriz dado el número de réplicas por nivel de tratamiento. Esta sugerencia surge de que la propia aproximación de Smilde et al. (2005) genera una matriz X con muchos datos constantes (columnas con los mismos valores) debido a las réplicas y que por ende no son informativos. No obstante, en la presente implementación se trabaja con la matriz de *coeficientes*, razón por la cual el usuario no tendrá que preocuparse por dicho número, dado que los componentes se encuentran directamente resumidos $\hat{B}_l$. Adicionalmente, el paquete `lmdme` (Fresno y Fernández, 2013a) ofrece diferentes alternativas de visualización para PCA/PLS, por ejemplo, `biplot`, `loadingplot` and `screeplot` (Peña, 2002), o estimación de los `leverage` (palancas), a los efectos de filtrar genes/filas como se realiza en Tarazona et al. (2012).

3.3.2. Evaluación

En esta sección se presentan dos aplicaciones concretas del paquete `lmdme` desarrollado en esta tesis (Fresno et al., 2014; Fresno y Fernández, 2013a) donde se evalúan, desde una perspectiva de la MD, las diferentes funcionalidades que el paquete

presenta. En la primera de ellas, se aplica a la búsqueda de patrones de interacción de expresión en genes, donde se hace foco sobre la **definición del modelo**, **descomposición ANOVA**, **análisis de PCA/PLS** y **visualización** de los resultados. En la segunda aplicación, la metodología es utilizada como estrategia de **control de calidad** en datos obtenidos de tecnologías de alto rendimiento, aplicado a un conjunto de datos de microarreglos de ADN (sección 2.3.1). A partir de aquí, algunas salidas han sido removidas por razones de claridad y los comandos han sido ejecutados utilizando `options(digits=4)`.

Búsqueda de patrones de interacción

Prado-Lopez et al. (2010) estudiaron la diferenciación de células madres de embriones humanos bajo hipoxia. El conjunto de datos originales se encuentra disponible en Gene Expression Omnibus (Edgar et al., 2002), con el número de acceso GSE37761 y como paquete de R llamado `stemHypoxia` (Fresno y Fernández, 2013d), disponible en el repositorio de Bioconductor (`www.bioconductor.org`).

Prado-Lopez et al. (2010) midieron la expresión de genes a diferentes tiempos, bajo condiciones controladas de oxígeno. Este experimento posee un típica estructura de ANOVA a dos vías donde el factor A representa el "*tiempo*", con $a = 3$ niveles $\{0{,}5;\ 1;\ 5$ días$\}$, el factor B representa la concentración de "*oxígeno*", con $b = 3$ niveles $\{1,\ 5,\ 21\,\%\}$, y $r = 2$ réplicas biológicas, dando un total de 18 muestras. El remanente de los datos ha sido removido, a los efectos de tener un diseño balanceado, como lo sugiere Smilde et al. (2005) para cumplir con los supuestos de ortogonalidad de la descomposición ANOVA.

Primero, es necesario cargar el paquete `stemHypoxia` para poder acceder al objeto de R que posee los datos, mediante el comando `data("stemHypoxia")`. Esta invocación deja disponibles en el espacio de trabajo, el diseño experimental (`design`) y los niveles de expresión de los genes en el objeto `M`.

```
> library("stemHypoxia")
> data("stemHypoxia")
```

Ahora se debe manipular el objeto `design`, para solamente dejar los tratamientos que generan un diseño balanceado. Luego, es posible cambiar los nombres de las filas de `M` (`rownames(M)`) para que se correspondan con cada `M$Gene_ID`.

```
> timeIndex<-design$time %in% c(0.5, 1, 5)
> oxygenIndex<-design$oxygen %in% c(1, 5, 21)
> design<-design[timeIndex & oxygenIndex, ]
> design$time<-as.factor(design$time)
> design$oxygen<-as.factor(design$oxygen)
> rownames(M)<-M$Gene_ID
> M<-M[, colnames(M) %in% design$samplename]
```

Una vez seleccionados los niveles correspondientes, la matriz de expresión `M` resultante es de dimensión $g = 40736$ filas (genes) y $n = 18$ columnas (muestras/microarrays). Por otra parte, el data.frame llamado `design` contiene las columnas con los efectos principales (*time* y *oxygen*), al igual que el nombre de las muestras (`samplename`). Luego, es recomendable explorar las cabeceras de estos dos objetos invocando a la función `head` como se muestra a continuación:

```
> head(design)
```

```
  time oxygen samplename
3  0.5      1    12h_1_1
4  0.5      1    12h_1_2
5  0.5      5    12h_5_1
6  0.5      5    12h_5_2
7  0.5     21   12h_21_1
8  0.5     21   12h_21_2
```

```
> head(M)[, 1:3]
```

```
             12h_1_1 12h_1_2 12h_5_1
A_24_P66027    7.182   7.512   8.225
A_32_P77178    6.385   6.035   6.440
A_23_P212522   9.562   9.390   9.211
A_24_P934473   6.288   6.397   6.265
A_24_P9671    12.007  11.995  12.282
A_32_P29551   10.176   9.273   9.360
```

Una vez terminada la adecuación de los datos experimentales, se debe cargar la librería invocando `library("lmdme")`. Esta instrucción automáticamente carga en memoria los paquetes requeridos: `limma` (Smyth et al., 2011) y `pls` (Mevik et al., 2011). Luego, la descomposición ANOVA de la sección 3.3.1 puede ser llevada a cabo utilizando (3.4), invocando a la función `lmdme` especificando como parámetros la formula en `model`, el conjunto de datos en `data` y el diseño experimental en `design`:

```
> library("lmdme")
> fit<-lmdme(model=~time*oxygen, data=M, design=design)
> fit

lmdme object:
Data dimension:  40736  x  18
Design (head):
  time oxygen samplename
3  0.5      1    12h_1_1
4  0.5      1    12h_1_2
5  0.5      5    12h_5_1
6  0.5      5    12h_5_2
7  0.5     21   12h_21_1
8  0.5     21   12h_21_2

Model:~time * oxygen
Model decomposition:
  Step       Names            Formula CoefCols
1    1 (Intercept)                ~ 1        1
2    2        time        ~ -1 + time        3
3    3      oxygen      ~ -1 + oxygen        3
4    4 time:oxygen ~ -1 + time:oxygen        9
```

El resultado de `lmdme` es almacenado dentro del objeto `fit`, el cual es una clase S4 de R. Invocando al objeto `fit`, es posible tener una pequeña descripción de los datos (`data`), el diseño utilizado (`design`), así también como el modelo (`Model`) aplicado y un resumen de la descomposición ANOVA realizada. Este último `data.frame`

describe las formulas aplicadas (`Formula`), el nombre (`Names`) para cada paso (`Step`), y la cantidad de coeficientes estimados para cada gen (`CoefCols`).

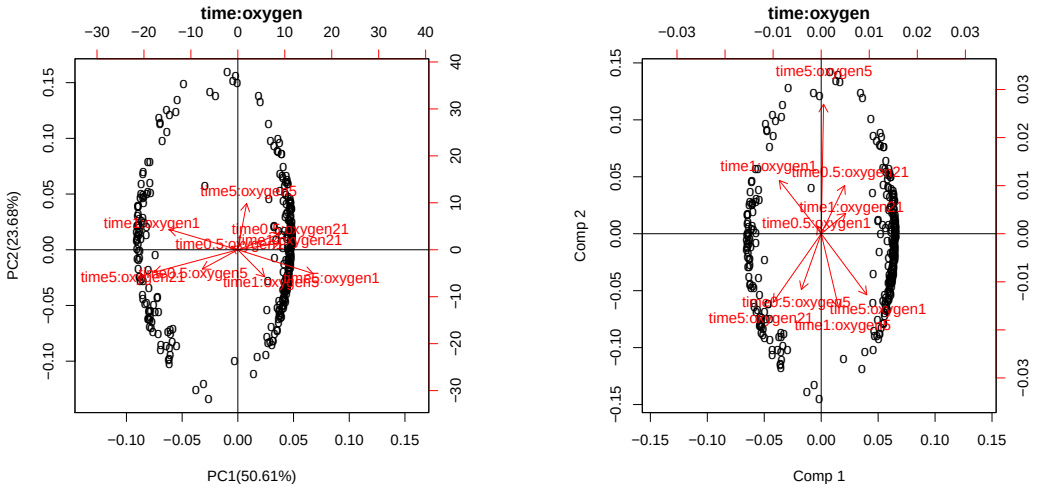
Completada la descomposición ANOVA, es posible elegir aquellos genes en los cuales al menos un coeficiente de interacción es estadísticamente diferente de cero (prueba F sobre los coeficientes) con un valor $p < 0{,}001$. Sobre estos genes es factible realizar un análisis tipo ASCA sobre los coeficientes (`type="coefficient"`) del término de interacción (`term="time:oxygen"`), y/o PLS (`decomposition="plsr"`) contra la matriz identidad de salida (opción por defecto).

```
> id<-F.p.values(fit, term="time:oxygen")<0.001
> decomposition(fit, decomposition="pca", type="coefficient",
+    term="time:oxygen", subset=id, scale="row")
> fit.plsr<-fit
> decomposition(fit.plsr, decomposition="plsr", type="coefficient",
+    term="time:oxygen", subset=id, scale="row")
```

Estas instrucciones realizan un análisis PCA y PLS (`decomposition`) sobre la versión escalada (`scale="row"`) de los 305 genes seleccionados (`subset=id`), almacenando los resultados en los objeto `fit` y `fit.plsr` respectivamente. Adicionalmente, se ha especificado de forma explícita `type="coefficient"` (valor por defecto), para indicar que la descomposición de varianza/covarianza se debe realizar utilizando la matriz de coeficientes del término de interacción `term="time:oxygen"` ($\hat{B}_{\alpha\beta}$). Una vez obtenido los resultados, es posible visualizar los `biplot`s asociados en las figuras 3.5 (a) y (b):

```
> biplot(fit, xlabs="o", expand=0.7)
> biplot(fit.plsr, which="loadings", xlabs="o",
+    ylabs=colnames(coefficients(fit.plsr, term="time:oxygen")),
+    var.axes=TRUE)
```

En las figuras 3.5 las etiquetas de los genes (`rownames(M)`) han sido reemplazadas por símbolos (`xlabs="o"`) para claridad visual. A su vez, el segundo eje es escalado (`expand=0.7`), para evitar que las flechas queden fuera de la gráfica. Por otra parte, el `biplot` 3.5(b) correspondiente al análisis de PLS, ha sido modificado (`which="loadings"`) para obtener un gráfico similar al del ASCA de la figura



(a) ANOVA simultaneous component analysis (b) ANOVA partial least squares regresion

Figura 3.5: Biplot realizado sobre los coeficientes del término de interacción (`tiempo:oxígeno`), para genes que poseen un valor $p < 0{,}001$ para la prueba F correspondiente. Note que la matriz de interacción en el modelo ASCA es de rango 9-1. Por lo tanto, se esperan 9 flechas y los 305 genes seleccionados son proyectados en espacio de las primeras dos componentes principales de la figura 3.5. Imágenes extraídas de Fresno et al. (2014).

3.5(a). Consecuentemente, las etiquetas del eje "y" (`ylabs`) son modificadas para que coincidan con las correspondientes a los coeficientes (`coefficients`) del término de interacción (`term="time:oxygen"`), y `var.axes=TRUE` para que muestre las correspondientes flechas.

En la figura 3.5(a) se muestra que las dos primeras componentes biplot del análisis ASCA explican más del 70 % de la varianza de los coeficientes. A su vez, los genes están dispuestos en una forma elíptica, y se puede observar que algunos de ellos tienden a interaccionar con diferentes combinaciones de tiempo y oxígeno. Un comportamiento similar se aprecia en el biplot del análisis por PLS de la figura 3.5(b).

El efecto de interacción del objeto `fit`, también puede ser visualizado utilizando la función `loadingplot` como se muestra en la figura 3.6. En ella se aprecia que para cada combinación de dos niveles consecutivos de factores (tiempo y oxígeno)

existe un efecto de interacción en la primera componente principal, la cual explica el 50.61 % de la totalidad de la varianza del término "`time:oxygen`".

```
> loadingplot(fit, term.x="time", term.y="oxygen")
```

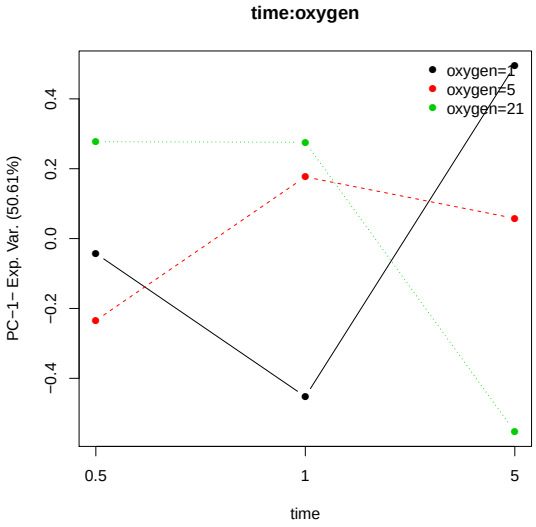


Figura 3.6: `loadingplot` del ANOVA-SCA para los genes que poseen un valor $p < 0{,}001$ para la prueba F realizada sobre los coeficientes del término de interacción (*tiempo* × *oxígeno*). Imagen extraída de Fresno et al. (2014).

En el caso que el usuario desee realizar un análisis de tipo APCA, sólo tiene que modificar el parámetro `type="residuals"` en la invocación a la función `decomposition`, y realizar una exploración similar, como se muestra en el siguiente ejemplo de aplicación de `lmdme` sobre microarreglos.

Control de calidad en microarreglos

Los datos utilizados en esta aplicación *no se encuentran publicados* aún, pero están disponibles en el sitio web `www.bdmg.com.ar` y en el anexo digital de la sección A.2. Se agradece especialmente al grupo del Dr. Osvaldo Podhajcer, del Laboratorio de Terapia Molecular y Celular de la Fundación Instituto Leloir, por permitir el uso de sus datos en la presente tesis.

En esta oportunidad se utilizaron *microarreglos* de dos colores (sección 2.3.1) para explorar los perfiles de expresión génica. Los niveles de expresión se midieron en diferentes puntos de tiempo, bajo diversas concentraciones de proteínas incluidas en los medios de cultivos independientes de una línea celular de melanoma. Este experimento también posee una estructura ANOVA a dos vías: donde el factor A representa el "*tiempo*" con $a = 3$ niveles $\{0{,}5;\ 4;\ 12\ horas\}$, el factor B representa la "*concentración*" con $b = 3$ niveles $\{0;\ 1;\ 10\ unidades\}$ para $r = 3$ replicas biológicas, dando un total de 27 muestras.

El grupo de investigación que generó los datos que se analizan en este ejemplo, conoce de experimentos previos que existen genes que presentan interacción de los factores *tiempo* × *concentración*. En particular, los investigadores están interesados en encontrar aquellos genes con una expresión diferencial con un valor $p < 0{,}05$ para la prueba F asociada al término de interacción. Resultados preliminares de análisis realizados por el grupo, utilizando el paquete `limma` (Smyth et al., 2011), no revelaron la existencia de patrones de interacción.

En este contexto, se mostrará que mediante un abordaje desde la MD utilizando el paquete `lmdme`, es posible identificar efectos técnicos inesperados que podrían dar una interpretación biológica sesgada. Justamente, esto no es viable mediante los análisis tradicionales, dado que no permiten realizar una exploración multivariada de los datos. Adicionalmente se demuestra cómo remover dicho artefacto, aplicando la librería desarrollada en esta tesis.

Una vez más, es necesario cargar la librería `lmdme` y los datos experimentales, los cuales han sido previamente guardados en un archivo. Invocando la instrucción `load(file="example2.RData")`, son cargados en memoria los objetos correspondientes al diseño experimental (`design`) y la matriz de expresión (`M`). Siempre es recomendable explorar estos objetos, para comprobar si se han cargado de forma correcta, utilizando la función `head`, de la misma manera que se realizó para los datos de `stemHypoxia` en el ejemplo anterior.

```
> library("lmdme")
> load(file="example2.RData")
> head(design)
```

```
  Time Conc SampleName HybridDate
1  0.5    0 221732.gpr        nov
2  0.5    0 338515.gpr        jan
3  0.5    0 339577.gpr        feb
4  0.5    1 221678.gpr        nov
5  0.5    1 338514.gpr        jan
6  0.5    1 339576.gpr        feb
```

```
> head(M)[, 1:3]
```

```
     221732.gpr 338515.gpr 339577.gpr
[1,]     0.1287     0.1181    0.72294
[2,]    -0.1653    -0.1080    0.10825
[3,]    -0.5227    -0.2300   -0.29959
[4,]     0.3142     0.5636    0.07366
[5,]     0.1519     0.2008   -1.10059
[6,]     0.2542    -0.1083   -0.40284
```

La dimensión de la matriz `M` es de $g = 2520$ filas (genes) y $n = 27$ columnas (muestras/microarreglos). A su vez, el `data.frame` que posee el diseño experimental (`design`) contiene las columnas de los efectos principales: tiempo (`Time`) y `Conc` para la concentración, el nombre de las muestras (`SampleName`) y la fecha en la cual los chips fueron hibridizados (`HybridDate`).

Invocando la función `lmdme` es posible ajustar el modelo utilizando el parámetro `model=`$\sim$`Time*Conc`, con una corrección empírica de Bayes (`Bayes=TRUE`) y el parámetro `verbose=TRUE`, para darle al usuario una realimentación sobre el progreso de la descomposición ANOVA. Adicionalmente, es factible comprobar si los resultados obtenidos por el grupo de investigación acerca de la inexistencia de genes que interactúan, son correctos o no.

```
> fit<-lmdme(model=~Time*Conc, data=M, design=design, Bayes=TRUE,
+    verbose=TRUE)
```

```
testing:   ~ 1
```

```
testing:   ~ Time -1
testing:   ~ Conc -1
testing:   ~ Time:Conc -1
```

```
> id.fit<-F.p.values(fit, term="Time:Conc")<0.05
> sum(id.fit)
```

```
[1] 0
```

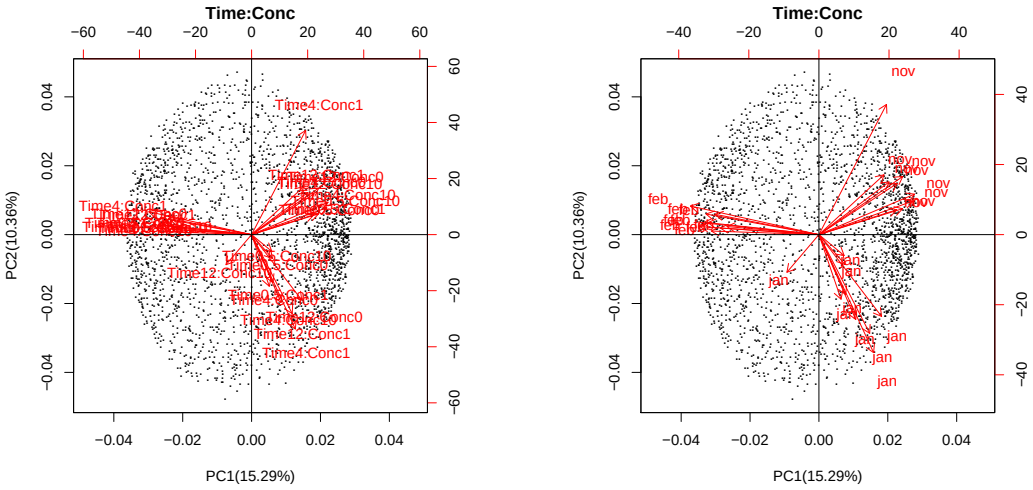
El resultado de `sum(id.fit)` igual a 0 es coincidente con los resultados previamente obtenidos por los investigadores. Sin embargo dado que se esperaban genes que interactúen, el resultado sugiere una exploración en profundidad de los datos. En este contexto, un abordaje del tipo APCA puede ser aplicado al objeto `fit`, para realizar una exploración visual del `biplot` del término `term="Time:Conc"` representado en la figura 3.7(a).

```
> decomposition(fit, "pca", scale="row", type="residual")
> biplot(fit, term="Time:Conc", xlabs='.', expand=0.9)
```

En la figura 3.7(a) se observa la presencia de un patrón correspondiente a una fuente de variabilidad no controlada, que pareciera agrupar los chips en tres grupos. La inspección del objeto `design` revela la existencia de una columna llamada `HybridDate`, no incluida en el modelo, que pueda estar relacionada con el agrupamiento que se observa en el `biplot` correspondiente. Para observar si esto es así, se pueden cambiar las etiquetas de las flechas del `biplot` de la figura 3.7(a) de manera tal de identificar a cada chip por su fecha de hibridación, utilizando el siguiente código:

```
> biplot(fit, term="Time:Conc", ylabs=design$HybridDate, xlabs='.',
+    expand=0.8)
```

Claramente, la figura 3.7(b) muestra que existiría una asociación entre los tres agrupamientos y la fecha de hibridización, es decir, los tres agrupamientos poseen etiquetas de fecha similares. Esta es una fuente de variabilidad no considerada en el



(a) Original `biplot` (b) Using `ylabs=design$HybridDate`

Figura 3.7: `biplot` del ANOVA-PCA de los residuos del término de interacción (`Time:Conc`). Imagen extraída de Fresno et al. (2014).

modelo original. De acuerdo con la evidencia, el usuario puede proponer un análisis del tipo PLS. En éste es posible definir una matriz de salida (`Omatrix`) personalizada para preguntar si los datos responden o no, a la estructura de los datos. Para ello, se define la estructura con ayuda de la función `model.matrix` utilizando como fórmula `~HybridDate-1` y la información del objeto `design`.

```
> decomposition(fit, "plsr", scale="row", type="residual",
+     term="Time:Conc", Omatrix=model.matrix(~HybridDate-1, design))
> biplot(fit, term="Time:Conc", which="loadings", xlabs='.',
+     var.axes=TRUE)
```

Una exploración visual del `biplot` de la figura 3.8, demuestra que la fecha de hibridación responde al patrón encontrado con anterioridad. Una conversación posterior con los autores de los datos, reveló que el experimento original tenía planificado realizar las hidridizaciones de las tres réplicas el mismo día. No obstante, debido a restricciones aduaneras en las importaciones, debió modificarse por cada recepción de grupo de chips. Así, por cada recepción se hibridizó una réplica para todas las combinaciones de tratamientos. Casualmente, la primera recepción fue en el mes de

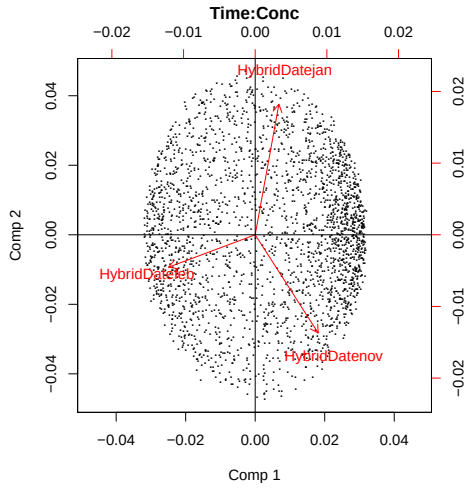


Figura 3.8: `biplot` del análisis PLS realizado sobre los residuos de interacción (`Time:Conc`), utilizando la fecha de hibridación como matriz de salida. Imagen extraída de Fresno et al. (2014).

noviembre (`nov`), la segunda en enero (`jan`) y la última en febrero (`feb`). La confirmación obtenida a partir de la exploración de los datos, junto con las restricciones en la aleatorización del experimento, sugieren que la variable `HybridDate` debe ser incluida en el modelo:

```
> fit.date<-lmdme(model="~HybridDate+Time*Conc, data=M, design=design,
+    Bayes=TRUE)
> id.fit.date<-F.p.values(fit.date, term="Time:Conc")<0.05
> sum(id.fit.date)

[1] 13
```

Así, con la inclusión de `HybridDate` en el modelo, es posible estimar y remover este efecto. Consecuentemente, la inferencia estadística sobre genes ha sido modificada revelando a 13 genes candidatos, afectados por los niveles de interacción de *tiempo* × *concentración*. Adicionalmente, el `biplot` del APCA ilustrado en la figura 3.9, muestra que el patrón observado en la figura 3.7(a) ha sido removido con éxito (no se aprecian tendencias).

```
> decomposition(fit.date, "pca", scale="row", type="residual")
> biplot(fit.date, term="Time:Conc", xlabs='.', expand=0.8)
```

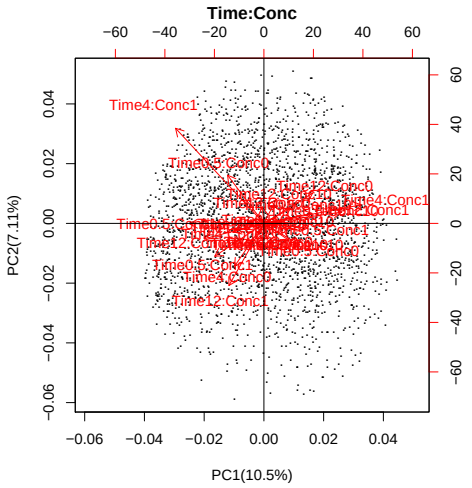


Figura 3.9: `biplot` del ANOVA-PCA sobre los residuos del término de interacción (`Time:Conc`), incluyendo la fecha de hibridación en el modelo. Imagen extraída de Fresno et al. (2014).

3.3.3. Comentarios finales

Considerando los ejemplos anteriormente analizados, puede notarse que la descomposición ANOVA en experimentos con diseños multifactoriales, a través de modelos lineales, ha demostrado ser una alternativa válida como herramienta de MD para la búsqueda de patrones multivariados en datos de expresión ómicos. En este sentido, la metodología propuesta tiene en cuenta la **información del diseño experimental**, permitiendo descomponer las diferentes **fuentes de variabilidad**, para su posterior exploración **multivariada**. En este contexto, se propone realizar un análisis de la estructura de varianza/covarianza de las matrices de coeficientes y/o residuos mediante PCA y/o PLS. Cabe destacar que este último método es novedoso para este tipo de aplicaciones, y permite indagar por la existencia (o no) de relaciones funcionales de interés, definidas por el usuario.

A través de las diferentes visualizaciones provistas en el paquete, es posible explorar la variabilidad que introducen los diferentes niveles de tratamientos controlados en el experimento. Mas aún, ha sido posible comprobar la *existencia de fuentes de variabilidad tecnológica* no tenidas en cuenta en el diseño experimental original y posterior remoción. La aplicación de abordajes univariados tradicionales, en el mismo contexto, no permitiría detectar este tipo de artefactos. Justamente, en el segundo ejemplo se mostró cómo la no remoción de estas fuentes *impacta en el inferencia estadística*, con la consecuente *interpretación biológica errónea de los datos*. En este sentido, el paquete `lmdme` (Fresno et al., 2014; Fresno y Fernández, 2013a) ha demostrado ser una alternativa válida para el *control de calidad* de datos de tecnologías de alto rendimiento. Este paquete se encuentra disponible para la comunidad científica en Bioconductor (`www.bioconductor.org`) y posee más de 2300 descargas según las estadísticas del repositorio (`bioconductor.org/packages/stats/bioc/lmdme.html`) desde su primera versión, en diciembre de 2012.

3.4. Conectividad al portal DAVID

Uno de los sistemas más accedidos por la comunidad científica para el análisis proteómico/genómico funcional, es la base de Datos para Anotación, Visualización y Descubrimiento Integrado (DAVID, Dennis Jr et al. (2003) y Huang et al. (2007)). Este recurso bioinformático tiene como objetivo proporcionar herramientas para la interpretación funcional de grandes listas de genes/proteínas (Huang et al., 2009b). A través de él, es posible realizar análisis de tipo SEA y/o MEA y luego explorar los resultados mediantes reportes de tipo *HTML* y gráficas bidimensionales con la evidencia existente entre *muchos-genes-a-muchos-términos*, como se describe en la sección 1.3.

El portal DAVID se accede principalmente a través de una página web (`david.abcc.ncifcrf.gov`). También existe una interfaz de programación para aplicaciones (API, del inglés application programming interface), basada en consultas mediante URLs para acceder a DAVID de forma programática. No obstante, la API en sí misma posee una serie de limitaciones entre las cuales es posible destacar: i) sólo funciona con la configuración predeterminada de DAVID, ii) la longitud del URL

está limitada a 2048 caracteres ($\leq$ 400 genes, dependiendo del ID), y iii) es posible realizar hasta 200 consultas por máquina, con 10 segundos de espera entre consultas. Pese a ello, la API puede ser utilizada a través de cualquier lenguaje para consultas livianas y en R existe el paquete llamado **DAVIDQuery** (Day y Lisovich, 2010) para tal fin.

Frente a las limitaciones de la API, se puso a disposición una interfaz de servicios web (*DAVID-WS*) para permitir el pleno acceso y control de todas las funciones (excepto la visualización), manteniendo algunas restricciones como por ejemplo la cantidad de consultas (Jiao et al., 2012). Si bien en la publicación original de Jiao et al. (2012) se provee de una serie de clientes para diferentes lenguajes de programación (Java®, Perl®, Python® y MATLAB®), potencialmente puede utilizarse desde cualquier lenguaje. No obstante, desde R es posible utilizar DAVID-WS mediante el paquete **SOAP** (Lang, 2012), situación que requiere elevados conocimientos de programación. Más aún, los resultados de cada consulta son muy difíciles de manejar, ya que son objetos XML que el usuario deberá decodificar (paquete **XML**, Lang (2013b)) o incluso clases Java (paquete **rJava**, Urbanek (2013)) si se utiliza el cliente nativo provisto por DAVID-WS.

Dadas estas limitaciones y complejidades de utilización de DAVID-WS, es que en esta tesis se desarrolló un paquete R llamado **RDAVIDWebService** (Fresno y Fernández, 2013b,c) que permite una **acceso programático y versátil** a DAVID, evitando que el usuario sea un experto en programación para hacer uso de los reportes que esta herramienta provee. Mediante `RDAVIDWebService`, es posible tener objetos nativos de R y expandir el análisis en uno de los lenguajes de programación más utilizados en bioinformática (R Core Team, 2013). Adicionalmente, el paquete supera las limitaciones de **visualización** de DAVID-WS, permitiendo obtener las habituales gráficas de evidencia de *muchos-genes-a-muchos-términos* provistas en DAVID. A esta capacidad de visualización se le incorpora la *vista de grafos de GO* (no disponible en DAVID), la cual es utilizada en otras herramientas, como se describió en la sección 1.3.5.

En las siguientes secciones se presentan las características de **implementación** del desarrollo junto con sus diferentes **funcionalidades** y dos **ejemplos de aplicación**. Uno aplicado a la *conectividad* con DAVID, donde se muestra cómo realizar

consultas desde R, y otro donde se *exploran y visualizan* los resultados obtenidos del análisis ontológico-funcional.

3.4.1. Implementación

El servicio web de DAVID se basa en una topología cliente-servidor para publicar las diferentes funcionalidades (Jiao et al., 2012). En la práctica, el desarrollo original de Jiao et al. (2012) está basado en lenguaje Java, y utiliza un mecanismo para invocar métodos de manera remota llamado **RMI** (del inglés *Remote Method Invocation*). Este mecanismo permite comunicación en aplicaciones distribuidas exclusivamente en Java. A través de RMI, un programa como DAVID puede exportar un objeto (usualmente llamado `skeleton`), permitiendo que dicho objeto sea accesible a través de la red y permanezca a la espera de peticiones de internet utilizando una dirección y un puerto TCP específico (del inglés *Transmission Control Protocol*). A partir de ese momento, desde el lado del cliente es posible acceder a las diferentes funcionalidades utilizando una interfaz (usualmente llamada `stub`), la cual puede conectarse e invocar los métodos proporcionados por el objeto `skeleton`. No obstante, también es posible conectarse utilizando otros lenguajes de programación, para lo cual debe utilizarse otras tecnologías como **CORBA** (del inglés *Common Object Request Broker Architecture*) o **SOAP** (del inglés *Simple Object Access Protocol*) en lugar de **RMI**, como lo hacen los diferentes clientes disponibles en la publicación original.

En la implementación del paquete `RDAVIDWebService`, se optó por utilizar el cliente nativo de Java provisto por Jiao et al. (2012) para acceder a las diferentes funcionalidades ofrecidas por DAVID-WS. De esta manera, el desarrollo tiene la ventaja de utilizar código estable y probado por los autores. Adicionalmente, cualquier nueva funcionalidad que ofrezca DAVID-WS no impactará en el paquete R, ya que sólo será necesario cambiar el `stub` correspondiente para disponer de dicha funcionalidad. No obstante, será necesario implementar una interfaz entre los dos lenguajes, R y Java. En R el paquete consta de dos módulos: uno de *conectividad* entre R y Java y otro que modela en *objetos nativos de R*, los resultados de DAVID.

Módulo de conectividad

El **módulo de conectividad** se basa en la clase `DAVIDWebService`, que es la interfaz desde R hacia DAVID-WS. Esta clase se encuentra implementada a través del paradigma *R5*, también conocido como "*clase de referencia*". Este paradigma permite crear clases persistentes en R, para mantener un único punto de conexión con el servidor. Para ello, la clase `DAVIDWebService` a su vez utiliza el paquete **rJava** (Urbanek, 2013) para establecer una comunicación con el cliente de Java, `DAVIDWebServiceStub`, quien es el responsable de establecer la comunicación con su contraparte Java (`skeleton`) en el servidor de DAVID. Así, mediante la clase `DAVIDWebService`, es posible acceder a DAVID-WS para realizar el **flujo de trabajo habitual**:

1. **Subir los identificadores** de genes/proteínas y lista de referencia.
2. **Comprobar el estado de DAVID** en lo que respecta a los genes/proteínas reconocidos por el sistema, búsqueda de las categorías disponibles, etc.
3. **Seleccionar la lista de referencia/especies y categorías** para utilizar en el presente análisis.
4. **Obtener los diferentes reportes** SEA/MEA en los que se incluyen la tabla de análisis funcional, agrupamiento de genes/términos, entre otros.

Sin embargo, el módulo de conectividad posee ciertas **restricciones** entre las cuales es posible nombrar:

- Un usuario o computadora puede realizar hasta 200 consultas en un día.
- El agrupamiento (clustering) de genes/términos puede incluir hasta un máximo de 3000 genes/términos.
- El equipo de DAVID se reserva el derecho de suspender cualquier uso indebido de DAVID-WS sin previo aviso.

Estas limitaciones no son impuestas por el paquete `RDAVIDWebService`, sino que son propias de DAVID, como se indica en la página `david.abcc.ncifcrf.gov/content.jsp?file=WS.html`.

El diagrama de clases del módulo de conexión entre R y Java se muestra en la figura 3.10. En ella se aprecia cómo la clase `DAVIDWebService` realiza sus peticiones a DAVID-WS, a través de la clase `DAVIDWebServiceStub`. Esta clase devuelve objetos nativos de Java, los cuales deben ser transformados en estructuras de datos de R, para su posterior utilización. Este proceso consume un tiempo de cálculo considerable (5-90 min.), debido a todos los controles internos que realizan ambos lenguajes. Como estrategia de reducción del tiempo de importación (<5 min.) se implementó una clase Java llamada `DAVIDParser` (figura 3.10), que genera un archivo de texto plano temporal que posee la misma estructura que los reportes generados en la página web de DAVID. Justamente, esta estrategia permite importar los resultados a R y **guardar localmente los reportes** para su posterior análisis/exploración. Esta es una *característica novedosa*, dado que al tener la **misma estructura de archivos**, es posible utilizar tanto los reportes obtenidos *desde R*, o los generados desde el *sitio web* de forma indistinta (no importa dónde se generaron). Por otra parte, todos los

Figura 3.10: Diagrama de clases del módulo de conectividad entre R y Java. `DAVIDWebService` utiliza el paquete **rJava** para acceder/controlar el servicio web de DAVID a través de la clase `DAVIDWebServiceStub` y decodificar los reportes hacia R mediante la clase `DAVIDParser`. Note que por simplicidad se han omitido las firmas de las funciones. Imágenes extraídas de Fresno y Fernández (2013c).

análisis realizados en DAVID a la fecha, deben hacerse utilizando una conexión a internet. Ahora el investigador puede **compartir sus resultados** con otros grupos de investigación y trabajar de forma colaborativa (*archivos de reportes de DAVID* u *objetos de R*), evitando así la engorrosa carga de datos a DAVID cada vez que se quiere realizar un análisis. Es decir, se pueden *explorar los resultados* del análisis ontológico-funcional **sin conectividad a internet**.

Módulo de objetos de R

El **módulo de objetos nativos de R** proporciona un **marco de trabajo uniforme** (en inglés se conoce como *framework*), para acceder directamente a las funcionalidades de DAVID desde R, sin la necesidad de decodificación *ad hoc* de los resultados de cada tipo de reporte. En este sentido, este módulo es el responsable de importar los resultados obtenidos por `DAVIDWebService` en las apropiadas clases *S4* según el análisis realizado.

En la figura 3.11 se muestra la jerarquía de clases del módulo, donde se aprecia que todos los resultados son genéricamente un `DAVIDResult`. Dependiendo de la precedencia del reporte, esta clase se irá especializando. En la izquierda de la figura se aprecia cómo los resultados obtenidos para un análisis de tipo *MEA* mediante un agrupamiento común en `DAVIDCluster`, terminan en las clases `DAVIDTermCluster` o `DAVIDGeneCluster` en caso de realizar un análisis de tipo "*Functional Annotation Cluster*" o un "*Gene Functional Classification*" desde DAVID, respectivamente. Por otra parte, en la región central de la figura 3.11 se muestran las clases `DAVIDGenes` y `DAVIDFunctionalAnnotationChart` las cuales poseen una herencia múltiple con el tipo de datos base de R denominado `data.frame`, dado que en esencia modelan a una estructura de este tipo. A su vez, la clase `DAVIDGenes` sirve de contenedor en la clase `DAVIDFunctionalAnnotationTable` la cual representan su equivalente reporte en DAVID.

La jerarquía de clases de R no sólo permite modelar el comportamiento de los diferentes reportes de DAVID, sino que también permite adaptar el comportamiento de diferentes funciones para incorporar dos capacidades de **visualización** no disponibles en DAVID-WS:

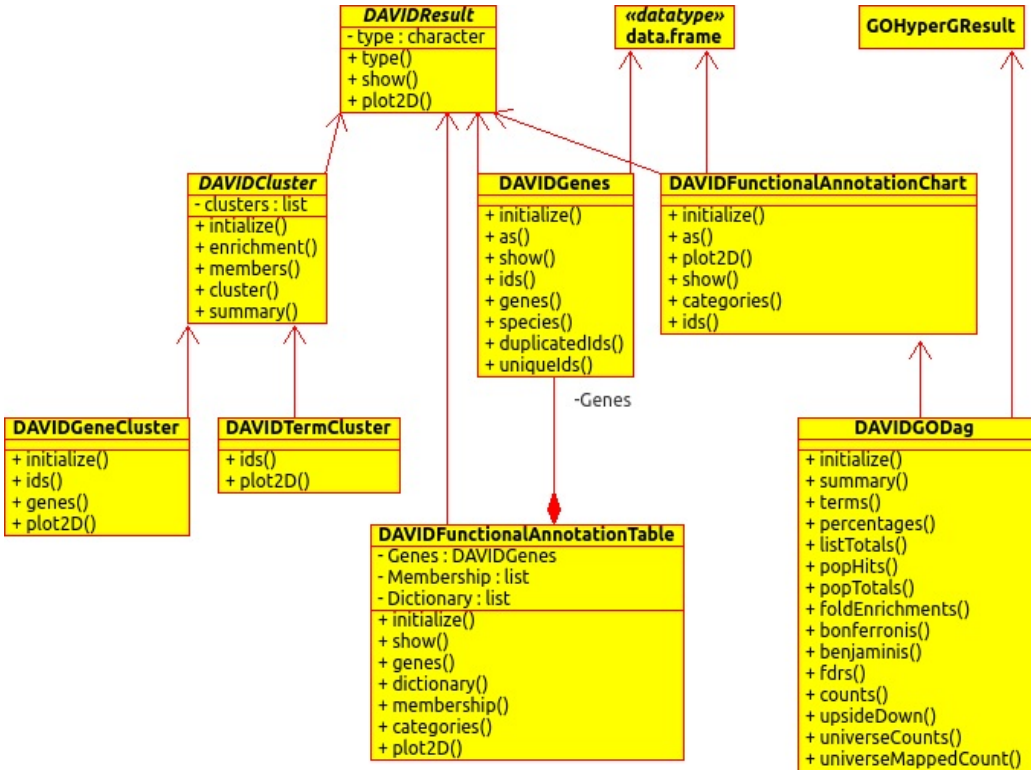


Figura 3.11: Diagrama de clases del módulo de R. Note que por simplicidad se han omitido las firmas de las funciones. Imagen extraída de la *vignette* de Fresno y Fernández (2013c)

Muchos-genes a muchos-términos: ahora están disponibles en R las habituales gráficas del sitio web de DAVID, de evidencias existentes entre términos y genes presentadas en la sección 1.3.5. Estas gráficas se encuentran disponibles invocando a la función `plot2D`, la cual utiliza toda la potencia del paquete **ggplot2** (Wickham, 2009). Adicionalmente, incorpora la función `plot2D` a las clases `DAVIDFunctionalAnnotationChart/Table`, como se observa en el diagrama de clases de la figura 3.11. Cabe destacar que ésta es una funcionalidad no disponible en el sitio web de DAVID. Ahora desde R el usuario posee la flexibilidad de seleccionar sólo aquellas relaciones de términos y genes que desee mostrar, frente a la habitual representación de resultados de tipo MEA (`DAVIDTerm/GeneCluster`) del sitio web (también disponible).

Grafos de Gene Ontology: los resultados de un reporte de tipo "Functional Annotation Chart" pueden ser representados en la estructura de grafos dirigidos acíclicos (GDA) de GO (Ashburner et al., 2000), de la misma manera que lo realizan otras herramientas como *GOMiner* y *GOstats* presentado en la sección 1.3.5. Para ello, se muestra en la figura 3.11 cómo la clase `DAVIDFunctionalAnnotationChart` es especializada por `DAVIDGODag`, quien a su vez hereda de la clase `GOHyperResult`. A través de ellas, los resultados obtenidos de DAVID son convertidos a una estructura de datos compatible con el paquete **GOstats** (Falcon y Gentleman, 2007). De esta manera, es posible **construir el GDA de enriquecimiento funcional** de las diferentes categorías principales de GO: procesos biológicos, funciones moleculares y/o componentes celulares, presentadas en la sección 1.1.1.

La clase `DAVIDGODag` permite **visualizar los valores EASE** en el contexto del GDA de GO, extendiendo las capacidades de DAVID desde una perspectiva de la MD. Utilizar la propia estructura de GO ha resultado ser una estrategia válida para resumir la información, en comparación con la habitual búsqueda en extensas tablas, problemática presentada en el capítulo 1. Esto no sólo permite la búsqueda de patrones funcionales de forma visual, sino que facilita la exploración de los resultados, como se verá en la sección 3.4.2.

3.4.2. Evaluación

En esta sección se describen dos aplicaciones típicas del paquete `RDAVIDWebService`. La primera de ellas se relaciona con la **conectividad** y manejo desde R sobre todas las funcionalidades disponibles por DAVID-WS (carga de identificadores, obtención de reportes, etc.). En el segundo ejemplo se muestra cómo se puede utilizar el paquete para **explorar los resultados**, independientemente de dónde hayan sido obtenidos (desde el sitio de DAVID o de forma programática con R).

Ejemplo de conectividad

Antes de poder usar el paquete `RDAVIDWebService`, el usuario debe registrar su e-mail institucional llenando el formulario provisto en la página `david.abcc.`

ncifcrf.gov/webservice/register.htm, para poder utilizar DAVID-WS. Una vez registrado, el usuario puede crear un objeto `DAVIDWebService` y establecer una conexión. A continuación, es posible subir la/s lista/s de identificadores indicando un nombre, tipo (genes o referencia) y clase de identificador. En este ejemplo, se utiliza la lista de identificadores proporcionada en el sitio web de DAVID (demoList1 con identificadores de Affymetrix®).

***Nota**: el siguiente código no funcionará a menos que cambie "user@inst.org" por el e-mail de una cuenta de usuario previamente registrada en DAVID.*

```
> library("RDAVIDWebService")
> david<-DAVIDWebService$new(email="user@inst.org")
> data(demoList1)
> result<-addList(david, demoList1,idType="AFFYMETRIX_3PRIME_IVT_ID",
+ listName="demoList1", listType="Gene")
> result
$inDavid
[1] 0.9695122
$unmappedIds
[1] "34902_at"   "1937_at"    "35996_at"   "32163_f_at" "32407_f_at"
```

La salida de `result` muestra que 96.95 % de la totalidad de la lista `demoList1` es reconocida en DAVID (`$inDavid`). Adicionalmente, este objeto contienen los cinco ids no mapeados (`$unmappedIds`). Por otra parte, el estado de la conexión es guardada en el objeto `david` y puede ser consultada en cualquier momento:

```
> david
DAVIDWebService object to access DAVID's website.
User email: user@inst.org
Available Gene List/s:
       Name Using
1 demoList1     *
Available Specie/s:
               Name Using
1 Homo sapiens(155)     *
```

```
Available Background List/s:
          Name Using
1 Homo sapiens     *
```

La salida del objeto `david` muestra a los 155 genes correspondientes a la especie *Homo sapiens* presentes en la lista de genes llamada `demoList1`. Adicionalmente, se muestra que el genoma completo de la especie es seleccionado por defecto como lista de referencia. No obstante, el usuario puede subir una lista de genes y modificar el tipo de lista a `listType="Background"`, para especificar uno personalizado.

Por otra parte y en caso de que lo requiera, el usuario también puede seleccionar las categorías de anotación de su interés para realizar el análisis, como por ejemplo solamente con `GOTERM_BP_ALL`, `GOTERM_MF_ALL` y `GOTERM_CC_ALL`, como se muestra a continuación:

```
> setAnnotationCategories(david, c("GOTERM_BP_ALL", "GOTERM_MF_ALL",
+   "GOTERM_CC_ALL"))
```

Una vez establecida la configuración, es decir, lista de genes candidatos, referencia y categorías de interés, el usuario puede comenzar a realizar el análisis. Para ello, puede solicitar los diferentes reportes para su uso inmediato o para guardarlos en archivos para su posterior uso. Por ejemplo, se puede obtener el agrupamiento de términos ("*Functional Annotation Clustering*" en DAVID) y guardarlo en el objeto `termCluster`, o guardar los resultados en el archivo `termClusterReport1.tab` invocando los siguientes comandos:

```
> termCluster<-getClusterReport(david, type="Term")
> getClusterReportFile(david, type="Term",
+   fileName="termClusterReport1.tab")
```

En este caso, las dos alternativas mencionadas han sido puestas en práctica. A su vez, es posible obtener el agrupamiento de genes modificando el parámetro `type="Genes"`, o invocar alguna otra funcionalidad descripta en la ayuda del paquete, dependiendo del análisis que desee realizar el usuario, como se muestra en el siguiente ejemplo.

Ejemplo de exploración

En lo sucesivo, se utilizarán los reportes correspondientes a la utilización de la lista de genes llamada `demoList1`, guardados en el paquete `RDAVIDWebService`. No obstante y sin pérdida de generalidad, es posible utilizar los resultados contenidos en el objeto `termCluster` provenientes del ejemplo de conectividad.

Una vez más, el usuario debe cargar la librería para poder utilizar las diferentes funcionalidades. A continuación, es posible cargar los resultados del agrupamiento de términos almacenados en el archivo `termClusterReport1.tab`, e inspeccionarlos utilizando el siguiente código:

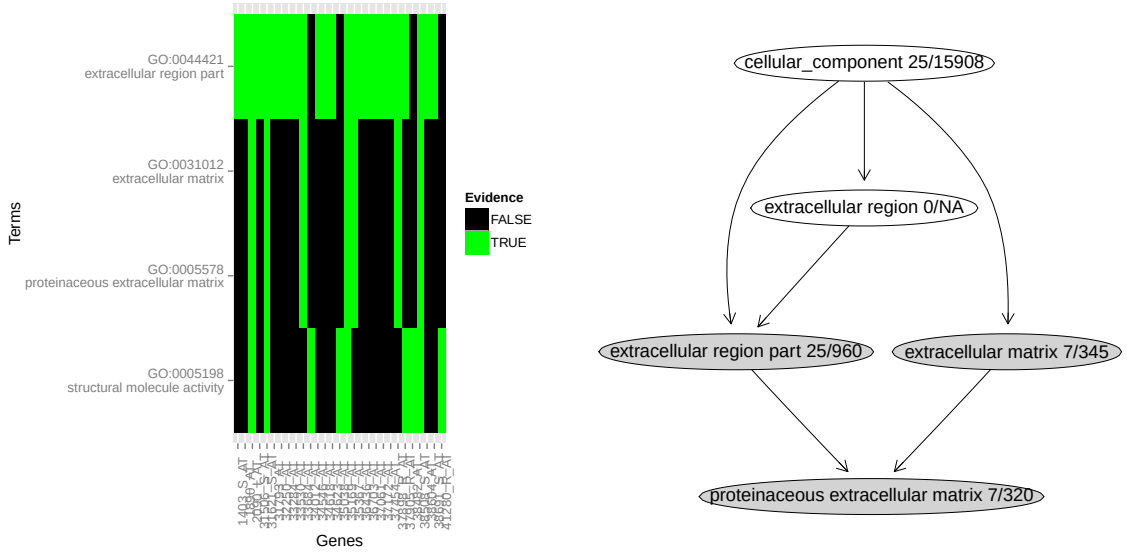
```
> library("RDAVIDWebService")
> fileName<-system.file("files/termClusterReport1.tab.tar.gz",
+    package="RDAVIDWebService")
> untar(fileName)
> termCluster<-DAVIDTermCluster(untar(fileName, list=TRUE))
> termCluster

DAVID Result object
Result type:  AnnotationCluster
Number of cluster:  28

> head(summary(termCluster))

  Cluster Enrichment Members
1       1      2.904      14
2       2      2.135       4
3       3      2.059      10
4       4      1.977      14
5       5      1.501       4
6       6      1.347       4
```

El objeto `termCluster` es una instancia de la clase `DAVIDTermCluster` con el correspondiente reporte (`AnnotationCluster`), para la lista de genes llamada `demoList1`.



(a) Evidencia de términos/genes

(b) Grafo de enriquecimiento

Figura 3.12: Exploración del segundo agrupamiento de anotaciones (a) Vista de muchos-términos a muchos-genes (b) Grafo de GO inducido por los términos enriquecimiento en (a). Los términos/nodos con un puntaje $EASE < 0{,}1$ se muestran en gris. Adicionalmente se muestra el cociente entre los genes de la lista contra los presentes en la referencia. Note que en los casos donde no se cuenta con la información del nodo (por no pertenecer al grupo) se introducen "`NA`" (dato no disponible). Imágenes obtenidas de Fresno y Fernández (2013c).

En este análisis se han encontrado 28 grupos, y es posible inspeccionar de manera superficial la cabecera del resumen de información de dicho objeto invocando a `head(summary(termCluster))`. Esta salida muestra un objeto de tipo `data.frame`, donde se aprecia el puntaje (`Enrichment`) obtenido en cada grupo (`Cluster`) y la cantidad de miembros que hay en cada uno de ellos (`Members`). Adicionalmente, se pueden explorar de forma visual las relaciones existentes entre los términos y genes de un grupo particular, por ejemplo el número 2:

```
> clustNumber<-2
> plot2D(termCluster, clustNumber)
```

En la figura 3.12(a) se aprecia que los cuatro términos de este agrupamiento comparten todos los genes en "*extracellular region part*" (fila superior). Sin embargo,

a medida que descendemos hacia la fila inferior ("*structural molecule activity*") sólo nueve genes poseen evidencia relacionada a ellos. Este tipo de representación no utiliza la estructura jerárquica de GO. Tampoco es posible discriminar si los términos se encuentran enriquecidos o no, dado que por defecto se utilizan todos los términos asociados a los genes de `demoList1`.

En este sentido, `RDAVIDWebService` permite extender la capacidad de análisis de DAVID mediante la transformación de los resultados obtenidos en el agrupamiento y construir el GDA de enriquecimiento con la clase `DAVIDGODag`, para un nivel de significancia dado (`pvalueCutoff=0.1`):

```
> davidGODag<-DAVIDGODag(members(termCluster)[[clustNumber]],
+    pvalueCutoff=0.1, "CC")
```

En el ejemplo se utiliza la categoría de componentes celulares ("`CC`"), pero ello no restringe a que se pueda realizar sobre otra de las categorías de GO (PB o FM). Mediante este abordaje es posible utilizar la propia estructura de GO para dar contexto a los resultados de enriquecimiento y explorarlas de forma visual invocando a la función `plotGOTermGraph` del paquete *GOstats*:

```
> plotGOTermGraph(g=goDag(davidGODag),
+    r=davidGODag, max.nchar=40, node.shape="ellipse")
```

En la figura 3.12(b) se muestra el GDA obtenido utilizando los datos del agrupamiento número dos. En esta figura se destacan los términos enriquecidos en color gris y el cociente entre los de genes de `demoList1` y los pertenecientes al genoma, en aquellos casos donde se cuente con la información necesaria. De esta manera, la propia estructura de GO permite dar contexto y resumir funcionalmente al agrupamiento, situación que no es posible ver en la figura 3.12(a). En este ejemplo, definitivamente la evidencia del agrupamiento apunta a un enriquecimiento de la matriz extra celular, más específicamente al término "proteinaceous extracellular matrix".

3.4.3. Comentarios finales

El paquete `RDAVIDWebService` (Fresno y Fernández, 2013b,c) ha demostrado ser una alternativa válida para la **conectividad programática** con el portal DAVID

desde R. En este sentido, provee una interfaz para realizar las mismas operaciones que se pueden realizar desde el sitio web de DAVID. Adicionalmente, brinda un **marco de trabajo uniforme** a través de diferentes objetos nativos de R, que permiten **importar los reportes** obtenidos desde R o incluso desde el mismo portal DAVID. A su vez, se ofrecen diferentes **alternativas de exploración visual** de los resultados, como se mostró en la figura 3.12.

Este paquete permite que los resultados del análisis funcional sean fácilmente importados a R y se encuentren listos para ser usados con el/los paquetes de CRAN (Hornik, 2012) o Bioconductor (Gentleman et al., 2005) favoritos del usuario. `RDAVIDWebService` se encuentra disponible para la comunidad científica en Bioconductor (`www.bioconductor.org`) y posee más de 450 descargas según las estadísticas del repositorio desde su primera versión en julio de 2013 (`bioconductor.org/packages/stats/bioc/RDAVIDWebService.html`).

3.5. Integración y contraste de múltiples referencias

En el contexto del análisis ontológico-funcional, la etapa de MD correspondiente a *modelado* puede realizarse mediante alguna de las diferentes metodologías como SEA, GSEA o MEA descriptas en la sección 1.2.1. No obstante, un análisis de tipo SEA se ve influenciado por la **selección de la lista de referencia** (LR) como se describió en la sección 1.2.2. Esto último impacta en el valor que toma la prueba estadística (*prueba exacta de Fisher, prueba* χ^2, etc.), con el consecuente *sesgo en la interpretación biológica* de los resultados, debido a una *elección inapropiada de la LR*. Frente a esta problemática, el flujo de análisis tradicional de la figura 3.2 propone una etapa de "*evaluación*" como estrategia para eliminar enriquecimiento espurio, basada en alguna de las alternativas de corrección por comparaciones múltiples como por ejemplo, a través de **FDR**, como se describió en la sección 2.1.2. Si bien esta alternativa ha mostrado ser de gran utilidad, no contempla la problemática de la selección de la LR, como tampoco nos da un indicio de **cuán sensible es el enriquecimiento** frente a esta elección.

En esta tesis se propone una metodología conocida como **MRCM** (del inglés *Multi-Reference Contrast Method*, Fresno et al. (2012)), para complementar los abor-

dajes tradicionales de la etapa de evaluación de MD. Esta metodología se basa en dos conceptos:

Integración y contraste de múltiples LRs. Una alternativa valiosa a la hora de *enriquecer y asistir* a los investigadores en la *exploración de los resultados de SEA*, es mediante la integración y contraste de los resultados obtenidos de múltiples LRs. De esta manera, la propia integración/contraste de la información es explotada para obtener **conocimiento biológico a través de las discrepancias/consensos**. Para ello, es posible utilizar la propia estructura de GO para resumir y explorar un **único GDA**, a través de un contraste visual basado en un patrón de colores para cada una de las categorías principales (PB, FM y CC).

Sensibilidad del enriquecimiento. Mediante simulaciones de tipo **bootstrap** (sección 2.5), es posible evaluar la robustez (sensibilidad) del enriquecimiento de cada término, frente a problemática de la elección apropiada de la LR. De esta manera, el investigador posee un valor indicador de la **potencia** de cada término, para **asistir la exploración y elección de términos** candidatos para la posterior *validación biológica*.

En las próximas secciones se presentan los dos análisis propuestos por la metodología. Adicionalmente, se presentan tres conjuntos de datos obtenidos de tecnologías de alto rendimiento: uno aplicado a proteómica basada en geles 2D-DIGE (sección 2.3.1), y dos de transcriptómica con microarreglos de un color (sección 2.3.1). Sobre estos datos se ponen en práctica los dos aspectos propuestos como aporte a la etapa de *evaluación de MD*, en el contexto del *análisis ontológico-funcional*.

3.5.1. Análisis de múltiples LRs

La metodología propuesta utiliza **DAVID** (ver sección 1.3, Dennis Jr et al. (2003) y Huang et al. (2007)) como motor para el cálculo del análisis tipo **SEA** de las múltiples LRs a *integrar/contrastar*.

El punto de partida del análisis considera que el usuario ha seguido las diferentes etapas del KDD y MD descriptas en el capítulo 2, hasta obtener una **lista de**

candidatos y una **LR** con las proteínas/genes presentes en el experimento, como se encuentra esquematizado en la sección 3.1. A partir de estas listas, el procesamiento sigue el diagrama de flujo propuesto en la figura 3.13, para lo cual se deben llevar a cabo los siguientes pasos:

1. **Subir** la lista de **candidatos** obtenidos en el experimento (por ejemplo, control vs tratamiento) a DAVID. Definir la LR para el análisis. La plataforma posee por defecto **LR-I** (el genoma) y **LR-II** (los genes de un microarreglo). Sin embargo, **LR-III** (proteínas/genes presentes en el experimento) debe ser subida por el usuario.

2. **Seleccionar la anotación** completa de Gene Ontology, para cada una de las categorías principales (*GOTERM_BP_ALL, GOTERM_MF_ALL y GOTERM_CC_ALL*).

3. **Obtener** los *reportes de anotación funcional* completos, para cada una de las LRs (LR-I, LR-II y LR-III). Es decir, todos los términos que tengan al menos una proteína/gen candidata y cuyo valor p para la prueba estadística sea menor o igual a uno. Esto equivale a seleccionar las opciones `Count=0` y `EASE=1` en las opciones de filtro avanzadas del sitio web de DAVID.

Los pasos anteriores pueden realizarse de forma manual, desde el sitio web de DAVID. No obstante, esto implica que para cada una de las tres LRs es necesario seleccionar las categorías de anotación y obtener los reportes de anotación funcional requeridos. Sin embargo, este proceso se puede realizar de forma programática con `RDAVIDWebService`, como se describió en la sección 3.4 bajo el recuadro de línea discontinua de la figura 3.13. Los resultados obtenidos del análisis funcional, ya sea desde el sitio web de DAVID o con `RDAVIDWebService`, se almacenan a nivel local y se procesan utilizando el lenguaje R (R Core Team, 2013), junto con diferentes paquetes de Bioconductor (Gentleman et al., 2004).

El primer procesamiento que se realiza sobre los resultados es la **identificación de los términos enriquecidos**. Esto requiere de la definición, por parte del usuario, de un umbral de enriquecimiento, utilizando lo que se conoce en DAVID como un valor **EASE** (de las siglas en inglés, Expression Analysis Systematic Explorer, Hosack

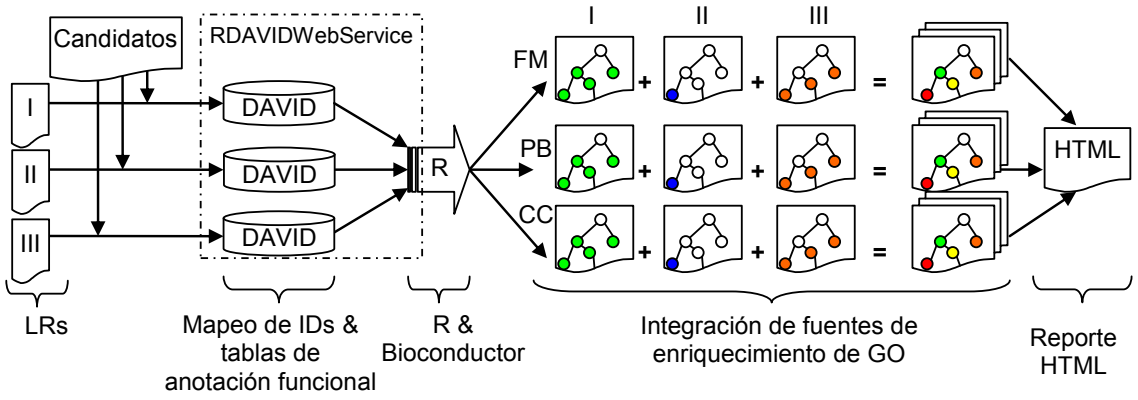


Figura 3.13: Diagrama de flujo del método de contraste de múltiples referencias. Adaptación de la imagen de Fresno et al. (2012).

et al. (2003)). Este último representa un valor p ajustado, donde se ha penalizado en una unidad a los candidatos que pertenecen a cada término (Término$_i$) de la tabla de contingencias 3.1. Es decir, en vez de que dicha celda contenga n_i candidatos como en la tabla 1.1 presentada en la sección 1.2.1, ahora se utiliza $n_i - 1$ para realizar la prueba exacta de Fisher, resultando en un valor llamado $EASE$. Huang et al. (2009b) sugieren utilizar un $EASE \leq 0{,}1$ para encontrar aquellos términos enriquecidos.

Una vez identificados los términos enriquecidos, es posible utilizar la estructura jerárquica de GO para representar visualmente los resultados. Para ello, mediante `RDAVIDWebService` se obtiene el **grafo de enriquecimiento** para cada una de las LRs, de la misma manera que se mostró en la sección 3.4.2. De esta manera, los términos enriquecidos (o no) son representados como nodos en el grafo, por cada categoría principal de GO (PB, FM y CC) dando contexto a los resultados, como se muestra en el diagrama de flujo de la figura 3.13. No obstante, estas estructuras deben ser integradas en un único GDA a los efectos de obtener **conocimiento biológico a través de las discrepancias/consensos**.

En la figura 3.14 se aprecia el proceso de **integración de los resultados** de *enriquecimiento funcional*. Para ello se obtiene un GDA que posee la estructura global del experimento, es decir, aquella que contiene todos los nodos y arcos presentes por los diferentes resultados. Luego, por cada una de las LRs se construye el GDA global

Tabla 3.1: tabla de contingencia 2x2 para el i-ésimo término de interés

	Término$_i$	**Término**$_i^c$	**Total**
Candidatos	$n_i - 1$	$N_{Candidatos} - n_i$	$N_{Candidatos}$
Candidatos^c	$n_{Término} - n_i$	$(N - N_{Candidatos}) - (n_{Término} - n_i)$	$N - N_{Candidatos}$
Total	$n_{Término}$	$N - n_{Término}$	N

El total de genes de la lista de referencia (N) se encuentra dividido en filas en caso de pertenecer o no a la lista de candidatos (Candidatos o $Candidatos^c$); las columnas determinan la pertenencia (o no) de los genes al término de interés (Término$_i$ o Término$_i^c$). Note la penalización propuesta por Hosack et al. (2003), para obtener un valor EASE.

e identifican los nodos enriquecidos en color. Posteriormente, los diferentes grafos de enriquecimiento son integrados en una única estructura, mediante el **código de colores** esquematizado por el diagrama de Venn de la figura 3.14. En la estructura de discrepancia/consenso se pueden identificar tres tipos de términos:

Nodo consenso: aquel término identificado como enriquecido por todas las LRs.

Nodo discrepante: término enriquecido en al menos una LR, pero no en todas.

Nodo no enriquecido: nodo interno del GDA que posee un valor $EASE$ mayor al definido por el umbral de enriquecimiento en todas las LRs.

En la figura 3.14, el MRCM resume los **nodos consenso** con el color *rojo*. Los **nodos/ramas discrepantes** son automáticamente resaltados: en *naranja* para aquellos sólo enriquecidos por la LR-I, en *amarillo* para los compartidos por LR-I y II, en *azul* para los únicamente presentes en la LR-II y los exclusivos de la LR-III, en color *verde*. Con este código de colores es posible identificar nuevos términos con relevancia biológica, que se pierden si se utiliza cualquier otra de las referencias habituales (LR-I o LR-II).

A través del MRCM los **nodos consenso** representan nodos **confiables**, en el sentido de que se encuentran *consistentemente enriquecidos, independientemente del análisis llevado a cabo*. A su vez, los **nodos discrepantes** pueden poseer enriquecimiento **espurio**, producto de la longitud de la referencia y no por ello respetar los supuestos de la prueba estadística (LR-I y/o LR-II), como se describió en la sección 1.2.2.

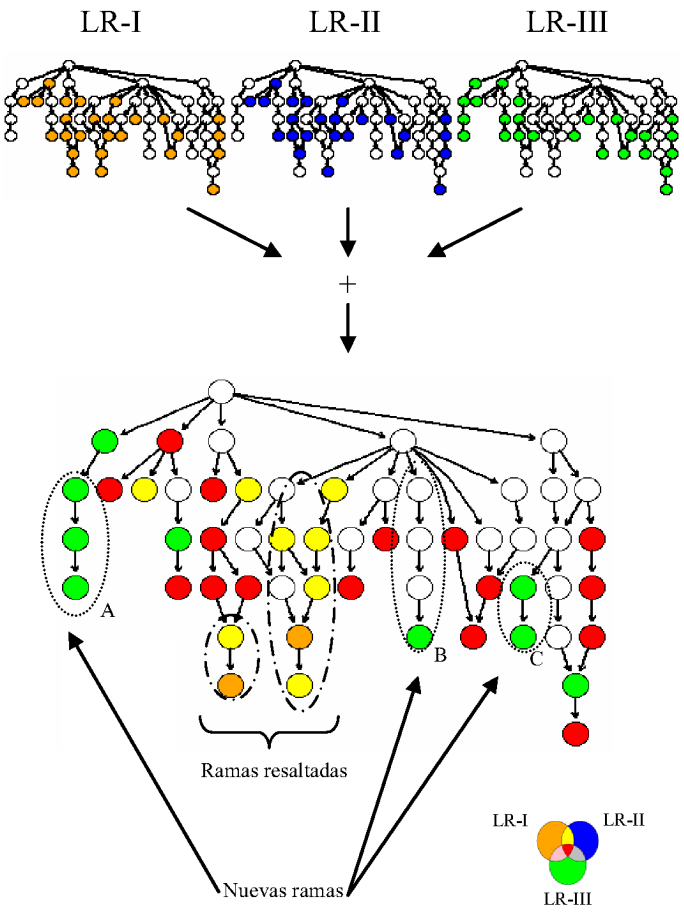


Figura 3.14: Esquema del contraste de múltiples referencias para los datos de Packer et al. (2007) correspondientes a Funciones Moleculares de Gene Ontology. Los nodos enriquecidos se muestran en color, para cada grafo de la correspondiente lista de referencia (LR). La combinación de los resultados se resume en una única estructura, siguiendo el patrón de colores del diagrama de Venn. El método resalta las ramas centrales obtenidas por la LR-I y LR-II, mientras que en A, B y C emergen sólo con la LR-III. Adaptación de la imagen de Fresno et al. (2012).

Los resultados obtenidos de la integración/contraste de las diferentes LRs utilizadas por el MRCM son documentados en un **reporte HTML**, como muestra el diagrama de flujo de la figura 3.13. Dicho reporte no necesita conectividad a internet para su exploración, y de forma interactiva permite navegar los diferentes GDA in-

tegrados de cada categoría de GO. En la sección 3.6 se describe con detalle el aporte realizado en el contexto de la MD a los reportes del *análisis ontológico-funcional*.

3.5.2. Análisis de estabilidad

Un *análisis ontológico-funcional* tradicional usualmente se realiza mediante un SEA, utilizando una única LR. Si bien en la etapa de MD correspondiente a *evaluación* el valor obtenido de la **prueba estadística** es ajustado por comparaciones múltiples, este valor sólo permite establecer sí el término se encuentra enriquecido o no. Es decir, **no brinda información acerca de la sensibilidad o estabilidad del enriquecimiento de cada término**, frente a la LR seleccionada.

Frente a esta problemática, es posible definir una **LR patrón** y realizar simulaciones de tipo **bootstrap** como una alternativa de "**validación por simulación**", como se presentó en la sección 2.5. Justamente, la aplicación de este tipo de metodologías permite aumentar la fiabilidad sobre los resultados en términos de **potencia estadística** en el sentido de detectar enriquecimiento, cuando el efecto "*verdaderamente existe*". Es decir, permite reducir la posibilidad de enriquecimiento "*espurio*", producto de artefactos que puedan sesgar los resultados funcionales obtenidos. Sin pérdida de generalidad, se propone a la LR-III como la referencia patrón para las simulaciones bootstrap. Esta LR cumple con todos los supuestos del estadístico asociado, es decir, todos los candidatos pueden estar en cualquier celda de la tabla de contingencias 1.1 o 3.1.

Una vez definida la LR patrón, la idea subyacente es introducir una **pequeña perturbación** tanto en el tamaño del término, $n_{Término}$, y longitud, N, tratando de mantenerla lo más cercano a la LR utilizada a los efectos de identificar los términos verdaderamente enriquecidos. Para ello, se obtienen diferentes LRs mediante un **muestreo con reposición** (bootstrapping) sobre la LR patrón, manteniendo siempre presente la totalidad de proteínas/genes candidatos. En este sentido, las LRs bootstrap contendrán proteínas/genes repetidos, que son descartados en la construcción de las tablas de contingencias antes de realizar la prueba estadística correspondiente. Consecuentemente, los resultados de la simulación permiten proporcionar una medida de **estabilidad** (potencia) del enriquecimiento para cada término según (3.5):

$$potencia = \frac{cantidad\ de\ veces\ que\ es\ enriquecido}{número\ de\ simulaciones} * 100 \tag{3.5}$$

donde la **potencia** representa el porcentaje de veces que un término se enriquece, sobre un elevado número de simulaciones. En este sentido, mayor potencia implica una mayor estabilidad en el enriquecimiento del término.

Cabe destacar que la idea de validación por simulación en análisis ontológicos-funcionales fue introducida por Zeeberg et al. (2003), a través de valores q en *GoMiner* (sección 1.3). En esta herramienta la perturbación se realiza sobre la lista de proteínas/genes candidatos, situación que responde a *cuán estables son los resultados funcionales con respecto a los candidatos utilizados*. Por el contrario, en esta tesis se asume que la selección de los candidatos es la apropiada y la problemática a abordar radica en *cuán estables son los términos dependiendo de la LR utilizada en el análisis*. Si bien ambas validaciones evalúan la problemática de la estabilidad, en la presente tesis se hace énfasis en las propias características de la LR utilizada y no en la lista de proteínas/genes candidatos. En este contexto, **esta metodología no se encuentra disponible en las herramientas bioinformáticas actualmente disponibles**. Más aún, su implementación es *computacionalmente intensiva* dado que requiere generar un numero elevado de LRs bootstrap para su posterior análisis de SEA. En el caso de utilizar DAVID como motor de cálculo, siguiendo los lineamientos de la sección 3.5.1, se convierte en una tarea impracticable dada la abrumadora intervención en el sitio web. Más aún, esta tarea se encuentra propensa a errores no forzados por parte del usuario. Sin embargo, es posible utilizar `RDAVIDWebService` de forma programática para obtener los resultados (sección 3.4), con la limitación de poder realizar hasta 200 simulaciones por cuenta de usuario por día, o incluso haciendo uso de varias cuentas de usuario en un mismo día.

3.5.3. Bases de datos de ejemplo

El funcionamiento del MRCM se pondrá a prueba utilizando una base de datos de proteómica que utiliza geles de electroforesis bidimensional (2D-DIGE) y tres estudios de microarreglos de ADN del repositorio Gene Expression Omnibus (GEO, `www.ncbi.nlm.nih.gov/geo`). En particular, se centra el interés en las hojas (nodos

sin nodos hijos), dado que contienen información biológica más específica y permiten explicar términos ancestrales. A tales efectos, se estudiarán sólo los nodos discrepantes y serán validados mediante búsqueda de artículos científicos en PubMed (www.ncbi.nlm.nih.gov/pubmed).

Proteómica

Las proteínas expresadas diferencialmente se obtuvieron a partir de un experimento 2D-DIGE para el análisis de secretomas (proteínas secretas hacia el exterior de la célula) de dos líneas celulares de melanoma, donde se varió el nivel de expresión de la proteína protumoral SPARC (Sosa et al. (2007) y Girotti et al. (2011)). La base de datos fue pre-procesada mediante un modelo lineal mixto de dos etapas, como se explica por Fernández et al. (2008), obteniendo 120 manchas (spots) con expresión diferencial.

En este tipo de experimento, las proteínas subyacentes no son conocidas "*a priori*". Más aún, las restricciones biológicas en este diseño experimental sólo permiten ver un subconjunto de las proteínas que realmente está presente en el proteoma bajo estudio, es decir, comprende solamente las proteínas extracelulares. Consecuentemente, no se encuentra disponible la LR-II para el análisis. A su vez, la LR definida por el usuario (LR-III) fue construida utilizando diferentes técnicas dando un total de 3154 proteínas (ver Tabla 3.2 y Fresno et al. (2012)). Esta lista consta de 72 proteínas únicas (46 sobreexpresadas y 26 subexpresadas), obtenidas a partir de la identificación de los spots diferenciales en los geles 2D-DIGE analizados y de 3082 proteínas identificadas en la muestra referencia mediante LC-MS/MS usando Orbitraps (Girotti et al., 2011).

Microarreglos de ADN

Se analizaron tres estudios publicados de Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo), donde se utilizaron microarreglos de Affymetrix® (ver Tabla 3.2). El primer estudio fue llevado a cabo por Packer et al. (2007), donde se obtuvieron los perfiles de expresión génica en 12 réplicas biológicas de tipo salvaje y 23 mutantes para p14ARFs en seres humanos. El objetivo del trabajo fue buscar efectores noveles aguas abajo de la p14ARF.

Tabla 3.2: Descripción de las bases de datos utilizadas para probar el MRCM

Base de datos		Affymetrix			Proteínas/genes diferenciales	
Autor	GEO ID	Nombre del chip	Criterio de calidad #call="P" $\geq$	Total	Criterio	Genes
Packer et al. (2007)	GSE7152	HG-U133 plus 2.0	6 sal. y 12 mut.	11986	FDR<0.05	165 (68↑, 97↓)
Spira et al. (2004)	GSE994	HG-U133A	4 fum. y 4 ctrl.	4128	FDR<0.05 y $\lvert log_2 FC \rvert > 0{,}4$	116 (73↑, 43↓)
McGrath-Morrow et al. (2008)	GSE7310	Mouse Genome 430 2.0	3 fum. y 2 ctrl.	12905	FDR<0.05	118 (10↑, 108↓)
Girotti et al. (2011)	Modelo mixto de 2 etapas Fernández et al. (2008) e identificación por LC-MS/MS			3154	120 spots, 72 proteínas (46↑, 26↓)	

La base de datos de Girotti et al. (2011) no respeta los encabezados de la tabla dado que es un experimento proteómico de 2D-DIGE y por simplicidad, se incluye con el resto de los estudios analizados.

El segundo estudio pertenece a Spira et al. (2004), donde se analizaron los efectos sobre el epitelio bronquial en 20 personas fumadoras y 20 que nunca habían fumado. Los autores llegaron a la conclusión de que fumar induce respuesta xenobiótica, regulación de redox, expresión de varios oncogenes y disminución de la expresión de varios genes supresores tumorales, al igual que de moduladores de la inflamación en vías aéreas.

El último estudio corresponde al realizado por McGrath-Morrow et al. (2008). Los autores analizaron la expresión génica de tejido pulmonar de 6 ratones neonatos expuestos a 14 días de humo de cigarrillo y 4 ratones control. Los autores mostraron que los pulmones perinatales eran particularmente susceptibles a los efectos dañinos de la exposición, inhibiendo la inmunidad innata y perjudicando ligeramente el crecimiento postnatal de los pulmones.

Todas las bases de datos se procesaron bajo el mismo flujo de trabajo utilizando el lenguaje R (R Core Team, 2013) y paquetes de Bioconductor (Gentleman et al., 2005). En primer lugar, la intensidad de las sondas se escaló utilizando el algoritmo MAS5 con los parámetros por defecto del paquete **affy** (Gautier et al., 2004). Esto permitió obtener las medidas de detectabilidad del fabricante (calls) y la señal de expresión. Se incluyeron en el análisis sólo aquellas sondas con anti-sentido único "_a" (Affymetrix, 2004) y confiablemente detectadas (call="P"), para un número mínimo de chips acorde con el diseño experimental de cada estudio. De esta manera, se utilizaron sólo los genes identificados confiablemente en casi todos los microarreglos, para construir la referencia definida por el usuario (LR-III). El paquete **limma** (Smyth, 2004) permitió identificar los genes con expresión diferencial; se aplicó FDR para controlar las comparaciones múltiples, obteniendo los resultados de la Tabla 3.2.

3.5.4. Evaluación

Con el fin de evaluar la robustez del MRCM e independencia de las bases de datos presentadas en la sección 3.5.3, se evaluó la presencia de términos enriquecidos para las principales categorías de GO. La figura 3.15 muestra una visión global (unión) de los términos enriquecidos obtenidos para FM, PB y CC, en todas las bases de datos analizadas. El diagrama de Venn de la extrema izquierda muestra que la mayor parte

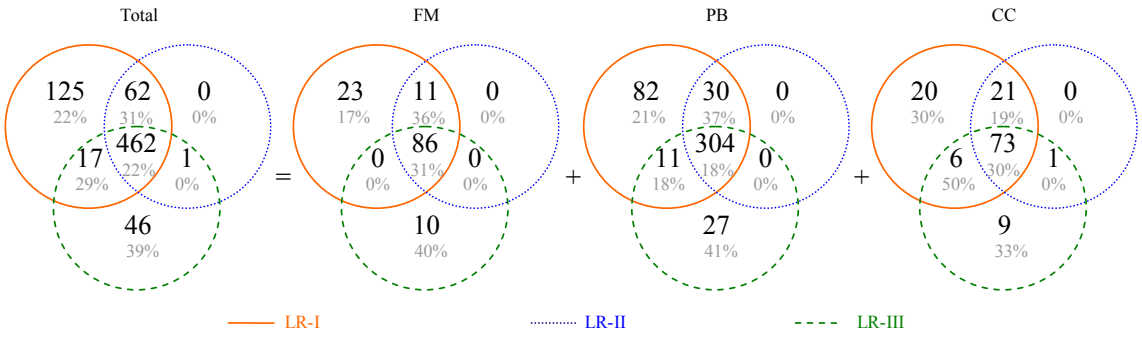


Figura 3.15: Diagrama de Venn para la distribución de términos enriquecidos encontrados en todas las bases de datos presentadas en la sección 3.5.3 para las diferentes categorías principales de Gene Ontology (Funciones Moleculares, Procesos Biológicos y Componentes Celulares). En números, la cantidad de términos enriquecidos para cada subconjunto y en porcentajes, aquéllos que corresponden a nodos hojas. Imagen extraída de Fresno et al. (2012).

de los términos enriquecidos (462) fueron compartidos por las diferentes LR, es decir, los nodos de consenso. A su vez, los resultados de la LR-II (genes del microarreglo) están contenidos en el conjunto de resultados de LR-I (genoma) para cada GDA de GO, a excepción de un término enriquecido en CC (diagrama de Venn de la extrema derecha). Este término se encuentra en la estructura interna del GDA por lo que no aporta nueva información biológica ya que puede ser explicado por nodos hojas enriquecidos (más específicos).

En la figura 3.15 también se aprecia que la LR-I (genoma) presenta el mayor número de términos enriquecidos. No obstante, la mayoría de ellos son nodos no consensuados (125) potencialmente debido a la longitud de la LR. Esto se debe, como se explicó en la sección 1.2.2, que a mayor longitud de LR, mayor oportunidad de que el término salga enriquecido (ver figura 1.3). Al analizar la LR-III (definida por el usuario) y a pesar de que sólo contiene un máximo del 43,4 % de los genes de la LR-I (véase tabla 3.3), 46 nuevos nodos no consensuados están enriquecidos y no son reconocidos en ninguna de las otras dos LRs. En la figura 3.15 se aprecia que el 39 % de ellos aporta nueva información biológica, soportada por la literatura según se detallada en Fresno et al. (2012), independientemente de cuál haya sido la base de datos o la tecnología/experimento.

Tabla 3.3: Población de genes en cada categoría principal de Gene Ontology de acuerdo con las tres listas de referencias utilizadas

Base de datos	Funciones Molecular			Procesos Biológicos		
	I	II	III	I	II	III
Girotti et al. (2011)	15143(100)	-	2561(16.9)	14116(100)	-	2381(16.8)
Packer et al. (2007)	15143(100)	14128(93.3)	6216(41.0)	14116(100)	13187(93.4)	5798(41.1)
Spira et al. (2004)	15143(100)	10886(71.9)	3212(21.2)	14116(100)	10391(73.6)	3089(21.9)
McGrath-Morrow et al. (2008)	15404(100)	12995(84.4)	6549(42.5)	14219(100)	11944(84.0)	6005(42.2)
Base de datos	Componentes Celulares					
	I	II	III			
Girotti et al. (2011)	15908(100)	-	2583(16.2)			
Packer et al. (2007)	15908(100)	14741(92.7)	6384(40.1)			
Spira et al. (2004)	15908(100)	11082(69.7)	3299(20.7)			
McGrath-Morrow et al. (2008)	15855(100)	13596(85.6)	6888(43.4)			

Población de genes para las distintas categorías de Gene Ontology y listas de referencias (I genoma, II chip y III definida por el usuario). Entre paréntesis, el porcentaje de la población respecto a los miembros de I. Cabe destacar que II es casi tan completa como el genoma (I), mientras que sería de esperar una relación más estrecha (próxima) como se muestra para en el conjunto de datos genómica de melanoma. Note también que el criterio de filtrado en III ha eliminado más de la mitad del total de los genes del genoma disponibles en cada categoría de Gene Ontology. Datos extraídos de Fresno et al. (2012).

Por otra parte, se analizó el desempeño del MRCM sobre cada base de datos de la sección 3.5.3. Para el caso de los resultados de Packer et al. (2007), se ha representado en la fig 3.14 un esquema del MRMC correspondiente a la categoría de FM. Un análisis detallado del grafo inferior de la figura mostró 35 términos enriquecidos, distribuidos como se muestra en la tabla 3.4, donde 16 de ellos son nodos de consenso. En este caso, a través de la utilización del MRCM, se identificaron tres nuevas ramas enriquecidas (sólo con la LR-III) directamente relacionados con el entorno experimental. La rama de la extrema izquierda (A) termina en un nodo de *actividad de receptor transmembrana*, la cual posee genes reportados en el estudio original relacionados con vías de transducción de señales célula-célula en receptores de superficie (Barnes, 2009). La nueva rama central (B) posee enriquecido el nodo de *unión de ion calcio*. Este nodo resultó ser un blanco potencial para la terapia de melanomas malignos (Charpentier et al., 2010). La última nueva rama (C), terminó en un nodo de *actividad de transporte de acido carboxílico* que contenía sólo dos genes (SCL16 y CTNS). Esta familia es fundamental para el metabolismo y la regulación del pH, según afirman Halestrap y Meredith (2004), pero no estaría directamente asociada

Tabla 3.4: Términos enriquecidos para las tres categorías de Gene Ontology y las tres listas de referencias utilizados en las cuatro bases de datos.

Base de datos	Nodo	Funciones Moleculares			Procesos Biológicos			Componentes Celulares		
		I	II	III	I	II	III	I	II	III
Girotti et al. (2011)	T	54	-	39	225	-	173	63	-	54
	nD	9	-	0	23	-	1	3	-	0
Packer et al. (2007)	T	26	24	25	127	114	116	38	33	24
	nD	1	2	3	7	4	7	6	4	3
Spira et al. (2004)	T	18	16	14	51	33	37	9	2	9
	nD	1	1	1	10	3	3	4	0	3
McGrath-Morrow	T	22	18	18	24	17	16	10	6	2
et al. (2008)	nD	3	0	0	3	1	0	0	0	0

T: cantidad total de nodos enriquecidos para una referencia dada (I genoma, II genes del microarreglo, III definida por el usuario). **nD**: nodos discrepantes al final de una rama (hojas), es decir, nodos sólo detectados por una o dos referencias. Datos extraídos de Fresno et al. (2012).

al melanoma.

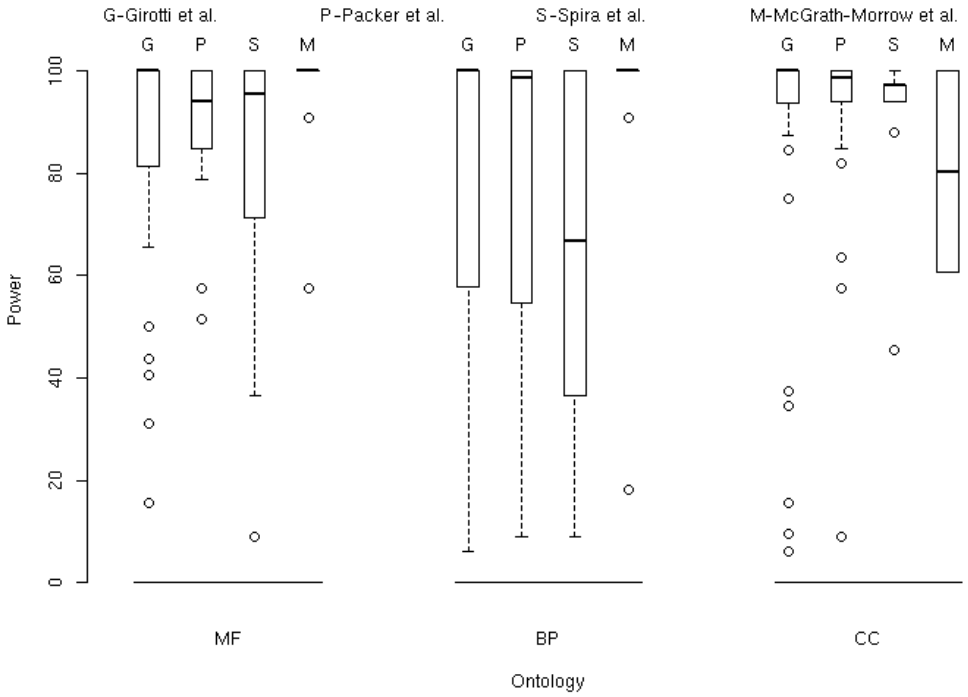


Figura 3.16: Diagrama de cajas de la potencia de los nodos enriquecidos en cada categoría principal de Gene Ontology, para las diferentes bases de datos. Imagen extraída de Fresno et al. (2012).

En el estudio de Girotti et al. (2011), a pesar del hecho de que la LR-III sólo tiene como máximo el 17 % de los miembros de la LR-I (tabla 3.3) y el menor número de candidatos expresados diferencialmente (tabla 3.2), obtuvo el mayor consenso entre todos los resultados (tabla 3.4). Este alto consenso válida los nodos enriquecidos, teniendo en cuenta que una LR larga tiende a producir valores EASE más bajos (significativos) que LRs más cortas. Adicionalmente, el análisis de potencia respecto a la longitud de la referencia, también mostró estabilidad en los resultados.

En la figura 3.16 es posible ver que los diagramas de caja de la potencia de los nodos enriquecidos, está por encima del 50 % para la mayoría de ellos. La potencia tiene una mayor varianza en la categoría principal de PB para todas las bases de datos, dado que es la categoría de GO que tiene la mayor cantidad de términos en

comparación con las otras dos (FM y CC, ver tabla 3.4). Consecuentemente, sólo por tener una mayor cantidad de términos, las perturbaciones presentes de las LRs bootstrap tienen una mayor posibilidad de producir cambios en $n_{Término}$ que en las otras dos categorías (FM y CC).

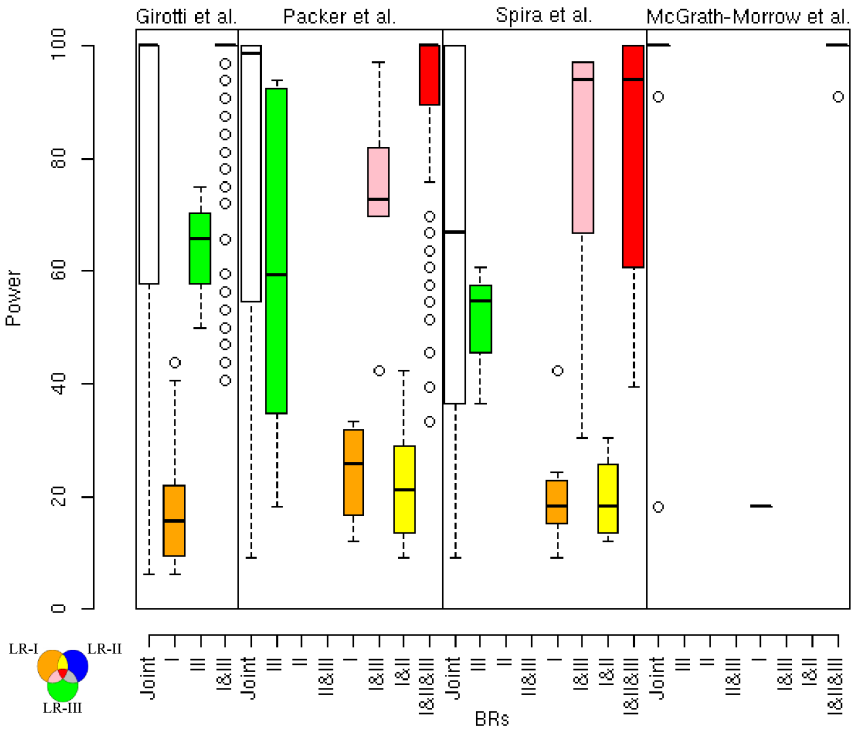


Figura 3.17: Diagrama de cajas de la potencia de los nodos enriquecidos en PB, codificados con la fuente de enriquecimiento del diagrama de Venn. Note que en blanco se encuentra el diagrama de caja conjunto ("Joint") de todos los nodos enriquecidos previamente visto en la figura 3.16. Imagen extraída de Fresno et al. (2012).

En la figura 3.17 se muestran los diagramas de caja de la potencia de nodos enriquecidos para PB, agrupados por la superposición (en color) del contraste de las tres LRs. En este caso, nuevos nodos enriquecidos aparecen para algunas LRs simuladas. Estos también fueron nodos discrepantes sólo en la LR-I. Sin embargo, ellos muestran valores de potencia por debajo del 40 %. Por el contrario, los nodos discrepantes únicamente encontrados por la LR-III alcanzan valores de potencia superiores al 50 % y los nodos de consenso alcanzan potencia más alta para este experimento.

En los experimentos de microarrays, los diagramas de caja para la potencia mostraron el mismo comportamiento (ver figura 3.17). Casi todos los nodos que se encuentran en la LR-III alcanzaron valores de potencia por encima del 50 %. Por otra parte, los nodos que aparecieron enriquecidos por bootstrapping y encontrados previamente por la LR-I o compartidos por LR-I y LR-II mostraron valores de potencia de menos de 40 % en todos los casos. Esto sugiere que los nodos enriquecidos encontrados por la LR-III fueron muy consistentes y potencialmente significativos; además, han sido validados por búsqueda bibliográfica por Fresno et al. (2012). Resultados similares se obtuvieron para la potencia de las otras dos categorías principales de GO (FM y CC). Sin embargo, el comportamiento de los valores de potencia es más evidente, ya que las cantidades de términos GO en estas categorías es menor que en PB.

3.5.5. Comentarios finales

A partir de los resultados obtenidos para los diferentes conjuntos de datos reales de la sección 3.5.3, *se ha demostrado que los mismos varían según la LR utilizada.* Esto podría potencialmente sesgar o contribuir a una *interpretación biológica engañosa de los resultados.* En este contexto, el **MRCM** se ha propuesto para facilitar la identificación de los términos enriquecidos por el contraste de los resultados. En este sentido, se pueden seguir dos abordajes para SEA mediante el MRCM: i) el **uso de más de una LR** o ii) seleccionar una LR definida por el usuario y realizar un **análisis de estabilidad**. En el primer caso, se encontró un elevado consenso, independientemente de la LR para las bases de datos de prueba utilizadas. Esto coincide con la afirmación de Hedegaard et al. (2009), quienes sugieren que *si los resultados biológicos (es decir, las proteínas/genes candidatos) son fiables, los resultados de*

las diferentes LRs deben ser comparables en cierta medida. Sin embargo, términos ontológicos informativos podrían perderse dependiendo de la LR o la visualización utilizada (por ejemplo formato tabular), haciendo dificultoso el proceso de MD de descubrimiento de la patrones biológicamente relevantes.

La inclusión de una referencia definida por el usuario (**LR-III**) permite encontrar *términos enriquecidos y no identificados por los enfoques tradicionales*. Más aún, el **código de color** utilizado en el MRCM ayuda a la identificación de términos biológicamente informativos. A su vez, esto permite una **visión global de los resultados del experimento**, lo que *facilita el análisis y la integración de la información* al destacar nodos y/o ramas del grafo, que en nuestro caso, sugirieron ser relevantes para el contexto experimental. La estrategia propuesta **asiste la inspección del GDA**, evitando mirar en extensas tablas al utilizar la *estructura jerárquica de GO como una estrategia de exploración y resumen visual.* Esto permite a los investigadores centrarse en los nodos hoja, que contienen la información biológica más rica, acelerando el análisis. Mediante el MRCM, los nodos consenso sugieren una visión global del experimento e información sobre la confiabilidad de los genes expresados (aparecen enriquecidos sin importar la LR que se utiliza). Esto permite rápidamente saber que el experimento funcionó en términos generales, dado que se observan los enriquecimientos esperados por la hipótesis experimental planteada. Por otro lado, los nodos discrepantes sugieren nueva información biológica. En este contexto, el uso de una referencia definida por el usuario (LR-III) permite la identificación de nuevos nodos/ramas enriquecidas muy representativos, no antes vistos cuando se utiliza el abordaje de una única referencia.

El uso del MRCM con **remuestreo utilizando la LR-III** permite explorar la estabilidad del enriquecimiento. Mediante el análisis de potencia se demostró que los nodos discrepantes, identificados únicamente por la LR-I y/o LR-II, son inestables, lo que sugiere enriquecimiento espurio. Por el contrario, nodos enriquecidos encontrados por la LR-III mostraron alta potencia, lo que sugiere mayor "**confianza**", haciendo a estos nodos buenos candidatos de exploración. En los conjuntos de datos aquí utilizados, los nodos enriquecidos encontrados por el MRCM fueron **validados por la literatura**, como se describe en el material suplementario de Fresno et al. (2012).

A diferencia de otras herramientas, el MRCM incluye toda la información "*a*

priori" (sin recortar el GDA de GO) y "*a posteriori*" (sin filtrado de los resultados obtenidos), con el fin de dejar que los GDA y el MRCM hablen por sí mismos. Los resultados sugieren que se obtiene más información utilizando DAVID y GO sin ninguna restricción. Por ejemplo, la nueva rama enriquecida que contiene al nodo "*actividad de transporte transmembrana de ácido carboxílico*" de la figura 3.14, resultó ser fundamental para el metabolismo y la regulación de pH. Este nodo no se identificaría utilizando a DAVID con la estrategia definida por defecto, es decir, excluyendo los términos con menos de 3 proteínas/genes. Esto es especialmente importante en los estudios proteómicos, donde términos que poseen pocas proteínas podrían quedar fuera del análisis, mientras que existe evidencia de la presencia de ellos en varias manchas expresadas diferencialmente en el gel, como en el estudio de Girotti et al. (2011).

3.6. Visualización y exploración de los resultados

La última etapa del flujo de trabajo del KDD corresponde a los "*reportes*", como se presentó en el capítulo 2 y se encuentra esquematizada en la figura 3.2. En esta etapa, el investigador cuenta con los resultados de *anotación, expresión, estadísticos* y *funcionales* de su experimento. En este punto se encuentra con el cuello de botella más grande de todos: "**la exploración de los diferentes reportes**".

En la sección 1.3.5 del capítulo 1 se mostró que el tipo de reporte *depende de la herramienta bioinformática utilizada*. Usualmente, la mayoría de las herramientas exportan los resultados en extensas **tablas de anotación, expresión y resultados funcionales**. En otros casos son reportes en formato de **páginas web, imágenes/gráficos predefinidos** con escasa o nula capacidad de interacción con el usuario. No obstante, cuando la interacción es posible, se encuentra circunscrita a un sitio web que *requiere conectividad a internet* como en el caso de *DAVID*, *GOstat*, *GoMiner*, etc. En estos casos, generalmente es necesario volver a analizar los datos cada vez que se desean explorar, *dificultando el procesos de búsqueda y análisis de patrones desde la MD*.

En este contexto, el investigador es el único responsable de *integrar las salidas obtenidas de la aplicación de diferentes herramientas* para obtener una visión com-

pleta del modelo biológico bajo estudio. Sin embargo, esto no es posible incluso para diseños experimentales simples (caso control-tratamiento) y mucho más dificultoso al analizar experimentos de mayor complejidad. De manera que la propia **complejidad de la integración de información**, es en sí misma un problema. A su vez, la falta de técnicas de **resumen visual** sobre los resultados **limita la capacidad de análisis**. *Esto impacta negativamente en la extracción de patrones que pueda realizarse aplicando técnicas de MD, sobre la información que pudiese estar disponible.*

El aporte de esta tesis en materia de reportes, en el contexto del *análisis ontológico-funcional*, consta de un reporte **HTML** denominado "**contraste ontológico**" (Fresno et al., 2011). El reporte permite un **análisis más estructurado y completo** de la información biológica existente de forma *interactiva*, sin necesidad de conectividad a internet. En él es posible **integrar de forma visual** los **resultados funcionales y de expresión**, utilizando la misma idea de la metodología del MRCM presentado en la sección 3.5, mediante grafos de enriquecimiento de GO como los de las figuras 3.13 y 3.14. De esta manera es posible integrar automáticamente el enriquecimiento de las *diferentes LRs, análisis de estabilidad*, incluso de resultados de *diferentes experimentos o diseños de mayor complejidad*, para visualizar mediante un patrón de colores las nuevas relaciones inferidas de la integración/contraste. Así, una vez obtenida la estructura, se generan dos vistas:

Vista de enriquecimiento. En esta vista se pueden navegar los grafos de *integración funcional* de las diferentes categorías de GO (PB, FM o CC). En cada uno de ellos es posible visualizar el *nombre* de los nodos enriquecidos (término) y en *color*, la/s fuente/s de procedencia. Así, la propia estructura de GO sirve tanto de estrategia de *resumen* de información, al igual que *guía para la exploración* de los *nodos terminales u hojas*, que son los que poseen la mayor especificidad biológica de la rama del grafo bajo exploración.

Vista de expresión. Adicionalmente a la información de GO, esta vista *integra la expresión de las proteínas/genes asociados al experimento*. Para ello, cada nodo se representa con un gráfico de *sección circular* la cantidad de proteínas/genes *sobre y/o subrexpresadas*, y un punto central con la *fuente de enriquecimiento*. A través de esta vista, automáticamente se muestra la existencia (o no) de **patrones de sobre o subexpresión** asociada a términos biológicos específicos;

es decir, si existen **ramas del grafo** que presenten sobre o subexpresión. Cabe destacar que esta vista no se encuentra presente en ninguna de las herramientas presentadas en la sección 1.3.

En ambos casos, el usuario puede seleccionar de forma interactiva un **nodo de interés** y acceder a su **información de anotación** asociada (*IDs, símbolos, vínculos a bases de datos externas, etc.*), **información de expresión** (*media de expresión en diferentes tratamientos, etc.*) e **información funcional** (*definición del término, valor EASE, etc.*). De esta manera, el usuario tiene la posibilidad de integrar diferentes fuentes de información e incluso *personalizar la información de expresión*, a los efectos de explorar en su conjunto el modelo biológico bajo estudio. Finalmente, se genera un **reporte HTML** que integra las diferentes vistas obtenidas (**enriquecimiento** y **expresión**) para cada una de las tres categorías de GO (FM, PB y CC) e incluso permite acceder a la información original obtenida por DAVID/`RDAVIDWebService`.

A los efectos de mostrar un caso de uso del reporte del contraste ontológico, se presentan los resultados obtenidos con el MRCM de la sección 3.5, para el experimento de *geles de proteínas 2D-DIGE* de Girotti et al. (2011) presentado en la sección 3.5.3 y tablas 3.2, 3.3 y 3.4.

3.6.1. Evaluación del contraste ontológico

En la figura 3.18 se muestran capturas pantallas de un reporte típico del contraste ontológico. El reporte posee, en el margen superior, una **barra de navegación** que permite acceder a diferentes fuentes de información:

Genes: posee la información de **anotación** y **expresión** de las proteínas/genes diferenciales del experimento en formato tabular.

KEGG: los resultados obtenidos en DAVID/`RDAVIDWebService` para las vías metabólicas de KEGG, a los cuales se les ha **incorporado la información de anotación y expresión** de las proteínas/genes sobre o subexpresadas.

BP, MF, CC: las tres categorías principales de GO. Se puede acceder a esta información a través de las tablas originales obtenidas por DAVID/`RDAVIDWebService`, o mediante las **vistas de enriquecimiento** (`Nombre`) y/o **de expresión** (`Pie`)

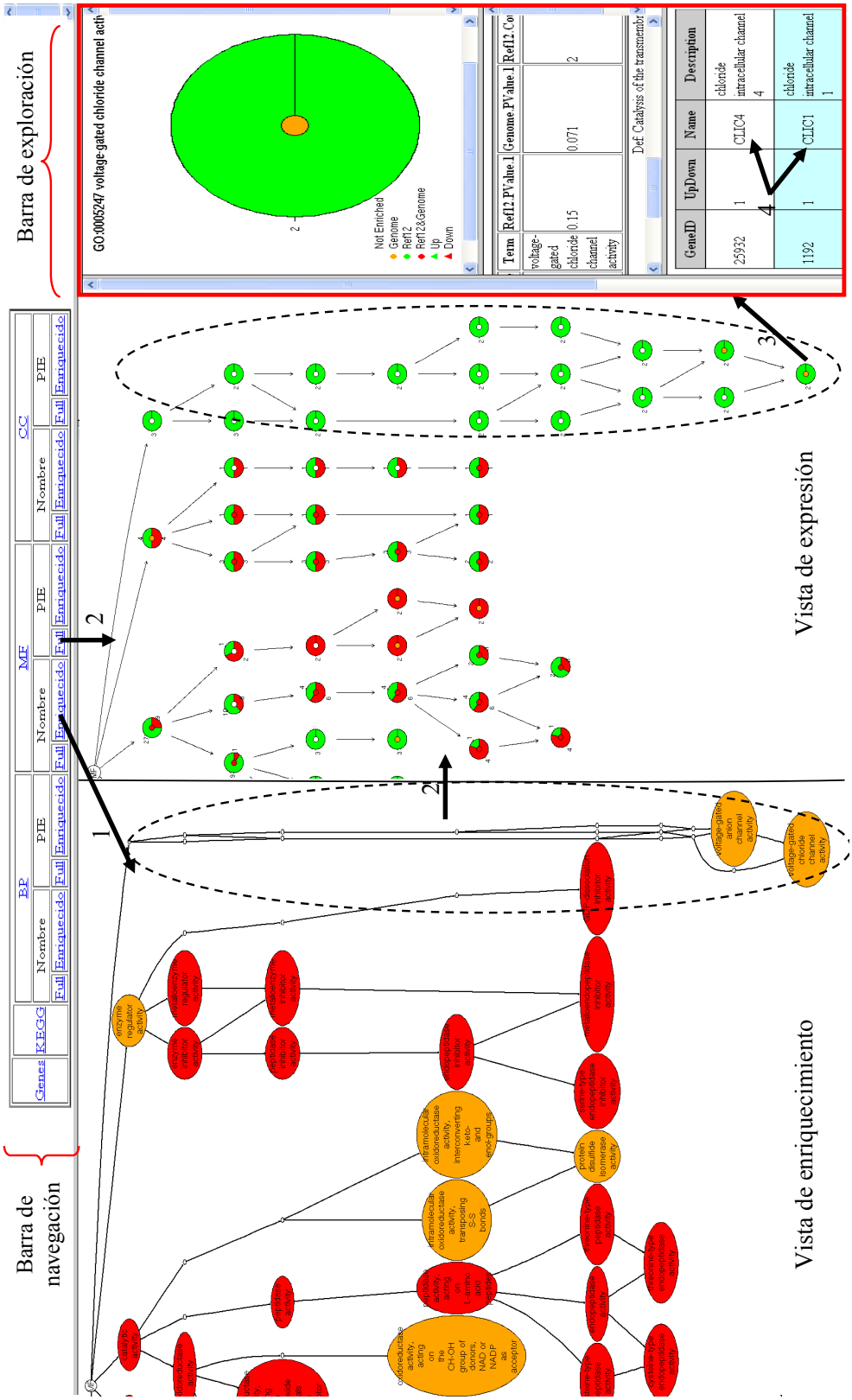


Figura 3.18: Capturas de pantalla del reporte HTML para el experimento de proteómica funcional de Girotti et al. (2011). Imagen extraída de Fresno et al. (2011).

en su versión resumida, que presenta sólo los nodos enriquecidos (`Enriquecido`) o el grafo completo de GO (`Full`).

En un **primer análisis exploratorio**, el usuario usualmente puede elegir seguir el esquema de los pasos numerados del 1 al 4 en la figura 3.18. Inicialmente en 1 seleccionó la **vista de enriquecimiento de MF**. En la sección del grafo mostrada se aprecian solamente aquellos nodos que resultaron enriquecidos, con su correspondiente nombre y grado de superposición (en color). En ella se observa un elevado consenso entre los dos contrastes utilizados (en rojo), es decir, nodos enriquecidos en ambas referencias (LR-I yLR-III) y esperables desde la biología del experimento (siempre enriquecidos). Al utilizar este tipo de representación, el usuario puede obtener rápidamente un **visión global del experimento** y *filtrar visualmente aquellos nodos más genéricos*, para focalizar su atención en los nodos enriquecidos más específicos (más profundos). De esta manera, se observa una rama de gran especificidad que termina sólo en dos nodos naranja, enriquecidos en el genoma o LR-I (elipse en trazo discontinuo, en la figura 3.18).

Motivado por esta observación, el usuario puede trasladarse a la **vista de expresión** *completa*, como es en el caso de continuar con el paso número 2. En esta vista se observa que *la totalidad de la rama se encuentra sobrexpresada* (sección verde completa), manteniendo en el punto central la fuente de enriquecimiento: *blanco* para los no enriquecidos y *naranja* para los observados en la vista de enriquecimiento.

En este punto, si el investigador quisiera saber cuál es la información funcional, expresión y anotación del nodo terminal de la rama sobreexpresada denominado "*actividad regulada por voltaje de canales de cloruro*", debe seleccionar el nodo correspondiente y navegar en la **barra de exploración**, como se presenta en el paso 3 de la figura 3.18. Esta barra se encuentra dividida en tres secciones:

Superior: presenta el mismo *diagrama de sección* del nodo correspondiente a la vista de expresión, donde adicionalmente se incluye la *leyenda* con las fuentes de enriquecimiento (punto central) y los niveles de expresión (secciones).

Media: presenta una *tabla con la información funcional* (nombre del término, valores p/EASE de enriquecimiento de las diferentes LRs, cantidad de proteínas/genes, definición del término, etc).

Inferior: presenta una *tabla con la descripción de anotación y expresión* de las proteínas/genes con expresión diferencial (IDs, expresión, símbolo/vínculos a PubMed, descripción, etc.) que resultaron asociados a dicho término biológico.

En particular, se observan dos genes de la familia CLIC para este nodo. El investigador puede continuar la búsqueda siguiendo con el paso 4 de la figura 3.18 y seguir el enlace de PubMed provisto por el reporte para acceder a toda la información disponible (requiere acceso a internet). De hecho, Fresno et al. (2012) realizaron una **validación bibliográfica** y relacionaron estos genes con *migración celular de melanomas*, Madeja et al. (2001a,b). Adicionalmente, utilizando una estrategia similar sobre el resto del grafo, resultó de interés el nodo *unión de proteínas no plegadas*. Este nodo se encontró relacionado con SPARC durante el desarrollo embrionario en estudios de deposición de colágeno tipo IV en la lámina basal y también fue mencionado como un chaperón molecular en el retículo endoplasmático, Martinek et al. (2007) y Pfaff et al. (1993).

Al utilizar la herramienta en la categoría de PB, el contraste ontológico automáticamente resaltó 10 nodos fuera del consenso (naranja), altamente específicos (hojas) y a su vez relacionados con SPARC. Por ejemplo, "*organización de filamentos intermedios del citoesqueleto*" es intrínsecamente afectado por SPARC (Alvarez, 2006). Del mismo modo, "*regulación positiva de migración y quemotaxis de leucocitos*", son directamente afectados por la expresión de SPARC según lo reportado por Alvarez et al. (2005) y Kelly et al. (2007). Términos de "*respuesta a stress celular, daño de axones y hormona esteroide*" también fueron resaltados y asociados a este gen (Au et al., 2007; Dieudonné et al., 2000; Luna et al., 2009; Sawhney, 2002; Schellings et al., 2004; Vadlamuri et al., 2003). Proteínas de la matriz celular como SPARC, también se encuentran involucradas en los términos resaltados de "*desarrollo de sistema nervioso y diferenciación celular*" (Chavey et al., 2006; Eroglu, 2009; Vincent et al., 2008). Por otra parte, es interesante observar que mediante el reporte propuesto, se encontraron términos resaltados en FM y PB donde genes de la familia CLIC también están presentes. Una relación entre SPARC y familia CLIC en "*regulación negativa de ubiquitinación de proteínas*" ha sido recientemente sugerida por Nakayama (2010) y Bellei et al. (2010). Por otra parte, "*transporte de aminas*" fue el único nodo que emergió en PB en el contraste ontológico, relacionado con la referen-

cia definida por el usuario LR-III (en color *verde*). Sin embargo, la evidencia en la literatura no resultó concluyente acerca de la relación de SPARC con este proceso, a pesar de que esta proteína posee capacidad para transglutaminarse debido a la alta densidad de ácidos glutámicos del extremo N-terminal (Hohenadl et al., 1995), pero las transglutaminasas de tejidos no están involucradas en este tipo de procesos.

En la última categoría principal de GO, CC, el reporte resaltó tres hojas en el genoma (*naranja*), "*membrana basal, complejo de enzimas de ubiquitinación y envoltura nuclear*" asociadas a expresión de SPARC, de acuerdo con las observaciones de Anwar et al. (2011); Sacks-Wilner y Freddo (1990).

3.6.2. Comentarios finales

El "**contraste ontológico**" ha mostrado ser un reporte capaz de abordar la problemática de sus predecesores en lo que respecta al **integración simultánea de diferentes fuentes de información y una visualización intuitiva para la inspección de los resultados obtenidos a partir de SEA**. A diferencia de otras herramientas de la sección 1.3.5, hace uso de la **estructura de GO** para presentar de forma *amigable* y facilitar la **exploración simultánea de los resultados**. En este contexto la propia estructura de grafo, permite que el usuario pueda con un simple filtrado visual acceder rápidamente a la información experimental relevante; en cambio, la metodología tradicional (por ejemplo, formato tabular), hace dificultoso el proceso de descubrimiento y exploración de información biológica. Adicionalmente, su utilización en conjunto con el MRCM de la sección 3.5, facilitó la identificación de términos biológicamente informativos, dando una rápida visión general de los resultados del experimento. El método resaltó automáticamente nodos y ramas del grafo que tuvieron relevancia biológica para nuestro contexto experimental. Más aún, la **vista de expresión** permitió extender el análisis a *ramas completamente sobre o subexpresadas*, aportando información propia de la biología al investigador. Así, la metodología propuesta facilita la inspección del grafo, evitando la búsqueda en detalle y ahorrando tiempo de análisis.

El reporte propuesto para **recuperación de información** (consulta simultánea a bases de datos) y **visualización**, como una herramienta de MD, permite **fácilmente ver información contextual enfatizando los nodos potencialmente**

relevantes mediante la codificación de colores, es decir, *identificando información novel en el panorama que presenta el grafo*. Por ejemplo, la rama enriquecida de "*actividad regulada por voltaje de canales de cloruro*" de la figura 3.18, resultó tener un elevado grado de importancia para el experimento de Girotti et al. (2011). Esta rama no hubiera sido identificada utilizando DAVID/`RDAVIDWebService` con sus parámetros por defecto (excluye términos con menos de 3 genes). Esto es especialmente importante en estudios de proteómica donde, en el ejemplo mostrado, un término que contiene solo dos genes (CLIC4 y CLIC1) resultó enriquecido, a pesar de que estas proteínas/genes se encontraron presentes diferencialmente en varias isoformas del gel de melanoma.

Capítulo 4

Aplicaciones

En este capítulo se muestra cómo a partir de la aplicación de las metodologías desarrolladas en esta tesis, descritas en el capítulo 3, ha sido posible encontrar información biológica en diferentes contextos experimentales, mostrando la utilidad práctica de las mismas.

En el experimento de Loreti et al. (2013) se evalúa el impacto funcional de la hormona folículo estimulante (FSH) en humanos. La actividad biológica de la FSH se encuentra dada por las diferentes configuraciones que puede adoptar dicha proteína, lo cual a su vez depende de la etapa del ciclo folicular. Además, la abundancia de las distintas configuraciones se ve alterada entre la pre y postmenopausia. La importancia de la FSH en clínica médica radica en que sólo se utiliza la configuración de mayor actividad biológica para tratamientos de inseminación artificial. En este contexto, los autores evaluaron el impacto funcional de las diferentes configuraciones de FSH. En particular se extiende la idea del **MRCM** presentado en la sección 3.5, para comparar diferentes contrastes de tratamientos (preguntas biológicas) obtenidos con `RDAVIDWebService` (sección 3.4) desde un punto de vista funcional. Luego, se utilizan los reportes obtenidos del **contraste ontológico** (sección 3.6), a los efectos de asistir a la exploración del impacto funcional de las diferentes variantes de FSH.

En el experimento de Denninghoff et al. (2014) se investiga el efecto protector/reparador del aceite de pescado sobre una lesión renal aguda inducida por dieta en un modelo de ratón. En particular, los autores están interesados en encontrar vías metabólicas relacionadas con efecto terapéutico del aceite de pescado sobre el riñón

y evaluar posibles efectos colaterales en hígado. En este contexto, se profundiza sobre la exploración multivariada utilizando `lmdme` como se presentó en la sección 3.3, a los efectos de inspeccionar los diferentes *biplots* en búsqueda de existencia de patrones de asociación entre genes y tratamientos. Adicionalmente, se integran/contrastan funcionalmente los diferentes tratamientos y órganos a través del **MRCM** y reportes del **contraste ontológico** (sección 3.5 y 3.6), siguiendo una estrategia similar a la utilizada sobre los datos de Loreti et al. (2013).

4.1. Impacto funcional de variantes de FSH

El crecimiento de los folículos ováricos es un proceso complejo regulado por gonadotropinas, esteroides y factores de crecimiento (Richards et al., 2002). La hormona folículo estimulante (FSH), juega un papel esencial durante la foliculogénesis ovárica y sus acciones tienen consecuencias importantes en la fertilidad, ya que los ratones hembra con deficiencia de subunidades β de FSH o receptor de FSH son infértiles (Abel et al., 2000; Kumar et al., 1997). Esta gonadotropina no sólo regula la proliferación de células de la granulosa y la producción de estradiol, sino que también previene la apoptosis de las células de la granulosa y la atresia folicular (Chun et al., 1996; Robker y Richards, 1998).

Al igual que otras hormonas glucoproteicas, la FSH se compone de una familia de variantes de glicosilación que difieren entre sí en la estructura del oligosacárido incluyendo la finalización de la síntesis de la rama, grado de ramificación y el contenido de ácido siálico. En modelos *in vitro* se ha demostrado que las **variantes de glicosilación** de la FSH tienen acciones complementarias y específicas sobre los folículos en desarrollo, y que se requiere un balance específico de glicoformas para un óptimo desarrollo del folículo (Barrios-de Tomasi et al., 2006; Ulloa-Aguirre et al., 1999; Vitt et al., 1998). Por otro lado, el grado de bioactividad de la FSH es inversamente proporcional al **contenido de ácido siálico** (Zambrano et al., 1996). En este contexto, el objetivo propuesto por Loreti (2012) y Loreti et al. (2013) fue determinar el posible *impacto funcional de la complejidad de oligosacáridos y contenido de ácido siálico en FSH recombinante humana (FSHrh), sobre células de granulosa humana en cultivo.*

4.1.1. Entendimiento de datos

El modelo biológico bajo estudio fue una línea celular tumoral, similar a una granulosa humana (KGN), la cual mantiene la expresión funcional del receptor de FSH y la capacidad de producir esteroides y expresar las subunidades α y βA inhibinas (Nishi et al., 2001). Las inhibinas son complejos protéicos que regulan a la baja la síntesis de FSH e inhiben la secreción de FSH. Esta línea celular es estimulada con un medio de cultivo que posee diferentes aislamientos obtenidos de la FSHrh comercial (NICHD, NIH; USA):

Análogos de carga: mediante isoelectroenfoque, se combinaron las fracciones recuperadas de diferentes preparados con pH 2,56 a 4,00 para obtener una mezcla más ácida (**FSHrh-AC**) y análoga en carga de ácido siálico. A su vez, las fracciones con un pH >5,00 se combinaron para obtener una mezcla más básica (**FSH-BA**), como se describe para ambas preparaciones en Loreti et al. (2013).

Isoformas con distinta complejidad de oligosacáridos: mediante cromatografía en Concanavalina A se separaron tres grupos de variantes glicosiladas de FSHrh de acuerdo a la complejidad de sus oligosacáridos retenidos por lectina:

No retenidos (NR): FSHrh con glicoformas que poseen complejos, triantenarios y bisectrices de oligosacáridos.

Débilmente retenidos (DR): FSHrh con glicoformas que poseen cadenas de carbohidratos biantenarios.

Fuertemente retenidas (FR): FSHrh con glicoformas que poseen un elevado contenido de oligosacáridos de manosa o de tipo híbrido.

De los tres grupos anteriores sólo se utilizaron las preparaciones que no retienen (FSHrh-NR) y aquellas fuertemente retenidas (FSHrh-FR), como se describe en Loreti et al. (2013).

Las células de KGN se cultivaron en un medio con la FSHrh comercial nativa, dos aislamientos de ácido siálico (FSHrh-AC y FSHrh-BA) y dos aislamientos de complejos de oligosacáridos (FSHrh-NR y FSHrh-FR), empleando una dosis de 20 ng/ml de cada preparado y dos réplicas biológicas por cada tratamiento durante

24 horas. Luego se extrajo, purificó e hibridizó el ARN utilizando el microarreglo "Human Gene 1.0 ST" de Affymetrix® siguiendo el protocolo del fabricante. Se utilizó el software del fabricante, Expression Console 1.1, para obtener los niveles de expresión de genes y las medidas de calidad/detectabilidad (call) del fabricante. Una vez obtenidos los valores de expresión del experimento, se abordaron las diferentes etapas del KDD descriptas en el capítulo 2, utilizando el flujo de trabajo de la sección 3.1:

- La *conversión e integridad de anotación* utilizó el aporte de la sección 3.2, empleando la anotación del fabricante en conjunto con el paquete de Bioconductor del fabricante y una actualización de los datos asociados a los IDs con e-utiles.
- El *filtrado* de datos consideró sólo aquellas sondas que poseen anotación, que codifican alguna proteína y que han sido detectadas en la totalidad de los microarreglos, según las métricas del fabricante obtenidas por Expression Console®, dado que se poseen sólo dos réplicas biológicas por tratamiento.
- La *normalización* de los datos no ha sido necesaria, dado que se utilizó el algoritmo RMA-SKETCH para obtener la señal de intensidad de las sondas. Este algoritmo aplica una transformación a los valores de intensidad de manera de dejar a todos los microarreglos con la misma distribución (Affymetrix, 2004).
- La *reducción, proyección e integración* de datos, fue llevada a cabo mediante el ajuste del modelo lineal de la ecuación (4.1) para cada gen del microarreglo presente en esta etapa:

$$y_{ij} = \beta_{0i} + \beta_{1i}\tau_{AC}(i) + \beta_{2i}\tau_{BA}(i) + \beta_{3i}\tau_{NR}(i) + \beta_{4i}\tau_{FR}(i) + \varepsilon_{ij} \qquad (4.1)$$
$$i = 1, \ldots, N; j = 1, 2$$
$$\varepsilon_{ij} \sim N(0, \sigma_i^2) \quad \wedge \quad Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0 \quad \forall\ i \neq k \wedge j \neq l \qquad (4.2)$$

donde:

y_{ij} es el valor log_2 de expresión del i-ésimo gen, para la j-ésima réplica
β_{0i} representa el nivel de expresión para el control de FSHrh comercial
$\beta_{0i}, \beta_{1i}, \ldots, \beta_{4i}$ y σ_i^2 son parámetros desconocidos del modelo a estimar,

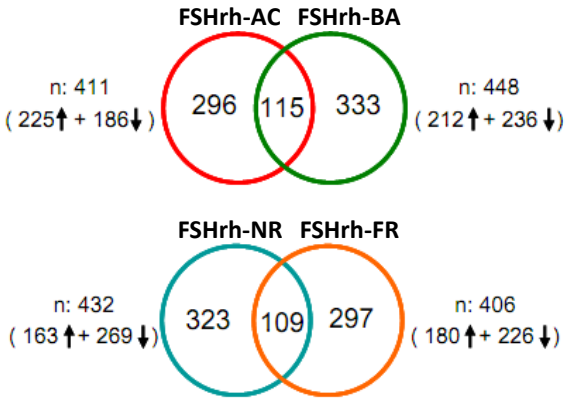
ε_{ij} es el error aleatorio no observable, sujeto a los supuestos de (4.2)

$\tau_{AC}(i), \tau_{BA}(i), \tau_{NR}(i)$ y $\tau_{FR}(i)$ son variables binarias para indicar la pertenencia o no (1 o 0), del i-ésimo gen al tratamiento FSHrh-AC, FSHrh-BA, FSHrh-NR y FSHrh-FR respectivamente.

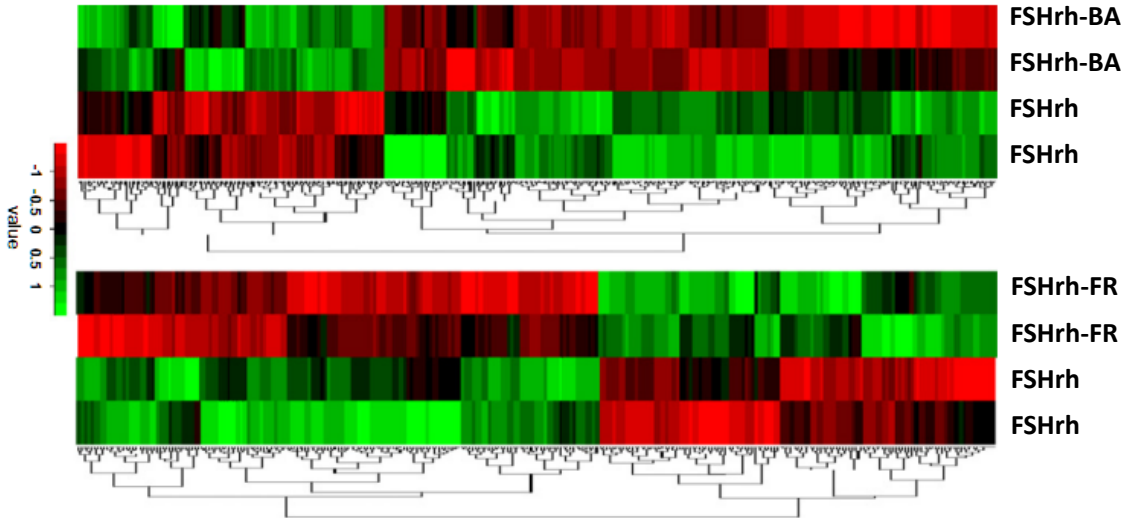
El modelo (4.1) se ajustó con la librería **limma** de R, utilizando una corrección empírica de Bayes (Smyth et al., 2011). Se seleccionaron aquellos genes expresados diferencialmente entre cada uno de los tratamientos respecto de la situación de control, es decir, $H0 : \beta_{ki} = 0$ contra $H1 : \beta_{ki} \neq 0$ para $k = 1, \ldots, 4 \ \ \forall i$. Adicionalmente, se compararon los tratamientos con diferente ácido siálico y complejidad de oligosacáridos. En todos los casos se utilizó un valor $p < 0{,}05$ y $|\beta_{ki}|$ de corte, que permitiera una correcta separación de los mapas de calor de los diferentes pares de contraste de tratamientos y suficiente información ($\approx$ 400 genes) para la etapa de modelado como se describe en Fresno et al. (2012).

En la figura 4.1 se muestran los resultados obtenidos en la etapa de reducción, proyección e integración de expresión para los diferentes tratamientos. En el panel (a) se observan dos diagramas de Venn, con los genes diferenciales obtenidos para las comparaciones de interés biológico respecto del control (FSHrh). En este sentido, el diagrama superior compara la diferencia de genes candidatos según el contenido de ácido siálico (FSHrh-AC vs FSHrh-BA), mientras que en el inferior los genes atribuidos a la complejidad de oligosacáridos (FSHrh-NR vs FSHrh-FR). En todos los tratamientos se obtuvo una cantidad de genes en el orden de 400 individuos y una distribución similar de genes sobre o subexpresados. Note que para ambos diagramas, los tratamientos comparten en el orden de 1/4 de los genes seleccionados. Una lista completa de los genes, descripción y valores de expresión se puede encontrar en el material suplementario de Loreti et al. (2013) y en la página web `www.bdmg.com.ar/?page_id=251`.

En la figura 4.1(b) se muestran dos mapas de calor, uno para verificar el contenido de ácido siálico y el segundo para la complejidad de oligosacáridos. En estas figuras, las filas representan los tratamientos y en columnas los genes de los microarreglos, para los tratamientos FSHrh-BA y FSHrh-FR respecto del control (FSHrh). En ambos mapas de color se aprecia el correcto agrupamiento de las réplicas biológicas de cada tratamiento. Además se puede ver el cambio en el nivel de expresión de un



(a) Genes diferenciales



(b) Mapas de calor

Figura 4.1: Reducción, proyección e integración de genes diferenciales. (a) Diagrama de Venn con los genes candidatos para las diferentes comparaciones de interés biológico, respecto de la condición de control (FSHrh). Las flechas indican el sentido de sobre o subexpresión de los genes. (b) Mapas de calor para la comprobación visual el agrupamiento de las réplicas biológicas para el criterio de selección utilizado. Adaptación de imágenes de Loreti et al. (2013).

mismo gen atribuido al cambio de tratamiento, es decir, pasa de sobre a subexpresión o viceversa. Un comportamiento similar presentan los otros dos tratamientos respecto del control, los cuales no son mostrados en la figura 4.1(b).

4.1.2. Modelado

La etapa de *modelado* se llevó a cabo mediante el MRCM presentado en la sección 3.5. En este sentido, se utilizó la idea de obtener conocimiento biológico por el consenso/discrepancia del enriquecimiento de diferentes listas de referencias (LRs), pero en este caso sobre resultados funcionales obtenidos de listas de genes provenientes de diferentes contrastes de tratamientos.

El MRCM se utiliza de forma habitual para realizar el análisis ontológico-funcional de cada una de las listas de genes empleando las tres LRs propuestas: el genoma de la especie (LR-I), los genes impresos en el microarreglo (LR-II) y aquellos detectados de manera confiable (LR-III). Posteriormente, se obtiene el grafo de enriquecimiento para cada LR a los efectos de integrar los resultados en un grafo "ampliado", operando por columnas sobre los grafos que poseen un mismo color (verde, azul o naranja), como se muestra en la figura 4.2 para los contrastes que involucran diferentes conte-

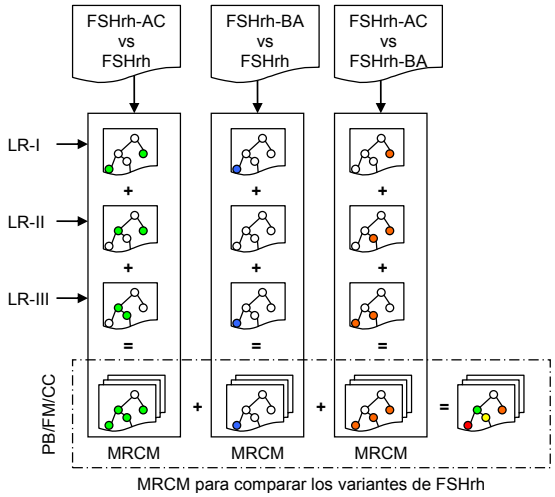


Figura 4.2: Contrastes de enriquecimiento funcional de las variantes de FSHrh relacionadas a diferente contenido de ácido siálico: ácidas (FSHrh-AC), control (FSHrh) y básicas (FSHrh-BA). Imagen adaptada de Fresno et al. (2012).

nidos de ácido siálico (FSHrh-AC, FSHrh-BA y FSHrh). De esta manera, la totalidad de información de enriquecimiento de un mismo contraste es integrada en un único grafo. Un esquema similar de procesamiento se utilizó para los tratamientos que poseen diferente complejidad de oligosacáridos (FSHrh-NR, FSHrh-FR y FSHrh).

Una vez obtenido el grafo que integra los resultados funcionales de las tres LRs para cada lista de genes, se busca comparar a nivel funcional aquellos contrastes de interés biológico. Para ello se emplea la misma idea del MRCM, donde ahora el consenso/discrepancia se aplica sobre la última fila de la figura 4.2, es decir, sobre la suma de los resultados parciales obtenidos para las columnas (grafos ampliados de contrastes de variantes de FSHrh). Ahora el patrón de colores permitirá identificar visualmente la especificidad funcional del diferente contenido de ácido siálico o complejidad de oligosacárido de la/s variante/s de FSHrh utilizadas. Justamente, la posibilidad de integrar información funcional de diferentes listas de genes candidatos, no se encuentra disponible en ninguna de las herramientas bioinformáticas presentadas en la sección 1.3. Esta es una característica novel del MRCM, donde, a través de los reportes del Contraste Ontológico de la sección 3.6, es posible explorar de una manera rápida, visual y eficaz los resultados de las diferentes variantes glicosiladas de la FHSrh, sin que ello se torne en una tarea tediosa para el investigador, como se describió en el capítulo 1.

4.1.3. Evaluación

El uso conjunto del MRCM y los reportes del Contraste Ontológico, mostraron que en el análisis de enriquecimiento funcional tanto *contenido de ácido siálico* como *complejidad de oligosacáridos*, modulan la expresión de genes implicados en la actividad funcional de las células KGN como se describe en Loreti et al. (2011), Loreti (2012) y Loreti et al. (2013).

En la figura 4.3 se muestran los resultados del contraste funcional para diferente **contenido de ácido siálico** en el grafo de PB de GO. En la figura se aprecia que con el MRCM fue posible descubrir ramas funcionales asociadas a la FSHrh-AC (en color *azul*), nodos hoja que se diferencian entre las variantes ácidas y básicas (en color *naranja*), ramas y nodos hoja asociadas a la FSHrh-BA (en color *verde*) y unos pocos nodos consensuados por todas las variantes (en color *rojo*). Esta visualización

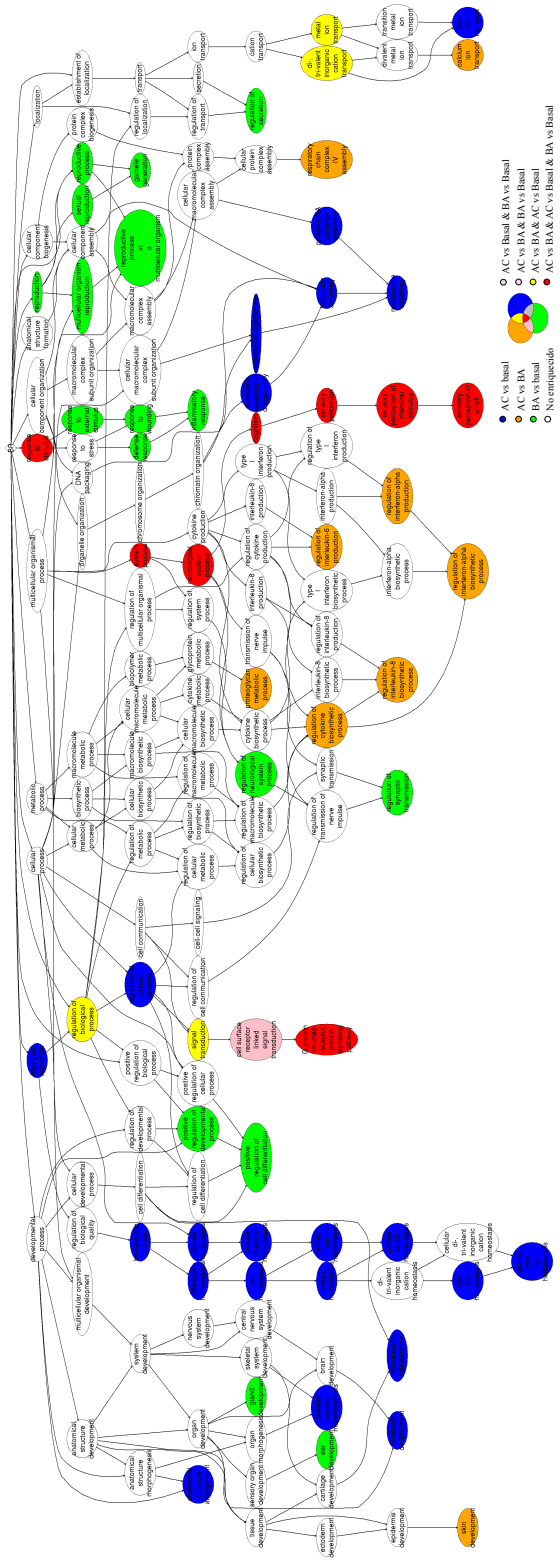


Figura 4.3: Adaptación del MRCM para contrastar el enriquecimiento funcional de Procesos Biológicos de Gene Ontology, de las variantes de FSHrh con diferente contenido de ácido siálico: ácido (FSHrh-AC), control (FSHrh) y básico (FSHrh-BA). Imagen extraída de Loreti et al. (2011).

es de gran utilidad para la búsqueda de patrones, desde una perspectiva de la MD, donde la estructura de GO y los colores permiten una visión funcional global de las variantes de FSHrh asociadas al contenido de ácido siálico.

En particular se encontraron diferentes nodos enriquecidos por la FSHrh-AC como "*homeostasis celular*", "*organización del nucleosoma*", "*ensamblado de esteroides*" y "*transporte de iones de hierro*", los cuales están relacionados con el modelo biológico bajo estudio. A su vez, la contraparte con menor contenido de análogos de ácido siálico (FSHrh-BA), se asoció a términos de GO enriquecidos principalmente en aspectos importantes del "*proceso de reproducción*" tales como "*generación de gametos*", "*regulación de la diferenciación celular*" (factores de crecimiento) y "*regulación de la secreción celular*". Estos términos soportan las observaciones de Zambrano et al. (1996) donde el grado de bioactividad de la FSH es inversamente proporcional al contenido de ácido siálico.

Un grafo similar se obtuvo para las FM de GO y el diferente contenido de ácido siálico. En este sentido la FSHrh-AC se asoció a términos enriquecidos en "*actividad esteroide deshidrogenasa*", "*actividad transmembrana de glucosa*", "*actividad de canales de calcio*", "*unión de esteroides y ácido nucleico*" y "*actividad de receptores nucleares dependiente del ligando*", los cuales son rápidamente identificados de forma visual en el grafo. Bajo la misma metodología de exploración, los términos enriquecidos por la FSHrh-BA fueron: "*ligando de factores de crecimiento*", "*unión de iones de calcio*", "*actividad regulada por la síntesis de óxido nítrico*", "*actividad del inhibidor de endopeptidasa del tipo de serina*" y "*actividad de la proteína tirosina quinasa*". En este grafo los nodos consensos y aquellos asociados entre las variantes ácidas y básicas, son nodos internos a la estructura, por lo que son explicados por algunos de aquéllos biológicamente más específicos de FSHrh-AC o FSHrh-BA nombrados con anterioridad.

La exploración de la **complejidad de oligosacáridos** a nivel funcional fue la que brindó mayor cantidad de información sobre el modelo biológico bajo estudio. En este sentido, los grafos contenidos en el reporte del Contraste Ontológico poseen una cantidad elevada de nodos enriquecidos. Esto se atribuye directamente al grado de especificidad funcional de las variantes involucradas en la complejidad de oligosacáridos. A los efectos de la evaluación, en la figura 4.4 se muestran aquellos nodos

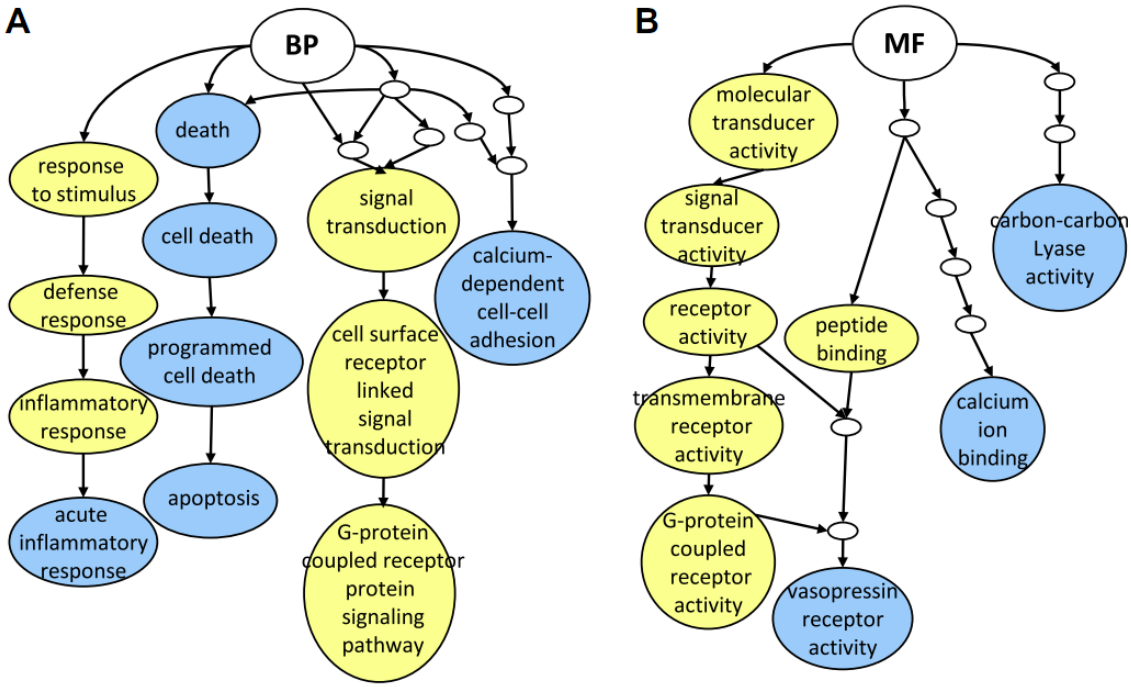
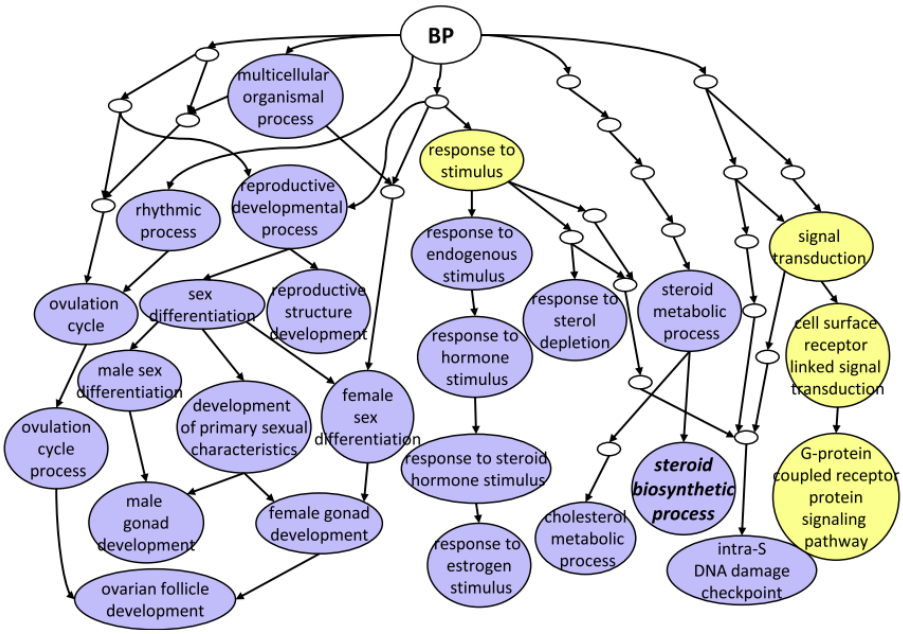


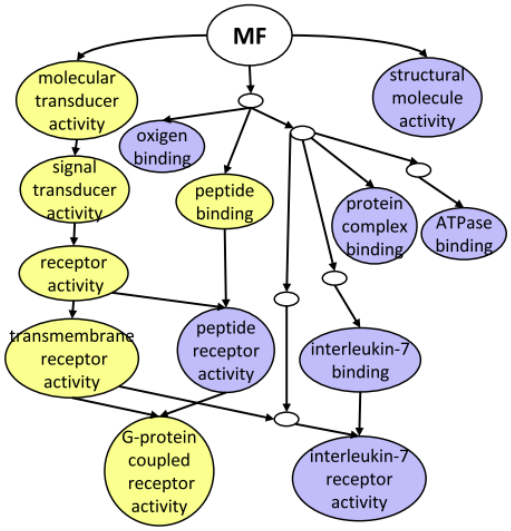
Figura 4.4: Extracto del grafo de enriquecimiento funcional asociado a la complejidad de oligosacáridos. En el panel A se presenta para Procesos Biológicos y en el B para Funciones Moleculares M. En color *amarillo* los nodos enriquecidos para todas las variantes de FSHrh y en *celeste*, aquellos específicamente relacionados a oligosacáridos no retenidos por lectina (FSHrh-NR). Imagen extraída de Loreti et al. (2013).

relacionados específicamente con la FSHrh-NR para PB y FM en el panel A y B respectivamente. En estos grafos el MRCM resalta los nodos enriquecidos comunes a las diferentes variantes de FSHrh en color *amarillo*, mientras que en *celeste* resalta los específicos de FSHrh-NR. En el panel A se resaltan los nodos de "*apoptosis*" así como la "*respuesta inflamatoria aguda*" y la "*adhesión célula-célula dependiente de calcio*" vinculados a la FSHrh-NR. A su vez, en el panel B la misma variante enriquece los términos de "*actividad de liasa de carbono-carbono*", "*unión de iones de calcio*" y "*actividad de receptores de vasopresina*", como se describe en Loreti et al. (2013).

Por otra parte, en el panel A de la figura 4.5 se observa cómo la variante de FSHrh-FR afecta genes que enriquecen nodos de PB como la "*biosíntesis de esteroides*",



(a) Procesos biológicos



(b) Funciones moleculares

Figura 4.5: Extracto del grafo de enriquecimiento funcional asociado a la complejidad de oligosacáridos. En el panel A se presenta para Procesos Biológicos y en el B para Funciones moleculares M. En color *amarillo*, los nodos enriquecidos para todas las variantes de FSHrh y en *violeta*, aquellos específicamente relacionados a oligosacáridos fuertemente retenidos por lectina (FSHrh-FR). Imagen extraída de Loreti et al. (2013).

la "*respuesta a estímulos de estrógeno*", "*punto de control de daño intra-S en el ADN*", "*desarrollo del folículo ovárico*" y "*metabolismo del colesterol*". Los nodos enriquecidos por esta variante en FM, se encuentran asociados a "*unión de ATPasas*", "*unión a receptores de interleukina-7*" y "*unión de oxígeno*" como se muestra en el panel B de la figura 4.5.

Curiosamente, en las figuras 4.4 y 4.5 las ramas enriquecidas comunes a todas las condiciones experimentales estudiadas terminan en el nodo hoja de la "*vía de señalización de receptores de proteínas-G acoplados*" en el grafo de PB o en la "*actividad de receptores de proteínas-G acoplados*" para el grafo de FM; es decir, que tanto las variante del contenido de ácido siálico como la de complejidad de oligosacáridos enriquecen dichos nodos. De hecho, esta observación fue detectada en el grafo que incluye la totalidad de variantes de complejidad de oligosacáridos, que por razones de tamaño del grafo no ha sido incluido y que a Loreti et al. (2013) les sugiere un par de hipótesis para explorar en trabajos futuros.

Los resultados obtenidos para los grafos de CC asociados a las diferentes variantes glicosiladas de FSHrh no fueron concluyentes. En este sentido, no fue posible detectar patrones desde la MD sobre las "vistas de enriquecimiento" o "de expresión" del reporte de Contraste Ontológico que fueran de utilidad para generar nuevo conocimiento sobre el comportamiento de las KGN.

A partir de los resultados anteriores y del conocimiento biológico previo del modelo biológico, Loreti et al. realizaron un selección sobre los nodos hojas enriquecidos sobre PB para elegir genes candidatos para la posterior "**validación biológica**". En este contexto, se seleccionó el nodo de "*biosíntesis de esteroides*" relacionado con actividad específica de la FSHrh-FR como se muestra en el panel A de la figura 4.5. Dentro de este nodo se realizó una selección de potenciales candidatos, mediante la exploración del reporte de Contraste Ontológico mostrado en la sección 3.6.1. En este sentido, se tuvo en cuenta la integración de información de expresión de las diferentes variantes glicosiladas de FSHrh y la evidencia de artículos científicos relacionados los modelos de las células KGN, siguiendo los vínculos de PubMed incluidos en el reporte. De esta manera, resultaron como candidatos los genes con los símbolos STAR, HSD3B2, CYP19A1 y HSD17B para validación por RT-PCR (sección 2.5). Para todos los candidatos seleccionados, se obtuvieron resultados de expresión simi-

lares a los reportados para la expresión realizada por microarreglos, como se describe en detalle para cada uno de ellos en Loreti et al. (2013).

4.1.4. Comentarios finales

La integración de la información experimental, expresión y ontológica-funcional utilizando los aportes de esta tesis en lo que refiere al MRCM y Contraste Ontológico presentados en la sección 3.5 y 3.6, ha demostrado ser de mucha utilidad para el estudio del funcionamiento de la línea celular tumoral, similar a una granulosa humana. En este contexto, el experimento de Loreti et al. (2013) es el primero en el cual ha sido posible estudiar el impacto funcional en PB y FM tanto del *contenido de ácido siálico*, como de la *complejidad de oligosacáridos* que poseen la familia de variantes de glicosilación de la FSHrh.

A través de los grafos de enriquecimiento, el MRCM permitió identificar fácilmente las condiciones experimentales asociadas a determinados términos biológicos relevantes al funcionamiento de las células de granulosa. Justamente, la posibilidad de integrar y explorar resultados provenientes de diferentes análisis de tipo SEA no se encuentra disponible en ninguna de las herramientas presentadas en la sección 1.3. Esta integración, en conjunto con las vistas tanto de *enriquecimiento* como de *expresión* de los reportes del Contraste Ontológico presentados en la sección 3.6.1, permiten una exploración amigable e integral de los resultados del diseño experimental, expresión y ontológico-funcional. A partir de la exploración fue posible identificar nodos candidatos y una posterior selección de genes candidatos. Estos genes fueron seleccionados por su nivel de expresión diferencial entre las variantes glicosiladas de FSHrh y existente evidencia en la literatura para su posterior validación biológica, como se describe en detalle en Loreti (2012) y Loreti et al. (2013).

Los resultados obtenidos a partir de los diferentes aportes metodológicos introducidos en esta tesis, apoyan aún más el concepto de que el contenido de glicano específico en la estructura molecular de la FSH influencia selectivamente la expresión de los genes necesarios para una adecuada función y crecimiento de los folículos ováricos humanos. Se están realizando nuevos estudios para determinar el impacto que pueda tener el aspecto novel de la acción de la hormona en el desarrollo folicular y la calidad de los ovocitos.

4.2. Efecto protector del aceite de pescado en la insuficiencia renal aguda

En las últimas décadas las investigaciones relacionadas con enfermedades críticas se han centrado cada vez más en el pronóstico y los resultados a largo plazo. Pocos estudios han descrito el desenlace a largo plazo de la **insuficiencia renal aguda** (IRA), a pesar de ser un trastorno común entre los pacientes hospitalizados. Ella representa entre $3-7\,\%$ de los pacientes que ingresan al hospital y un $25-30\,\%$ de los pacientes en la unidad de cuidados intensivos (Bagshaw, 2006; Brenner, 2004). Si bien tanto el tratamiento como la gestión técnica de la IRA han cambiado drásticamente en las últimas décadas, la tasa de mortalidad parece haber permanecido sin cambios, en alrededor de un $50\,\%$ (Bellomo, 2006; Ympa et al., 2005). Algunos pacientes nunca van a recuperar por completo la función renal, derivando en una insuficiencia renal crónica, que requiere diálisis de por vida o incluso un trasplante de riñón (Webb y Dobb, 2007).

Los mecanismos referidos a la etiología, al igual que a la progresión de enfermedades renales, no se comprenden en su totalidad. En este contexto, el posible rol patogénico de los cambios en los lípidos renales ha sido estudiado repetidamente, sin clara evidencia de una correlación entre un cambio lipídico en particular y la histología renal asociada. No obstante, se sabe que la cantidad y calidad de lípidos de la dieta pueden modular las lesiones renales en ratas alimentadas con una dieta deficiente en colina (Fewster y Hall, 1967; Monserrat et al., 1974; Simon et al., 1968).

El aceite de coco es rico en ácidos grasos saturados y tiene un efecto protector que se asocia con su contenido de ácido mirístico. El aceite de pescado también es rico en ácido mirístico y además posee ácido eicosapentaenoico y docosahexaenoico. Estos ácidos pueden influir en la composición de ácidos grasos renal y en el metabolismo del ácido araquidónico, el cual desempeña un papel clave en la fisiopatología renal (Courrèges et al., 2002; Monserrat et al., 2000, 1995; O´Neal et al., 1961). En este contexto, el objetivo propuesto por Denninghoff et al. (2014) *fue investigar el posible efecto protector del aceite de pescado en riñones, basado en un modelo nutricional de IRA*. Adicionalmente, se incluyen los resultados no publicados de Denninghoff et al., donde al análisis se agrega el *posible efecto colateral del aceite de pescado en hígado*.

4.2.1. Entendimiento de datos

El modelo biológico bajo estudio fueron 24 ratas Wistar machos de 21 días recién destetadas, las cuales se dividieron en cuatro grupos. Cada grupo se alimentó con la misma dieta específica por seis días antes que los animales se sacrificaran, como se describe en Denninghoff et al. (2014). Las dietas comprenden una estructura de tratamientos de dos factores con dos niveles cada uno:

Colina: un preparado deficiente en colina (CD) y otro suplementado en colina (CS).

Aceite: un preparado con aceite vegetal (AV) y otro con aceite de pescado (AP/AM).

Cabe destacar que en este modelo alimenticio, la dieta normal comprende la combinación de CS y AV (CSAV), mientras que la ausencia de colina en la dieta (CDAV) desarrolla IRA con alteraciones morfológicas, que comprenden desde necrosis tubular focal hasta necrosis cortical masiva y, en la mayoría de los casos, muerte por IRA (Monserrat et al., 1981). En cada grupo de ratas se extrajo suero para cuantificar diferentes biomarcadores y muestras de tejido del riñón izquierdo, para validación biológica del modelo experimental y comprobación de la histopatología de la IRA, como se describe en Denninghoff et al. (2014). El riñón derecho e hígado se utilizaron para analizar el nivel de expresión génica. Para ello, para cada combinación de *colina* × *aceite* se extrajo el ARN de cada tejido y se crearon dos preparados (*pool* en inglés) con las muestras de tres ratones cada uno, donde los preparados no se encuentran apareados entre los dos tejidos. Luego se extrajo, purificó e hibridizó el ARN utilizando el microarreglo "Rat Gene 1.0 ST" de Affymetrix® siguiendo el protocolo del fabricante. Para cada combinación de tratamientos se hibridizaron tres chips para hígado y dos para riñón, debido a restricciones de calidad en la obtención en la muestra por presencia de necrosis renal. Los datos de microarreglos de riñón pueden ser accedidos en Gene Expression Omnibus (GEO) utilizando el código de acceso GSE34139, mientras que los datos de hígado aún no se encuentran publicados en el repositorio.

Los valores de intensidad de expresión de las sondas se obtuvieron utilizando el software Expression Console® 1.1. En el caso del chip Rat Gene 1.0 ST, no es posible obtener las medidas de calidad/detectabilidad (call) del fabricante como

usualmente se obtienen con el algoritmo MAS5, dado que este chip no posee las sondas apropiadas. Sin embargo, este software permite procesar las sondas a nivel de "*exones*", a los efectos de tener una estimación del ruido que hay en la señal. De esta manera, es posible calcular lo que se conoce como puntaje DABG (*Detection Above Background* en inglés), a partir del cual se pueden obtener las medidas de calidad/detectabilidad. Así, una vez adquiridos los valores de expresión y calidad del experimento, se abordan las diferentes etapas del KDD descriptas en el capítulo 2, utilizando el flujo de trabajo de la sección 3.1 con pequeñas modificaciones debidas a la obtención de datos a nivel de exones:

- La *conversión e integridad de anotación* utilizó el aporte de la sección 3.2, empleando la anotación del fabricante del microarreglo a nivel de exones. No obstante, posterior a los filtrados por control de calidad y anotación, la señal es resumida a nivel de genes, siendo necesario incorporar la información del fabricante a nivel de genes y una actualización de los datos asociados a los IDs con `e-utiles`.

- El *filtrado* de datos consideró sólo aquellas sondas a nivel de exones que:

  1. Poseen anotación en la base de datos EntreZ Gene ID (sección 2.3.2).
  2. Se encuentran en la base de conocimiento de DAVID (sección 1.3).
  3. Codifican alguna proteína, es decir, no pertenecen a ningún control.
  4. No presentan hibridación cruzada, es decir, son únicas (sección 2.3.2).
  5. Se encuentran confiablemente presentes en la totalidad de los microarreglos del riñón (2 de 2 chips) y al menos en 2 de los 3 chips de hígado, según las métricas del fabricante obtenidas a partir del puntaje DABG. En el caso de los microarreglos de hígado que poseen un dato ausente o marginal, este se considera como un valor faltante dado que la intensidad obtenida por el escáner no es confiable.

- La *integración de información* contempló consolidar/resumir la señal de expresión de las 99.283 sondas obtenida a nivel de exones que codifican para un mismo gen. Para ello se utilizó la anotación del fabricante a nivel de genes,

para identificar el conjunto de exones presentes en cada gen y promediar los valores de expresión, resultando en 17.256 genes.

- La *normalización* de los datos no ha sido necesaria, dado que se utilizó el algoritmo RMA-SKETCH para obtener la señal de intensidad de las sondas. Este algoritmo aplica una transformación a los valores de intensidad de manera de dejar a todos los microarreglos con la misma distribución (Affymetrix, 2004).

Control de calidad y exploración multivariada

Continuando con el *entendimiento de datos*, es posible realizar un **control de calidad multivariado** de los datos. Para el presente diseño experimental, se esperaría que las réplicas biológicas provenientes de los mismos tratamientos (chips) se comporten de una manera similar. Así, es posible corroborar si la variabilidad total de los genes logra diferenciar la combinación de *tejido* × *colina* × *aceite*. Para ello se propone realizar un análisis de componentes principales (PCA, Peña (2002)), considerando como individuos a los chips (tratamientos) y atributos a los genes (va-

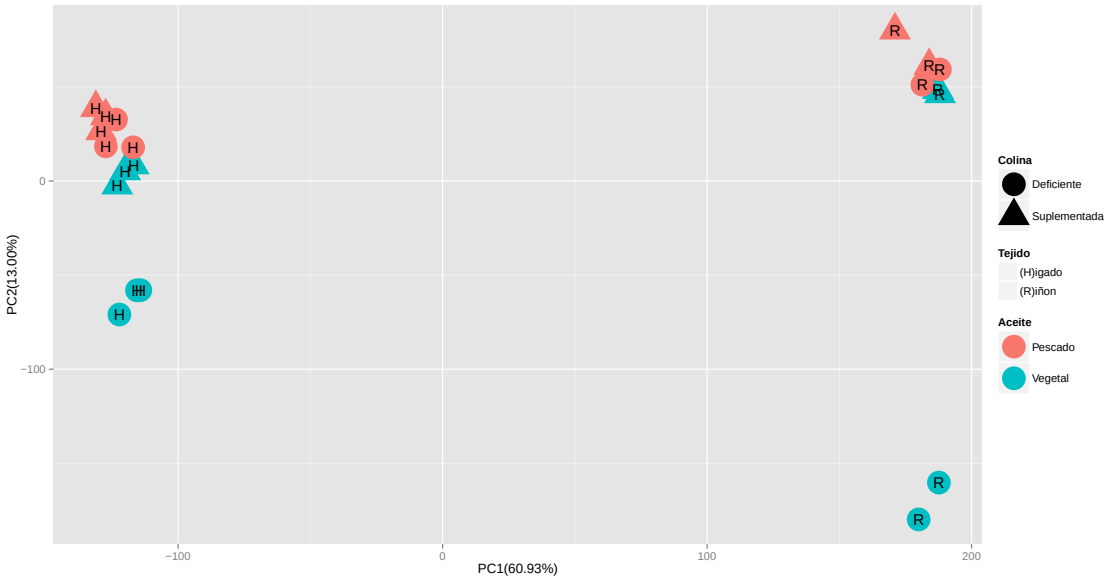


Figura 4.6: Análisis de componentes principales de la transpuesta de la matriz de expresión. La gráfica muestra los datos transformados en las dos primeras componentes principales obtenidos para los microarreglos (combinación de tratamientos).

riables), es decir, utilizar la transpuesta de la matriz de expresión y explorar el plano de las componentes principales. Cabe destacar que este tipo de control de calidad no es usual en aplicaciones de microarreglos; corrientemente se realiza una comprobación de los supuestos de normalidad, dejando fuera la información del diseño experimental.

En la figura 4.6 se muestran los chips en el plano de las dos primeras componentes principales que explican, en conjunto, el 73,93 % de la variabilidad total. La primera componente, PC1, induce dos agrupamientos en donde se encuentran chips del mismo tejido. Es decir, en la derecha de la figura están los chips de riñón (R) y a la izquierda los correspondiente a hígado (H). Adicionalmente, la segunda componente (PC2) separa para cada tejido la dieta de colina deficiente (forma de círculo) y aceite vegetal (color cyan) o CDAV, es decir, la dieta que produce necrosis en riñón y su equivalente en hígado, de las dietas que no producen necrosis en riñón. Más aún, esta separación (variabilidad) es mayor para el caso del órgano blanco (riñón), en comparación al órgano de control (hígado), probablemente por la repercusión de la IRA en hígado, la que usualmente evoluciona a hígado graso (Brenner, 2004).

Los resultados del control de calidad de la figura 4.6 no muestran la existencia de artefactos aparentes en los diferentes microarreglos. Más aún, los resultados coinciden con lo esperable del diseño experimental. En este sentido, un control multivariado de este tipo podría dejar en evidencia algún defecto técnico en algún microarreglo que no es visible en la comprobación de supuestos de normalización de los niveles de expresión. En caso de encontrar un chip en un agrupamiento incorrecto es posible, de forma temprana en el análisis, indagar sobre la génesis de los datos y, si fuera necesario, excluirlo del análisis.

Una vez concluido el control de calidad es posible realizar una **exploración multivariada de la matriz de expresión**, para ver cómo se comportan los genes en su conjunto, frente a los diferentes efectos contenidos en el diseño experimental bajo estudio. Para ello se utiliza el aporte de "**lmdme**" presentado en la sección 3.3, donde el modelo ANOVA de la ecuación (3.1) equivalente para cada gen en este

experimento, se corresponde con el presentado en la ecuación (4.3):

$$y_{ijklm} = \mu_i + Tejido_j + Colina_k + Aceite_l + Tejido_j \times Colina_k + Tejido_j \times Aceite_l + Colina_k \times Aceite_l + Tejido_j \times Colina_k \times Aceite_l + \varepsilon_{ijklm} \quad (4.3)$$

con $i = 1, \ldots, N$, $\{j, k, l\} = 1, 2$ y $m = 1, \ldots, m(j)$ dependiendo del tejido; donde y_{ijklm} es el valor log_2 de expresión del i-ésimo gen para el j-ésimo tejido ($Tejido_j$), bajo el k-ésimo efecto de colina $Colina_k$, con el l-ésimo efecto de aceite ($Aceite_l$), para la m-ésima replica biológica; μ_i es el nivel de expresión medio del i-ésimo gen; las diferentes combinaciones dobles de factores ($Tejido_j \times Colina_k$, $Tejido_j \times Aceite_l$ y $Colina_k \times Aceite_l$) y la única combinación triple de factores ($Tejido_j \times Colina_k \times Aceite_l$); por último, $\varepsilon_{ijklm} \sim N(0, \sigma^2)$ es el término de error aleatorio.

A los diferentes términos de la descomposición ANOVA (4.3), con excepción de la media μ, se les realizó una prueba F para evaluar si alguno de los coeficientes para cada uno de los niveles es diferente de cero (ver sección 3.3.1). La tabla 4.1 muestra los resultados obtenidos para un valor $p < 0{,}05$. En ella se puede apreciar cómo la inclusión del efecto "$Tejido(T)$" como primer factor, remueve dicho efecto dado el elevado número de genes detectados por la prueba (9.643). No obstante, de los dos efectos principales restantes, el "$Aceite(A)$" posee 3.890 genes frente a la "$Colina(C)$" con 1.826, es decir, el aceite posee un efecto biológico mayor que la colina, dado que se obtienen más del doble de genes. Adicionalmente, en este modelo se incluyen las interacciones dobles y la única triple con tejido. En la tabla 4.1 se muestra que existe mayor interacción para "$T \times A$" seguido de "$T \times C \times A$" y "$T \times C$" con 1.061, 769 y 295 genes respectivamente. Por último, la única interacción doble donde no participa el tejido, "$C \times A$" muestra un efecto intermedio entre la colina y el aceite, con 2.836 genes.

Tabla 4.1: Número de genes para una prueba F sobre la descomposición (4.3)

Tejido (T)	Colina (C)	Aceite (A)	$T \times C$	$T \times A$	$C \times A$	$T \times C \times A$
9.643	1.826	3.890	295	1.061	2.836	769

Los genes de la tabla poseen al menos un nivel del factor distinto de cero, para un $p < 0{,}05$ de la prueba F correspondiente, como se describe en la sección 3.3.1.

Utilizando los genes identificados en la tabla 4.1, es posible realizar un análisis de tipo ASCA, como se describió en la sección 3.3.1. Es decir, realizar un PCA sobre los coeficientes de cada término de los genes identificados. En esta oportunidad se han excluido del análisis los efectos principales (Tejido, Colina y Aceite) dado que sólo poseen dos coeficientes. Ello se debe a que por restricción del modelo lineal (Graybill, 2000), siempre los coeficientes poseen signos opuestos, dejando así ambas flechas del biplot en extremos opuestos, razón por la cual la PC1 siempre los separa.

En la figura 4.7 se incluyen los biplots del análisis ASCA de las interacciones dobles y la única triple del modelo (3.3.1). En las interacciones dobles se aprecia que la primera componente explica más del 91 % de la variabilidad en cada biplot. A su vez, la segunda componente explica para estos casos menos del 5 % de la variabilidad restante. En los paneles superiores de la figura 4.7 se presentan las interacciones dobles que involucran al tejido (T:C y T:A). En ellas se observa en la primera componente cómo la variabilidad del hígado (TH) es menor a la del riñón (TR), ya que aquellas interacciones que involucran TH:CD/CS o TH:AV/AM se encuentran más cerca del origen (flechas cortas) en las gráficas T:C y T:A respectivamente. Por otra parte, si bien el eje de la PC1 presenta dos agrupamientos a la izquierda y derecha del origen de coordenadas, estos no permiten separar aquellas interacciones con el mismo tejido o mismo nivel de colina/aceite en ambas gráficas. Sin embargo, la PC2 en el panel superior derecho de la figura 4.7 (T:A) agrupa los diferentes niveles de aceite, separando hacia arriba del origen el aceite de pescado o menhaden (AM) y hacia abajo, el aceite vegetal (AV).

En el panel izquierdo inferior de la figura 4.7 se muestra la interacción de *colina* × *aceite* (C:A). Curiosamente la PC1 agrupa, a la derecha del origen, el efecto protector del aceite de pescado frente a la falta de colina (CD:AM) junto con la dieta control (CS:AV), respecto de las dos combinaciones restantes en el agrupamiento de la izquierda (CD:AV y CS:AM).

En el panel derecho inferior de la figura 4.7 se observa el biplot de la interacción triple de $Tejido \times Colina \times Aceite$ (T:C:A), el cual se encuentra amplificado para una mejor visualización en la figura 4.8, donde se observan dos paneles que emplean la misma codificación (panel izquierdo) o con números (panel derecho). En esta interacción la PC1 explica el 85,35 % de la variabilidad total. Esta componente agrupa

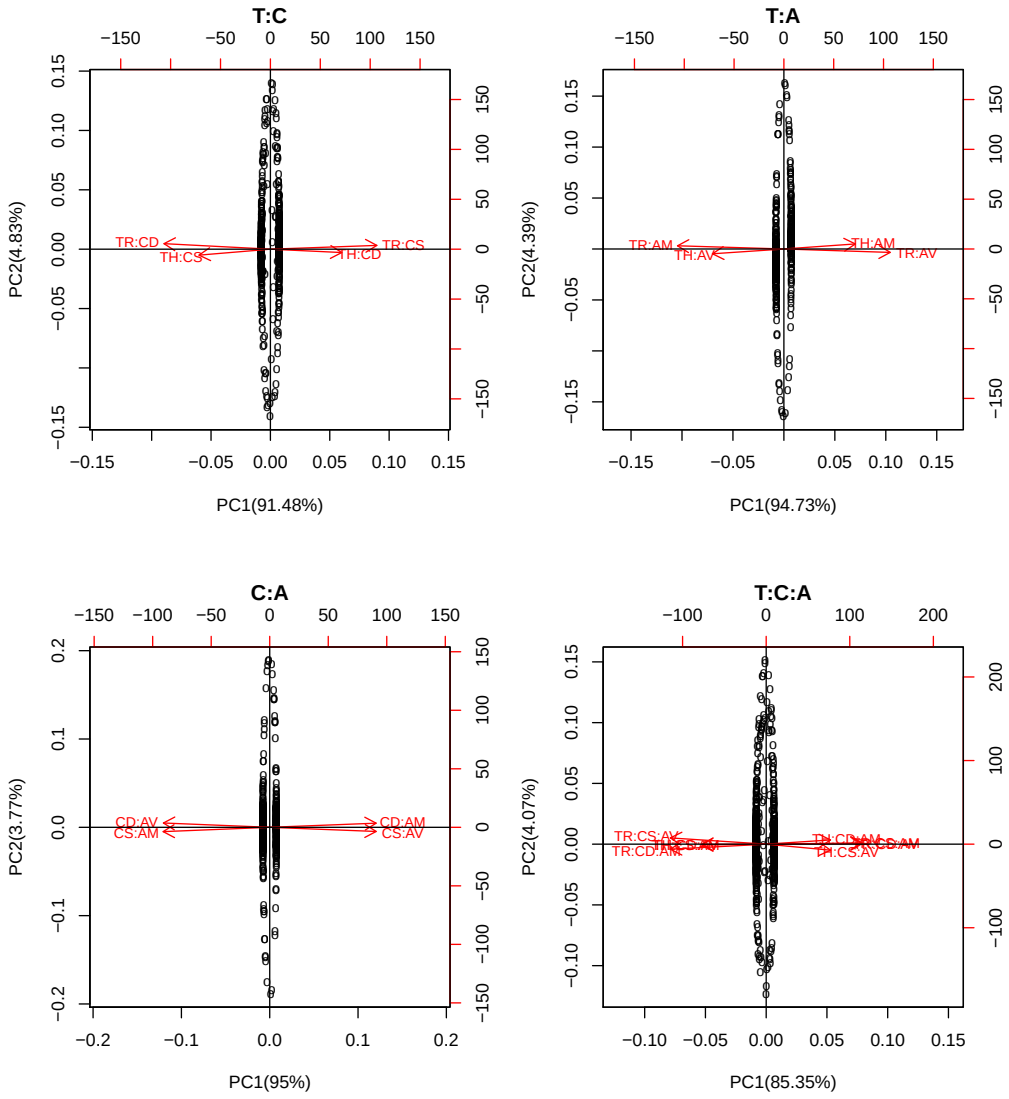


Figura 4.7: Biplots del análisis ASCA de los términos de interacción doble y el único triple del modelo (4.3), para los genes obtenidos para la prueba F con un valor $p < 0{,}05$ mostrados en la tabla 4.1. Note que la PC1 explica más del 91 % de la variabilidad de cada biplot. En particular los paneles de T:C, T:A y T:C:A muestran una menor variabilidad del hígado (TH) con flechas más cortas en la PC1. En el panel de C:A se agrupan a la derecha del origen el control (CS:AV) y el efecto protector de aceite de pescado (CD:AM).

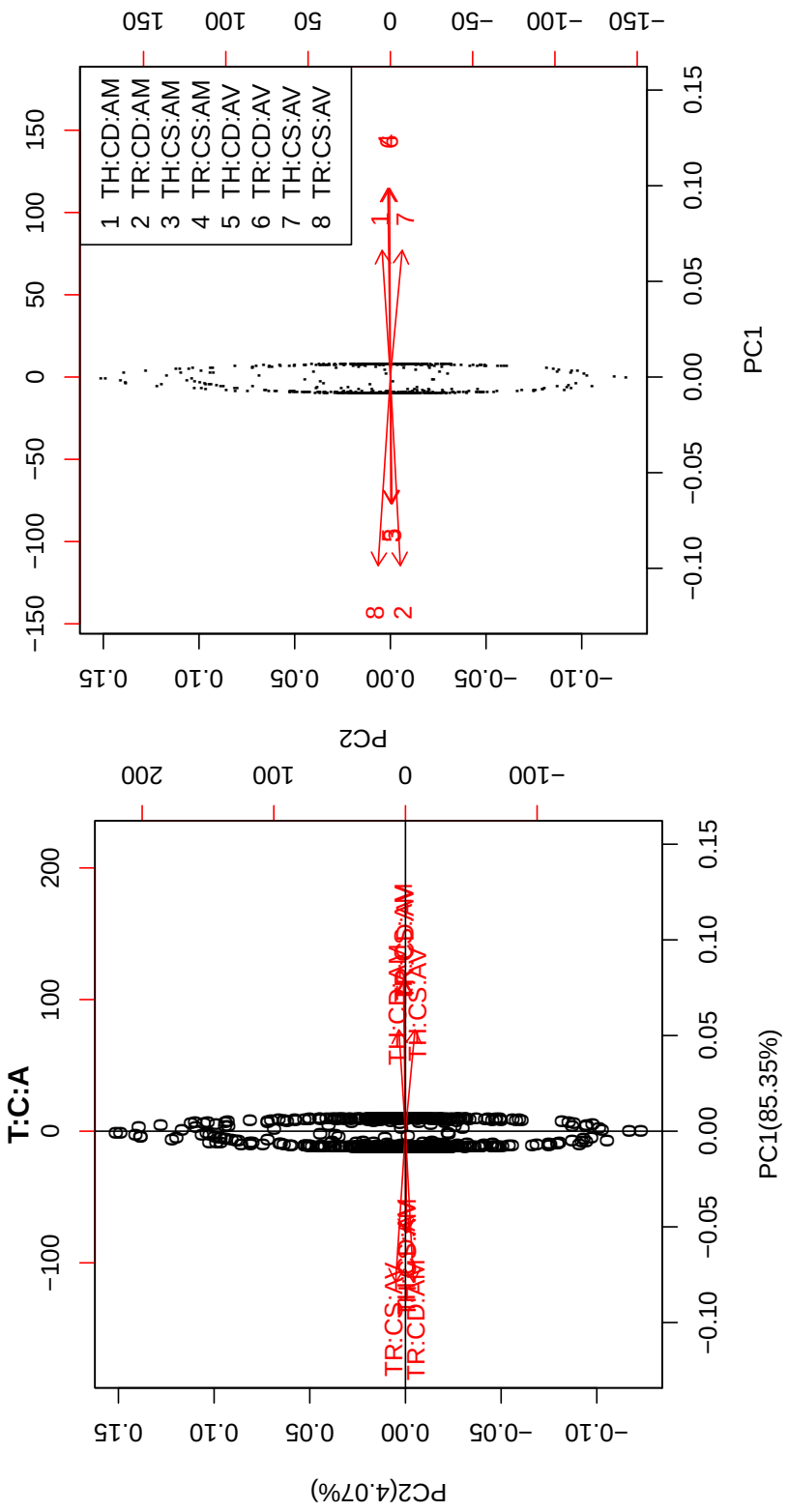


Figura 4.8: Biplots para el término de interacción triple $Tejido \times Colina \times Aceite$. En el panel izquierdo se muestra una ampliación del biplot de la figura 4.7 utilizando la combinación de tratamientos. En el panel de la derecha se recodificaron la combinación de tratamientos con número para poder apreciar cuáles se encuentran correlacionados.

las combinaciones que presentan hígado (TH), más cercanas al origen (menor variabilidad) respecto a las del riñón (TR). A su vez para el riñón, se agrupan del lado izquierdo las mismas combinaciones de *Colina* × *Aceite* presentes en la interacción doble (control y tratamiento con aceite de pescado) contra las restantes. Curiosamente, los mismos niveles para el hígado se encuentran en el agrupamiento opuesto (derecho). Estos resultados sugieren, que en la interacción triple se aprecia el efecto protector visto en la interacción doble de C:A (panel izquierdo inferior de la figura 4.7), pero ahora considerando la respuesta tejido-específica del riñón, mientras que el hígado presenta una variabilidad secundaria menor; dado que éste no es el órgano blanco, es esperable que no se agrupen con las mismas condiciones del riñón.

Reducción, proyección e integración de datos

Motivados por los resultados de la exploración con `lmdme`, realizamos una reducción de datos a los efectos de encontrar aquellos genes que tienen diferencias significativas en su expresión, cuando se comparan los diferentes efectos del diseño experimental. Para ello se utilizó un modelo lineal equivalente al descompuesto por ANOVA en la ecuación (4.3), obteniendo para cada gen (4.4):

$$gen_{ij} = \beta_{0i} + \beta_{1i}TH + \beta_{2i}CD + \beta_{3i}AM + \beta_{4i}TH:CD + \beta_{5i}TH:AM + \beta_{6i}CD:AM + \beta_{7i}TH:CD:AM + \varepsilon_{ij} \quad (4.4)$$

con $i = 1, \ldots, N$, $j = 1, \ldots, j(i)$ dependiendo del tejido; donde gen_{ij} es la expresión de i-ésimo gen para la j-ésima replica biológica; el modelo se ha parametrizado para que la media global para cada gen (β_{0i}) corresponda a la combinación de factores del caso control, es decir, riñón bajo dieta de colina suplementada con aceite vegetal (TR:CS:AV); $\beta_{0i,\ldots,7i}$ son coeficientes a estimar por máxima verosimilitud; TH, CD, AM son variables indicadoras que toman los valores 0 o 1 para indicar hígado, colina deficiente y aceite de pescado respectivamente; $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ es el término de error aleatorio.

El modelo (4.4) se ajustó con la librería **limma** de R, utilizando una corrección empírica de Bayes (Smyth et al., 2011). Se seleccionaron aquellos genes expresados

Tabla 4.2: Cantidad de genes diferenciales dependiendo del criterio de corte utilizado ($FDR < \alpha$)

α	Azar	T	C	A	T:C	T:A	C:A	T:C:A	C_R	A_R	C:A$_R$	C_H	A_H	C:A$_H$
0.05	863	12244	4091	5156	3650	571	3524	1796	3377	3884	2946	2445	3790	2085
0.01	173	10438	2408	3436	2140	200	1848	883	1895	2465	1620	1139	2183	650
0.001	17	8309	1114	2010	1038	67	888	312	955	1341	820	361	988	148

La columna α indica el nivel de corte utilizado, mientras que la denominada "Azar" indica la cantidad de genes esperados por error para la totalidad de las 17.256 genes bajo análisis. El resto de las columnas indican el efecto utilizado en la prueba de hipótesis, donde cada letra indica tejido (T), colina (C), aceite (A) y los subíndices indican la hipótesis marginal realizada sobre el riñón (R) o hígado (H).

diferencialmente para el comportamiento de los factores principales, interacciones dobles, la única triple y las marginales para cada tejido bajos las hipótesis:

- **H0**: la expresión de la comparación específica es igual a cero.
- **H1**: la expresión de la comparación es distinta de cero (gen diferencial).

En todos los casos se utilizó un valor corregido por comparaciones múltiples, de manera de reducir las tasas de falsos positivos mediante el método de False Discovery Rate (FDR) (Benjamini y Hochberg, 1995), obteniendo los resultados de la tabla 4.2 para diferentes valores de corte. En este contexto se decidió utilizar $\alpha = 0{,}001$, que si bien genera una cantidad de genes diferenciales elevada para los efectos principales del modelo completo, la cantidad de genes esperados por azar es de sólo 17 genes. Además, para este corte aún se obtiene una cantidad de genes sobre los contrastes marginales de cada tejido que permite un modelado apropiado, es decir, no se confunden las condiciones en la comprobación visual mediante mapas de calor.

En la figura 4.9 se muestra el mapa de calor correspondiente a la prueba de hipótesis de la triple interacción (T:C:A), para los 312 genes diferenciales reportados en la tabla 4.2. En la figura se observan dos agrupamientos de tratamientos bien definidos: i) aquellos pertenecientes a la IRA con colina deficiente y aceite vegetal en riñón (DVR), ii) el resto de los tratamientos. A su vez, en este último agrupamiento se subdivide por tejido, es decir, todas las muestras de hígado (DVH, DMH, SMH y SVH) de las restantes de riñón (DMR, SMR, SVR). Cabe destacar que el hígado tiene valores de expresión menores (son más oscuros) en comparación a los de riñón. Estos resultados concuerdan con los obtenidos para el análisis ASCA realizado por el aporte de `lmdme` propuesto en esta tesis, como ha sido mostrado con anterioridad en los *biplots* de las figuras 4.7 y 4.8.

El resto de los mapas de calor de los diferentes genes candidatos de la tabla 4.2 mostraron resultados similares, es decir, que las condiciones experimentales se separan correctamente en cada contraste para el valor de corte utilizado.

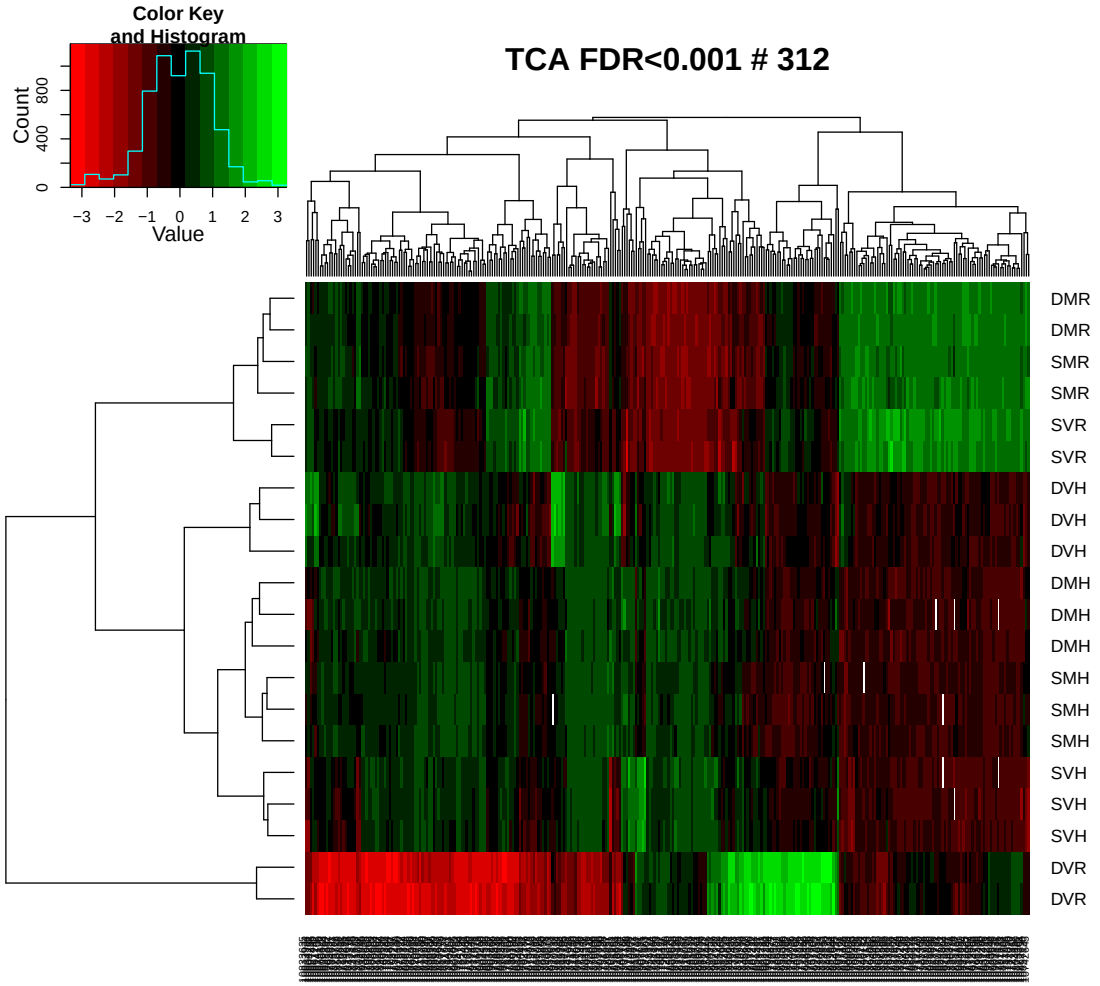


Figura 4.9: Mapa de calor para la triple interacción $Tejido \times Colina \times Aceite$ (TCA) para los genes diferenciales seleccionados para un $FDR < 0{,}001$ dando un total de 312 genes. En columnas los genes y en filas, las réplicas biológicas para cada condición experimental de colina deficiente (D) o suplementada (S), aceite de pescado (M) o vegetal (V) y tejido de riñón (R) o de hígado (H).

4.2.2. Modelado

La etapa de *modelado* se llevó a cabo de forma iterativa y progresiva, como parte del proceso de búsqueda de patrones desde la perspectiva del KDD y MD. En este contexto, el diseño experimental modelado en (4.4) es rico en estructura, dado que es un factorial $2 \times 2 \times 2 = 2^3$ (Walpole et al., 1999). Esta particularidad le aporta complejidad al análisis, razón por la cual se comenzó con una exploración de la distribución de los genes diferenciales.

En la figura 4.10 se muestran cuatro diagramas de Venn correspondientes a los genes diferenciales obtenidos para el corte seleccionado en la tabla 4.2:

1. Efectos principales: T, C y A.
2. Interacciones dobles: TC, TA y CA.
3. Efectos marginales del modelo en riñón: C_R, A_R y CA_R.

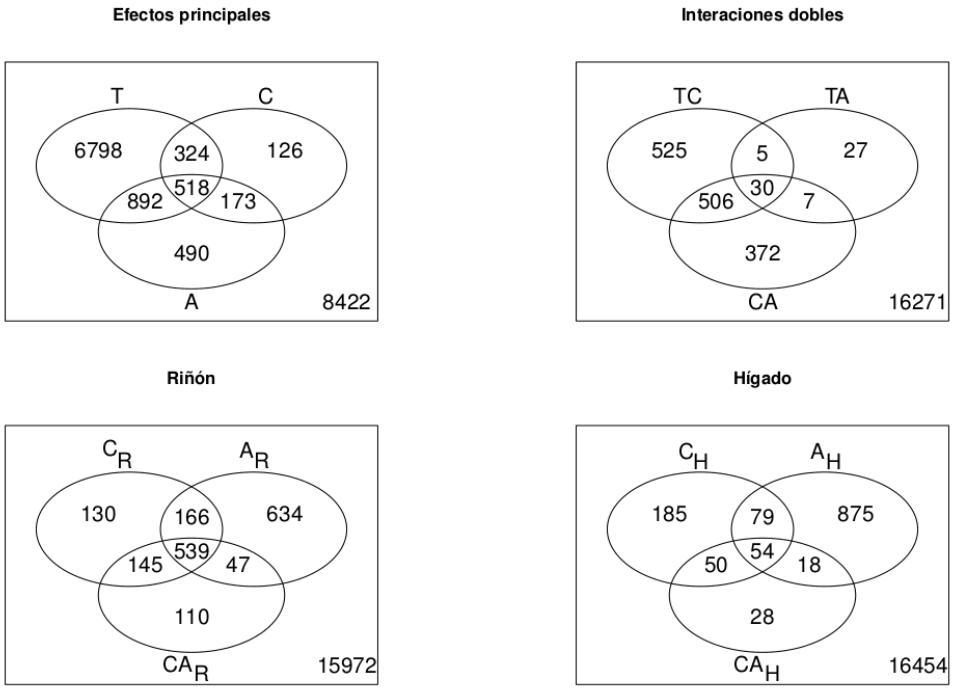


Figura 4.10: Diagrama de Venn para los diferentes genes diferenciales obtenidos con un $FDR < 0{,}001$ según se muestra en la tabla 4.2.

4. Efectos marginales del modelo en hígado: C_H, A_H y CA_H.

A partir de los resultados de cada uno de los diagramas de Venn, se realizó un análisis de enriquecimiento ontológico-funcional con el MRCM. Así, para los efectos principales mostrados en el panel superior izquierdo de la figura 4.10 se aplicó MRCM sobre los siguientes genes diferenciales:

I Efectos principales: $\#^1T = 8532$ genes, $\#C = 1141$ genes y $\#A = 2073$ genes.

II Genes pertenecientes exclusivamente a un efecto principal: $\#\{T \setminus (C \cup A)\}$[2]= 6798 genes, $\#\{C \setminus (T \cup A)\} = 126$ genes y $\#\{A \setminus (C \cup T)\} = 490$ genes.

III Genes pertenecientes a sólo dos efectos principales: $\#\{(T \cap C) \setminus A\} = 324$ genes, $\#\{(C \cap A) \setminus T\} = 173$ genes y $\#\{(A \cap T) \setminus C\} = 892$ genes.

IV Genes compartidos en los tres efectos principales $\#\{T \cap C \cap A\} = 518$ genes.

obteniendo los resultados ontológico-funcionales de la aplicación de diez MRCM. Adicionalmente, se utilizó una estrategia similar a la empleada en la aplicación de "*impacto funcional de variantes de FSH*" presentada en la sección 4.1. En este sentido, se realizó el contraste de los grafos unificados de los efectos principales de I de a pares (T vs C, T vs A y C vs A) y la única comparación triple de IV (T vs C vs A). De esta manera, para cada diagrama de Venn de la figura 4.10 se generan diez reportes del MRCM, junto con los cuatro contrastes de grafos unificados adicionales, ascendiendo a un total de $14\frac{reportes}{diagrama} \times 4\ diagramas = 56\ reportes$. En este punto la utilización del "**Contraste Ontológico**" fue crítico para manejar de forma eficiente la complejidad biológica y la inmensidad de información proveniente de los 56 reportes.

A partir de la exploración de los diferentes reportes, nos vimos sobrepasados por la dimensión/cantidad de información disponible para la exploración y consolidación de conocimiento, razón por la cual se iteró en la etapa de reportes a los efectos de generar un único grafo que contuviera toda la información de los diferentes reportes.

[1]El operador # es el cardinal del conjunto y devuelve el número de elementos del mismo.

[2]Para dos conjuntos A y B, la operación $A \setminus B$ devuelve aquellos elementos que pertenecen al conjunto A y que no se encuentran en el conjunto B. Los operadores $\cap$ y $\cup$ representan la intersección y unión de conjuntos respectivamente.

Si bien este grafo logró su fin, es decir, contiene la totalidad de los resultados, la posterior inspección no reveló un patrón biológico que fuera concluyente sobre el efecto protector del aceite de pescado. No obstante, existieron indicios de diferentes términos de GO compartidos en combinaciones específicas de *colina*×*aceite* y que no necesariamente se veían enriquecidos en los reportes de las correspondientes interacciones. Más aún, si pensamos que en riñón la deficiencia de colina es una precondición que debe existir para dar lugar a la IRA, en conjunto al complemento de aceite de vegetal o frente a su potencial protector (pescado). De esta manera, se decidió iterar sobre el modelado a los efectos de comparar de a pares las cuatro diferentes dietas en riñón: CSAV (control), CDAV (IRA), CSAP y CDAP. Justamente, las iteraciones planteadas en las diferentes etapas del KDD y MD al igual que la aplicación de los diferentes aportes de esta tesis (lmdme, MRCM y Contraste Ontológico), permitieron dilucidar el mecanismo de protección del AP como se presenta en la sección de evaluación.

4.2.3. Evaluación

La "**validación biológica**" del modelo nutricional bajo estudio se realizó mediante la determinación de las *alteraciones histopatológicas*, para establecer la presencia (o no) de necrosis tubular o cortical en el riñón izquierdo de cada rata. Los resultados mostraron que las ratas pertenecientes a los grupos de las dietas CSAV, CSAP y CDAP no mostraron alteraciones renales, mientras que las de CDAV presentaron necrosis renal cortical, como se describe en Denninghoff et al. (2014). Por otra parte, se midió en suero la *concentración de homocisteina*, *vitamina* B_{12} y *ácido fólico*. Se encontraron diferencias significativas en homocisteina (valor $p < 0{,}05$) para ratas alimentadas con AP (con y sin colina), niveles elevados de vitamina B_{12} en CDAV y no hubo diferencia en ácido fólico entre los grupos de ratas (Denninghoff et al., 2014).

Validado el modelo nutricional, se procedió al análisis de las comparaciones de pares de tratamientos de *colina* × *aceite* en riñón, obteniendo los resultados de la tabla 4.3 y diagrama de Venn de la figura 4.11. En ambos se aprecia cómo el efecto del aceite es marginal en dietas con CS, es decir, sólo 33 y 32 genes se expresan diferencialmente para las comparaciones B1 y B2 respectivamente. Esto sugiere un

Tabla 4.3: Comparación de diferentes dietas en riñón

Grupo	A1	A2	A3	B1	B2	B3
Comparación	CDAV vs CSAP	CDAV vs CDAP	CDAV vs CSAV	CDAP vs CSAV	CSAP vs CSAV	CDAP vs CSAP
Genes diferenciales	879	836	724	33	32	0
Necrosis Renal	X	X	X			
Colina deficiencia	X		X	X		X
Efecto protector del AP	X	X		X	X	

La dieta se compone de la combinación de colina deficiente (CD) o suplementada (CS), con aceite vegetal (AV) o de pescado (AP). Los genes han sido seleccionados con un $fdr < 0{,}01$ *y* $|log_2(FC)| > 1{,}5$*. Adaptación de Denninghoff et al. (2014).*

comportamiento similar entre CDAP y CSAV (control) y por ende, una potencial protección debida a la presencia de AP en la dieta. Por el contrario, la CD posee un efecto nocivo y el AV no produce mejoras (CDAV), como lo sugiere el elevado número de genes diferenciales al compararlo contra el control (CSAV) y dietas similares al control (CDAP y CSAP).

En la tabla 4.3 se muestra que el mayor efecto puede estar asociado a *necrosis renal*, debido a que las comparaciones A1-3 existen 542 genes en común, como se muestra en el diagrama de Venn de la figura 4.11. Estos genes enriquecen 3 vías metabólicas de KEGG (sección 1.1) fuertemente relacionadas con necrosis, como lo son "*cascada del complemento y coagulación* (15 genes)", "*interacción de receptores citoquina-citoquina* (14 genes)" y "*adhesión focal* (13 genes)". La comparación también muestra dos subconjuntos de 203 y 18 genes. Estos genes se encuentran alterados por la CDAV, dado que no presentan diferencias contra CSAV, CDAP y CSAP. Una posterior comprobación sobre el modelo (4.4) mostró, que los 203 genes se encuentran influenciados por el efecto principal de aceite. A su vez, el subconjunto de 18 genes se encuentra afectado por la CD. Por otra parte, existe un subconjunto adicional de 30 genes influenciados tanto por el efecto de la colina y aceite de pescado, lo que sugiere una interacción de tratamientos.

Motivados por los resultados anteriores se procedió a un análisis detallado de cada uno de los grupos de la tabla 4.3. En el anexo digital A.3 se encuentra una descripción completa de los genes y reportes del "Contraste Ontológico" utilizados

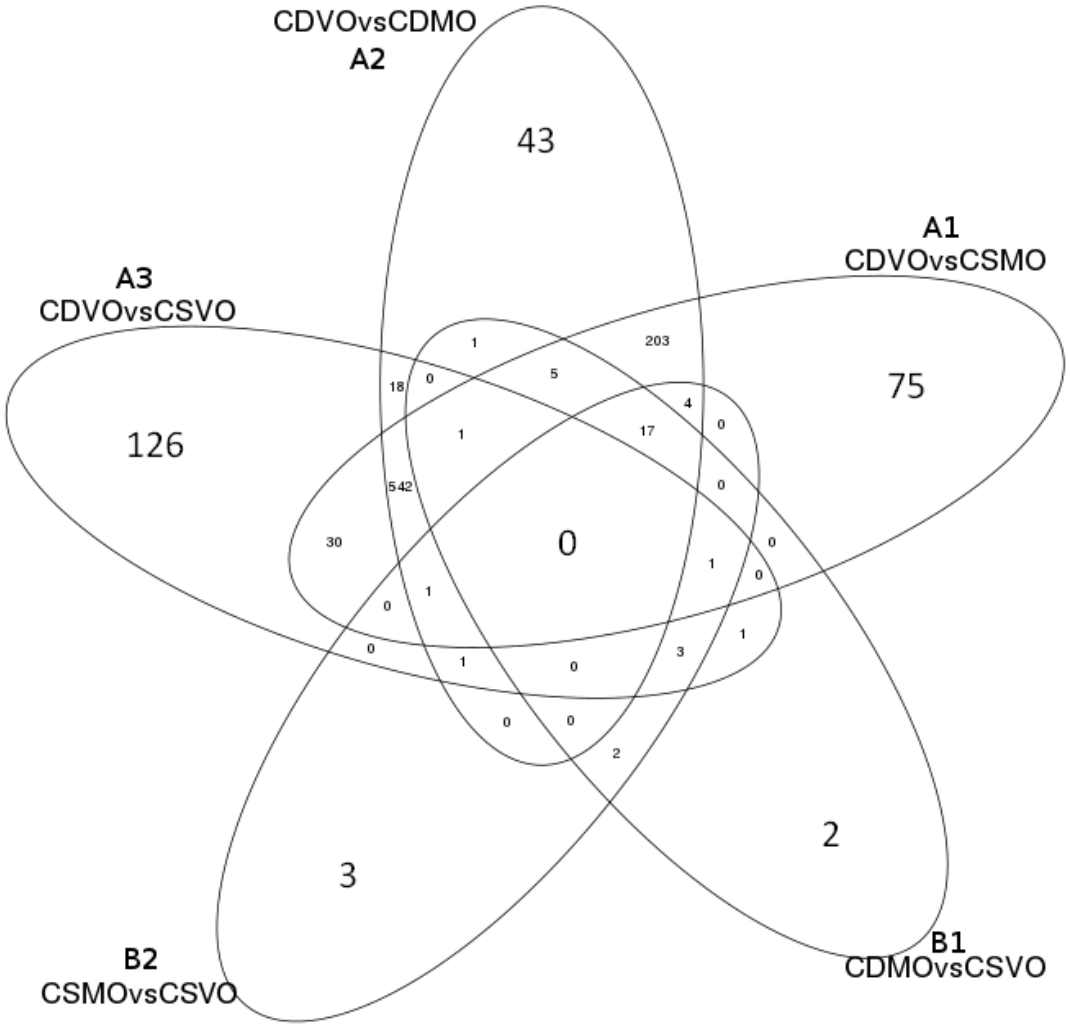


Figura 4.11: Diagrama de Venn para los diferentes genes diferenciales obtenidos con un $fdr < 0{,}01$ y $|log_2(FC)| > 1{,}5$ según se muestra en la tabla 4.3. La dieta se compone de la combinación de colina deficiente (CD) o suplementada (CS), con aceite vegetal (VO del inglés Vegetal Oil) o de pescado (MO del inglés Menhaden Oil). Imagen extraída de Denninghoff et al. (2014).

para cada grupo:

A1: La comparación de CDAV contra CSAP mostró el efecto combinado de aceite y colina. Si bien el grupo con CSAP no mostró necrosis renal, 879 genes se expresaron de forma diferencial al comparar su expresión contra CDAV. Denninghoff et al. (2014) no lograron una conclusión funcional acerca de este grupo, dado el efecto combinado de ambos factores.

A2: El contraste de CDAV contra CDAP mostró el efecto de aceite en CD. La combinación CDAP no presenta necrosis y se obtuvieron 836 genes diferenciales cuando se comparó contra CDAV, donde sí existe necrosis. El análisis con el MRCM para este grupo, permitió identificar vías metabólicas asociadas a la prevención de la necrosis renal por adición de AP en la dieta. En particular en PB se enriquecieron los términos relacionados con "*biosíntesis de metionina* (4 genes)", "*metabolismo de cisteina* (3 genes)", "*transulfuración de tirosina* (3 genes)", "*catabolismo de L-fenilalanina* (3 genes)" y "*biosíntesis de NAD* (5 genes)". La vista de expresión del "Contraste Ontológico" mostró que todos los genes se encuentran sobrexpresados en AP y a su vez, asociados al gen *Glutation-S-transferasa pi 1* cuyo símbolo es *Gstp1*, como se describe en Denninghoff et al. (2014). Por el contrario, los genes relacionados con *respuesta inflamatoria* o *respuesta inmune* se vieron sobrexpresados en AV, como por ejemplo en "*regulación positiva de hipersensibilidad tipo IIa* (9 genes)", "*fagocitosis* (13 genes)", "*regulación positiva de endocitosis* (13 genes)", "*regulación de diferenciación de macrófagos* (3 genes)", "*regulación positiva de diferenciación de mieloides* (3 genes)" y "*diferenciación de células gigantes de tromboblasto* (3 genes)". En el grafo de FM la sobrexpresión de genes en AV se relaciona a nodos de "*interleuquinas*" y "*citoquinas*", mientras que la sobrexpresión en AP a la "*actividad de ligando ácido-tiol*" y algunas funciones asociadas con el "*transporte de iones*" y "*ligando de glutation*". Sin embargo, en CC Denninghoff et al. (2014) no encontraron un patrón que permitiera establecer un lugar predominante para los PBs y FMs de interés.

A3: El contraste de CDAV contra CSAV mostró el efecto de colina bajo AV. El grupo alimentado con la dieta de control (CSAV), no presentó necrosis y se

obtuvieron 724 genes diferenciales al compararlo contra CDAV. Los resultados funcionales del MRCM se asociaron a términos genéricos de IRA y necrosis renal, los cuales no permitieron explicar el efecto protector de AP (no presente en esta comparación).

B1: El contraste CDAP contra CSAV mostró el efecto combinado de aceite y colina. En este caso, ambas condiciones experimentales no presentaron necrosis renal y sólo existe una diferencia de 33 genes diferenciales. Denninghoff et al. (2014) no lograron una conclusión funcional acerca de este grupo, dado el efecto combinado de ambos factores y lo reducida que es la lista de candidatos (sección 1.2.2). No obstante, el AP bajo CD logra un comportamiento cercano a la dieta de control (CSAV), por el reducido número de genes de esta comparación.

B2: La comparación CSAV contra CSAP mostró el efecto del aceite bajo CS. Ambas condiciones experimentales no presentan necrosis renal y sólo hay 32 genes diferenciales (11 sobre-expresados y 21 sub-expresados) como se describe en Denninghoff et al. (2014). Éste es el grupo "**biológicamente relevante**" dado que los 32 genes candidatos se encuentran solamente influenciados por el potencial efecto protector del AP y no se encuentran enmascarados por el cambio en colina como en B1. El análisis del MRCM no mostró términos enriquecidos pese a la reducida lista de genes candidatos, es decir que ambas condiciones se comportan de manera similar a nivel funcional por el reducido cambio de expresión en los genes. Motivados por estos resultados, Denninghoff et al. (2014) realizaron una búsqueda en la literatura sobre los 32 potenciales blancos terapéuticos. En este contexto, el conocimiento previo sobre el metabolismo de la colina, permitió identificar al gen *Gstp1* y comprender el rol crucial que su enzima puede tomar en el proceso de desintoxicación. Curiosamente, no existe en la literatura redes de interacción que relacionen a el *Gstp1* con el resto de los 31 genes encontrados en esta comparación.

B3: El último grupo compara CDAP contra CSAP, donde no se presentó diferencia en los niveles de expresión para el efecto colina bajo AP. Estas dos condiciones no mostraron necrosis ni diferencias en su transcriptoma, incluso relajando el corte a un $fdr < 0{,}05$. De esta manera, se podría hipotetizar que el AP

potencialmente protege la CD en este modelo, tanto a nivel morfológico como genético. La ingesta de AP en la dieta podría compensar la CD, lo que podría implicar que el AP protege tempranamente al riñón, previniendo la necrosis renal debido a la CD. Consecuentemente, ambas condiciones experimentales se comportan de manera similar en cuanto a la expresión de genes. Este resultado puede ser de relevancia clínica en pacientes con IRA.

A partir de los resultados anteriores, se volvió al modelo (4.4) a los efectos de obtener los valores de expresión esperados del *Gstp1* en cada una de las combinaciones de tratamientos de *colina* × *aceite*, los cuales se muestran en la figura 4.12. En este contexto, el gen posee un valor de expresión basal en la situación de control (CSVO) el cual se ve incrementado en una dieta CD (CDVO), y sorprendentemente mayor en presencia de AP (MO) con o sin colina. En estos últimos dos tratamientos la homocisteína en suero se ve aumentada, posiblemente por diferencia en el contenido de agentes exógenos y endógenos (Verhoef, 2007), producto de la ingesta de AP. También se puede atribuir a la síntesis endógena de colina a través de una triple metilación de fosfatidiletanolamina por la S-adenosilmetionina y la alteración en la vía metabólica de la tran-sulfuración (Denninghoff et al., 2014). Además, la sobreexpresión de *Gstp1* se correlaciona proporcionalmente con los elevados niveles

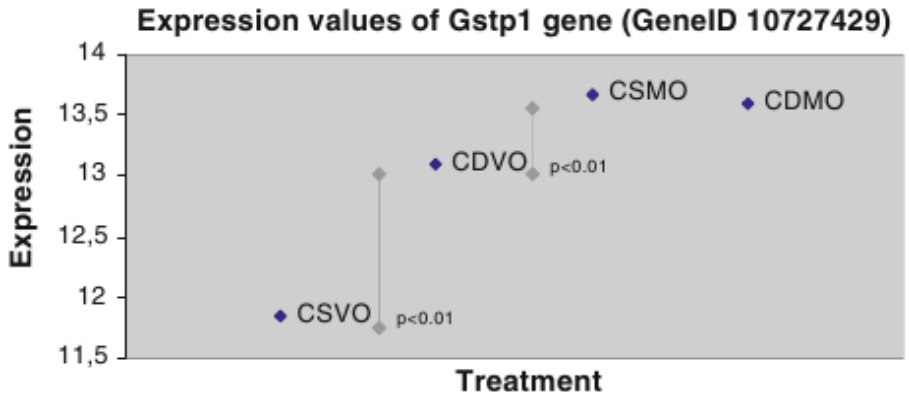


Figura 4.12: Nivel de expresión esperado del gen *Gstp1* en escala log_2, para las diferentes combinaciones de colina deficiente (CD) o suplementada (CS), con aceite vegetal (VO) o de pescado (MO). Note que la diferencia de niveles de CSVO contra CDVO y CDVO contra CSMO o CDMO son diferentes con un valor $p < 0{,}01$. Imagen extraída de Denninghoff et al. (2014).

de homocisteína, dado que es una de las enzimas involucradas en su catabolismo.

En la figura 4.12 también se muestra que existe expresión diferencial del *Gstp1* en presencia de IRA (CDVO) respecto del control (CSVO), con un valor $p < 0{,}01$. Esta diferencia se puede relacionar con el aumento de vitamina B_{12} en el suero, producto de la liberación de esta vitamina en túbulos renales necróticos (Scott et al., 1984; Ziegler y Filer, 1997). A la luz de los conocimientos de Denninghoff et al. (2014), existe una asociación entre la Gstp1 y estrés/daño oxidativo, el cual fue propuesto como mecanismo bioquímico de la IRA (Monserrat et al., 1969; Repetto et al., 2010), donde la CD se asocia con niveles elevados de peroxidación lipídica y daño oxidativo (Newberne et al., 1969; Ossani, 2012; Ossani et al., 2007; Repetto et al., 2010). Los ácidos grasos poliinsaturados; se ven afectados por el daño oxidativo en relación directa con el contenido de dobles ligaduras. En este sentido, el aceite de pescado es rico en ácidos grasos poliinsaturados, sin embargo, se produce una disminución en lugar de aumento del estrés oxidativo y lipoperoxidación en el riñón. Esto podría estar relacionado con el contenido intrínseco de antioxidantes y la sobreexpresión del gen Gstp1, en presencia de AP (MO) con o sin colina.

Por otra parte, es posible volver al modelo (4.4) a los efectos de analizar el comportamiento del gen *Gstp1* para cada efecto específico. En la tabla 4.4 se aprecia que para el **modelo completo**, existe un efecto de expresión diferencial a nivel de los tres efectos principales (T)ejido, (C)olina y (A)ceite y sólo para la interacción doble de C:A, es decir, que este gen no presenta interacción doble o triple dependiente del

Tabla 4.4: Efecto diferencial sobre el gen Gstp1

	Efecto						
Hipótesis	T	C	A	T:C	T:A	C:A	T:C:A
Modelo (4.4) completo	Si	Si	Si	No	No	Si	No
Marginales en riñón	-	Si	Si	-	-	Si	-
Marginales en hígado	-	No	Si	-	-	No	-

Las columnas indican el efecto utilizado en la prueba de hipótesis correspondiente a tejido (T), colina (C), aceite (A) y las correspondientes interacciones dobles y triples. En filas, el modelo en el cual se prueban las hipótesis sobre el gen *Gstp1* (modelo completo o marginales de cada tejido).

tejido en que se encuentre. A su vez, las hipótesis marginales en **riñón** muestran un comportamiento similar a los presentados en la figura 4.12. Adicionalmente, estos resultados son complementarios a los obtenidos para las comparaciones de las diferentes dietas de riñón mostrados en la tabla 4.3. Más aún, la Gstp1 es uno de los genes diferenciales mostrados para la interacción C:A en el biplot de la figura 4.7, obtenido en el análisis multivariado ASCA a través del aporte de `lmdme`, donde fue posible tempranamente asociar el efecto protector del AP con la dieta control, como se describió en la sección 4.2.1. En cambio en **hígado**, sólo se produce una diferencia en la expresión debido al efecto de aceite, probablemente por el cambio de mecanismos de generación de energía producto de la ingesta de AP en la dieta.

Los resultados anteriores muestran a la Gspt1 como un potencial "blanco" terapéutico órgano-específico para la IRA, relacionado con el efecto principal de aceite. En este sentido, es posible utilizar la idea del "MRCM" para contrastar las vías metabólicas de KEGG (sección 1.1) para el efecto aceite utilizando los genes candidatos del modelo completo, marginal para riñón e hígado de la tabla 4.2. El análisis mostró que las tres vías en las cuales participa la Gstp1, se encuentran enriquecidas por el consenso: "*metabolismo de drogas - citocromo P450 (30 genes)*", "*metabolismos de xenobióticos por citocromo P450 (23 genes)*" y "*metabolismo del glutatión (24 genes)*". En la figura 4.13 se muestra la vía del metabolismo de glutatión, donde se encuentran resaltados (en color rojo) los rectángulos de las reacciones enzimáticas, en las cuales participan alguno de los 24 genes expresados de forma diferencial. Particularmente, la Gstp1 participa en el recuadro con número enzimático 2.5.1.18, en el contexto del proceso de desintoxicación resaltado en la elipse de color. En este proceso, la Gstp1 emplea su capacidad para catalizar la conjugación de la forma reducida de glutatión con compuestos endógenos, sustratos xenobióticos y toxinas. Este precursor continúa una serie de transformaciones que culminan en la síntesis del ácido mercaptúrico, el cual es soluble y es excretado en orina como estrategia de desintoxicación.

En este contexto, Denninghoff et al. hipotetizan que la presencia de AP activa la vía de los xenobióticos y consecuentemente aumenta la expresión de Gstp1, la cual a su vez potencia la vía del metabolismo del glutatión, teniendo un rol crucial en el proceso de desintoxicación. De esta manera, la ausencia de colina en la dieta aunque con el suplemento de AP (CDAP), previene la aparición de IRA dado que está

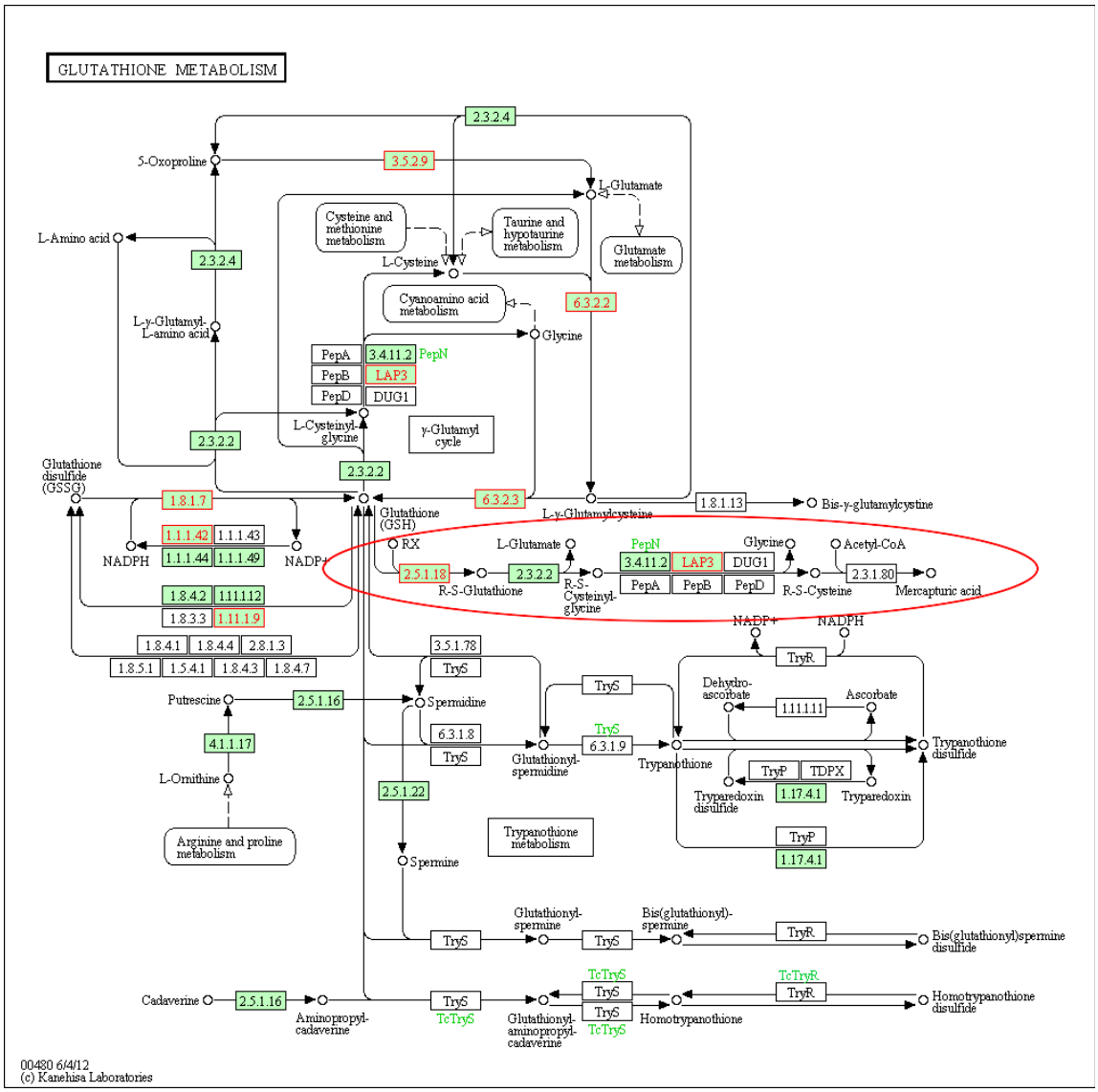


Figura 4.13: Metabolismo del glutatión según la ontología de KEGG (sección 1.1). En color rojo se encuentran resaltadas las enzimas donde participa alguno de los 24 genes diferenciales para el efecto aceite en el modelo (4.4). La elipse de color rojo resalta la participación de la Gstp1 en el proceso de desintoxicación, a través de la reacción enzimática 2.5.1.18, que termina en la síntesis de ácido mercaptúrico el cual es excretado en orina. Adaptación de la imagen de `www.genome.jp/kegg-bin/show_pathway?rno00480`.

incrementado el proceso de desintoxicación como lo sugiere la evidencia histológica. No obstante, para ello habría que medir el contenido de ácido mercaptúrico para las combinaciones de *colina* × *aceite*, para dar soporte a esta hipótesis. Los resultados previos han demostrado que el AP incluido en la dieta deficiente de colina produce un efecto protector en riñón. En la clínica, se podría incluir AP en la dieta de aquellos pacientes que son más propensos a desarrollar IRA, para prevenir su potencial aparición. Por otra parte, se continúa el análisis de potenciales efectos nocivos en hígado; a la fecha no se ha encontrado evidencia que contraindique su potencial uso.

4.2.4. Comentarios finales

La integración de los diferentes aportes metodológicos introducidos en esta tesis, han sido de gran utilidad para evaluar el *efecto protector del AP en el modelo nutricional de IRA* de Denninghoff et al. (2014) en las diferentes etapas del KDD.

En la etapa de *entendimiento de datos*, el control de calidad multivariado combinado con una descomposición ASCA a través de "**lmdme**" (sección 3.3), permitió en una etapa temprana del análisis explorar la adecuación de los datos y asociar diferentes niveles de *tejido* × *colina* × *aceite* con fenotipos y variabilidad esperadas por el diseño experimental, como se presentó en la sección 4.2.1.

La etapa de *modelado* hizo uso exhaustivo del **MRCM** (sección 3.5) para la exploración funcional de los efectos principales del modelo lineal (4.4), al igual que contraste a nivel funcional de los subconjuntos de genes diferenciales obtenidos de los diagramas de Venn de la figura 4.10 y su posterior contraste de grafos unificados. En este contexto, fueron de gran utilidad los reportes del "**Contraste Ontológico**" (sección 3.6) para navegar los cuarenta análisis preliminares. A partir de su exploración, se iteró sucesivamente sobre diferentes hipótesis que permitieron abordar diferentes aspectos de la *evaluación*. En este sentido, fue posible relacionar los resultados obtenidos de la "validación biológica" con la combinación de tratamientos de *colina* × *aceite* en riñón, para obtener mayor conocimiento del modelo nutricional. Este análisis permitió identificar una lista de posibles genes candidatos, donde el conocimiento previo del modelo permitió identificar a la proteína Gstp1 como potencial "blanco" terapéutico. Posteriormente, se relacionaron los niveles de expresión del gen *Gstp1* con los diferentes marcadores biológicos obtenidos en suero. Con esta

información en mente, se volvió sobre el modelo (4.4) para ver el comportamiento del gen *Gstp1*, es decir, poder discriminar qué efecto produce una expresión diferencial y se realizó un **MRCM** contrastando el efecto de aceite sobre el modelo completo, riñón e hígado, utilizando la ontología de KEGG. En el consenso se encontraron enriquecidas las vías metabólicas en las cuales participa la Gstp1. En particular, la vía del *metabolismo del glutatión* permitió plantear una nueva hipótesis sobre el mecanismo protector el aceite de pescado a través de la desintoxicación mediante ácido mercaptúrico en orina. A la fecha, esta hipótesis se encuentra bajo estudio, al igual que la búsqueda de efectos colaterales en hígado que contraindiquen su potencial uso en la clínica.

Capítulo 5

Conclusiones y trabajo futuro

En esta tesis se presentó la problemática del **análisis ontológico-funcional**, en el contexto de experimentos de alto rendimiento (capítulo 1). Para ello se introdujo el concepto de bases de datos con vocabulario controlado conocidas como *ontologías*, de las cuales se profundizó en GO y KEGG (sección 1.1), donde se mostró cómo se estructura la información en conceptos/categorías/términos/vías metabólicas. Luego en la sección 1.2 se mostraron las tres *metodologías de análisis* más conocidas en este campo (SEA, GSEA y MEA), para determinar qué funciones y/o vías metabólicas se encuentran modificadas (*enriquecidas*) por una lista de genes de interés del experimento, respecto de una lista de referencia. En el caso particular de SEA, se mostró cómo se construye la tabla de contingencia 1.1 y la problemática relacionada a la adecuada selección de la lista de referencia utilizada en el análisis. Especialmente en el caso de proteómica, donde no se cuenta a priori con esta última. Finalmente se presentaron tres *herramientas* para realizar SEA/MEA como lo son DAVID, GoMiner y GOstat/s (sección 1.3). En particular, se hizo hincapié en las destrezas/debilidades en lo que corresponde a la forma de acceso (con o sin conectividad a internet), carga de datos, metodologías disponibles, reproducibilidad de resultados (versión de bases de datos) y visualización (tablas, páginas web e imágenes). En este contexto, la problemática radica en la necesidad de utilizar las particularidades de múltiples herramientas para realizar un análisis, con la consecuente complejidad tanto de formateo/integración de datos (extensas tabla e imágenes estáticas) como en la exploración de reportes estáticos, sumado a la imposibilidad de abordar dise-

ños experimentales más allá de un típico caso-control. Justamente, estos problemas impactan negativamente en la extracción de patrones que pueda realizarse sobre la información biológica disponible, donde la aplicación de técnicas de minería de datos es de gran provecho en este campo.

En el capítulo 2 se describió el concepto de **minería de datos**, como un subproceso dentro de un contexto mucho más general como es el "descubrimiento de información en bases de datos", más comúnmente conocido por sus siglas en inglés KDD (sección 2.1). El KDD proporciona un marco de trabajo *ordenado*, *iterativo* e *interactivo*, *aportando* herramientas y *dirigiendo* el trabajo hacia la *búsqueda* de información relevante a través de la aplicación de algoritmos particulares, en el sentido de "revolver" y "escarbar" sobre los datos en la búsqueda de conocimiento. El abordaje en sí consta de distintas etapas que comprenden desde la conceptualización de los experimentos (sección 2.2), la obtención de los datos de entrada, el control de calidad, la adecuación e integración de distintas fuentes de información (sección 2.3), el análisis con las herramientas elegidas (sección 2.4), la evaluación de la solución (sección 2.5), hasta la presentación de los resultados mediante informes con visualizaciones apropiadas (2.6). Cada una de estas etapas ha sido particularizada en el contexto de la presente tesis, donde se ha hecho la revisión del estado del arte correspondiente.

En el capítulo 3 se describió brevemente el flujo de trabajo (sección 3.1) que involucra la minería de datos en análisis ontológico-funcionales. En particular el capítulo se focalizó en los diferentes **aportes realizados en esta tesis** en las diferentes etapas del KDD. En este sentido en la etapa de entendimiento de datos, se contribuyó en una alternativa frente a la problemática de *consistencia e integridad de anotación* a través de diferentes módulos para la conversión/actualización de identificadores en experimentos de alto rendimiento (sección 3.2). Para ello se desarrollaron diferentes módulos en lenguaje R, para incorporar plataformas de proteómica (2D-DIGE) y transcriptómica (microarreglos). Sin este aporte, un análisis tradicional puede sesgar el análisis sólo a aquellas proteínas/genes que se encuentren presentes en la base de datos de la herramienta correspondiente. Consecuentemente se excluyen, sin el consentimiento del usuario, potenciales candidatos en el análisis y/o incluyen identificadores obsoletos que no codifican proteínas. Por el contrario,

mediante la alternativa propuesta es posible tener una traza a lo largo del flujo de trabajo de cada proteína/gen, permitiendo así en una etapa temprana del análisis, tomar decisiones/estrategias para incorporar la mayor cantidad de información biológica disponible.

Continuando con la etapa del entendimiento de datos, se desarrolló una metodología para la *exploración multivariada y control de calidad* de datos de alto rendimiento (sección 3.3), mediante una descomposición ANOVA-PCA/PLS a través de modelos lineales (sección 3.3.1). Este aporte permite una exploración multivariada de los datos previo a la etapa de modelado, a los efectos de ver si los mismos se corresponden con lo que es esperable del diseño experimental (efectos principales, interacción, etc.), al igual que realizar un control de calidad multivariado de los datos (comprobar la inexistencia de fuentes de variabilidad no controlados), cómo se mostró con los dos ejemplos de evaluación de la sección 3.3.2. Más aún, el modelado mediante PLS permite ver si la variabilidad explicada responde a una estructura definida por el usuario, cómo se mostró para la fecha de hibridización en el ejemplo de control de calidad. Así este aporte, representa una alternativa válida frente a la adecuación tradicional de fuentes de variabilidad mediante técnicas de normalización (scale, RMA, etc.), que no permiten utilizar la información del diseño experimental. Adicionalmente, este desarrollo cuenta con una librería de R, `lmdme`, disponible en el repositorio de Bioconductor para la comunidad científica.

Pasando a la etapa de modelado del KDD, se desarrolló una librería en R que permite *conectividad con el portal DAVID* (sección 3.4), para realizar análisis ontológico-funcionales de forma programática. Esta característica no sólo permite estructuras de datos nativas de R, sino que también extiende el análisis al incorporar visualización de grafos de GO, como estrategia de resumen y exploración de los resultados, característica no disponible en DAVID. A su vez, la librería permite trabajar importando reportes obtenidos desde la plataforma web o desde R, facilitando el intercambio y reproducibilidad de resultados, sin necesidad de una conexión a internet. Sin este aporte, realizar análisis computacionalmente intensivos como validaciones de tipo bootstrap, resultan impracticables con intervención del usuario en el portal web. Adicionalmente, este desarrollo cuenta con una librería de R, `RDAVIDWebService`, disponible en el repositorio de Bioconductor para la comunidad científica.

Continuando con el modelado, se desarrolló una metodología de *integración y contraste de múltiples referencias* (MRCM, sección 3.5). Esta metodología permite integrar los resultados de *análisis realizados con múltiples referencias* (sección 3.5.1), a los efectos de ganar conocimiento biológico en el consenso/discrepancia de manera de asistir al investigador en la selección de términos biológicos de relevancia. Adicionalmente, es posible integrar/contrastar los resultados de diseños experimentales de mayor complejidad, de manera de dejar al descubierto términos específicos de un efecto principal, interacción, etc. Para ello, la metodología hace uso de `RDAVIDWebService` para obtener los resultados y se focaliza en emplear la estructura de grafo de GO, como estrategia de resumen visual (patrón de colores) y exploración de los resultados (nodos hoja). De esta manera, rápidamente es posible validar el diseño experimental al encontrar términos enriquecidos en el consenso y resaltar aquellos potenciales candidatos fuera del consenso. Sin este aporte, el investigador se ve obligado por una parte a explorar extensas tablas y por otra parte integrar los resultados provenientes de diferentes referencias/experimentos de forma no estructurada. Ambas situaciones tornan este tipo de análisis en una tarea tediosa y van en detrimento de la extracción de patrones que se pueda realizar, por la falta de aplicación de técnicas de minería de datos.

Por otra parte, haciendo uso de la capacidad de acceso programático de `RDAVIDWebService` se implementó un *análisis de estabilidad* utilizando validación por bootstrap (sección 3.5.2). A través de esta validación, el investigador cuenta con un valor de potencia para cada término, contando con un indicador de la robustez del mismo frente a perturbaciones introducidas a la lista de referencia. Esta es una característica sin precedentes en el análisis ontológico-funcional, que aporta información valiosa a la hora de seleccionar términos estables para la posterior validación biológica.

En conjunto el análisis de múltiples referencias y sensibilidad del MRCM, fue evaluado con datos reales de un experimento proteómico y dos transcriptómicos (sección 3.5.3 y 3.5.4). La posterior validación bibliográfica demostró que la metodología propuesta no sólo asiste al investigador en la exploración y selección de términos de interés, sino que recupera términos de relevancia biológica que en un análisis tradicional no se tendrían en cuenta.

Pasando a la etapa de reportes, para la *visualización y exploración de los resul-*

tados se desarrolló un reporte html llamado "Contraste Ontológico" (sección 3.6). Este reporte permite, a diferencia de otros, la exploración de los resultados sin conectividad a internet. Más aún, se puede personalizar la información que el usuario desee incluir en las tablas anexas. Por otra parte, incorpora la *vista de enriquecimiento* en la cual es posible explorar los grafos de GO de forma interactiva (ver qué genes participan de cada nodo, hipervínculos a PubMed, definición del término e información anexa). A esta vista se le suma la de *expresión* donde cada nodo muestra la cantidad de genes que posee sobre o subexpresados. Utilizando ambas vistas, el investigador puede navegar las categorías principales de GO (PB, FM y CC) de manera de obtener una visión unificada de los resultados funcionales de su experimento. Adicionalmente, este reporte incluye la totalidad de información disponible (sin filtro), siendo que en otras herramientas criterios ad hoc como la cantidad de genes podrían potencialmente excluir del análisis nodos relevantes para el experimento (sección 3.6.1).

Por último, en el capítulo 4 se abordaron dos **aplicaciones** reales: el impacto funcional de variantes de FSH en el experimento de Loreti et al. (sección 4.1) y efecto protector del AP en IRA en el modelo nutricional de Denninghoff et al. (sección 4.2). En cada una de estos experimentos se mostraron las diferentes etapas del KDD y MD involucradas en el análisis, haciendo *énfasis en la aplicación y versatilidad que ofrecen los diferentes aportes realizados en esta tesis*, presentados en el capítulo 3, para la búsqueda de información relevante en el contexto de la *minería de datos en el análisis ontológico-funcional*. En este sentido, cada una de las aplicaciones hizo hincapié en diferentes aspectos de los aportes de esta tesis, lo cual soporta la observación de Hedegaard et al. (2009), quienes sugieren que el análisis ontológico-funcional es un proceso exploratorio guiado por la biología, más que una certeza estadística. Más aún, en cada una de las aplicaciones, se mostró cómo los diferentes aportes contribuyeron a la ciencia en cada uno de los respectivos campos de aplicación.

En lo que respecta a trabajos futuros, en lo mediato se pretende incorporar los diferentes aportes realizados en la presente tesis como una extensión (*plug-in* del inglés) de la plataforma para la visualización e integración de redes biológicas, Cytoscape, Smoot et al. (2011), para darles mayor visibilidad en la comunidad científica a los mismos. Adicionalmente, se podrán incorporar al análisis ontológico-funcional

metodologías complementarias disponibles en Cytoscape como por ejemplo análisis de redes de interacción y predicción de funciones de genes con la extensión de `GeneMANIA` (Warde-Farley et al., 2010), o alguna otra del almacén de aplicaciones de Cytoscape, `apps.cytoscape.org`, Lotia et al. (2013).

En cuanto a nuevas líneas futuras pensando en un plan post-doctoral, se pretende continuar con el análisis ontológico-funcional aplicado a cáncer de mama. En este contexto la información de expresión de genes ha sido utilizada para clasificar diferentes tipos de cáncer; en particular, uno de los trabajos pioneros fue el de Perou et al., que en el año 2000 determinaron la clasificación no supervisada del cáncer de mama en subtipos intrínsecos, basada en experimentos de microarreglos (Perou et al., 2000). A partir de esta prueba de concepto surgieron numerosos trabajos que dieron origen a las conocidas "*firmas moleculares*" que se enfocaron en pronosticar la recurrencia del cáncer de mama y predecir la respuesta a la quimioterapia convencional (Haibe-Kains et al., 2012; Hu et al., 2009; Miller et al., 2005; van't Veer et al., 2002). Más aún, muchos trabajos han sido llevados a cabo para validar los subtipos intrínsecos de cáncer de mama (Basal, HER2, Luminal A, Luminal B y Normal) sobre diferentes bases de datos (Fan et al., 2006; Lusa et al., 2007; Sotiriou et al., 2006; Weigelt et al., 2010). Sin embargo, algunos trabajos han mostrado que varias de estas firmas tienen similares poderes pronósticos y/o predictivos entre ellas (Fan et al., 2006), mientras que otros han informado discrepancias en la clasificación de un mismo paciente entre las distintas firmas (Weigelt et al., 2010), dejando una fuerte controversia sobre su valor diagnóstico (Sørlie et al., 2001).

En cada uno de los distintos análisis existentes sobre firmas moleculares, la caracterización funcional es parcial y no hay un análisis integrador para validar si la segmentación vía firma molecular tiene un correlato funcional. En este contexto, en la literatura no se ha visto una caracterización sistemática de los diferentes tipos de cáncer a nivel funcional, es decir, donde el investigador relacione los niveles de expresión y la información clínica, con funciones biológicas y vías metabólicas conocidas (Bauer-Mehren et al., 2009). En realidad, se sospecha que la propia complejidad de fenómeno bajo estudio es abordada sólo parcialmente desde cada firma molecular, probablemente debido a que las distintas firmas moleculares seleccionan un subconjunto diferente de genes y así muestran retratos parciales de la red de interacciones

a nivel sistémico. Se hipotetiza que la integración de información en un contexto de funcionalidad biológica puede contribuir a describir los mecanismos que se activan, comparten y difieren entre los diferentes tipos de cáncer de mama. Justamente, a nivel funcional, será posible integrar información de diferentes poblaciones y similares tipos de cáncer, dejando atrás problemas técnicos de incorporación/compatibilidad al utilizar diferentes plataformas tecnológicas (Agilent®, Affymetrix®, etc.) cuando se realiza un análisis de expresión diferencial. De este modo se mejorará la comprensión del fenómeno biológico bajo estudio, integrando información de expresión genómica, clínica y ontológica, para establecer si existen características funcionales que distinguen los diferentes tipos moleculares definidos para el cáncer de mama. Para ello se utilizará la información disponible en repositorios de libre acceso como Gene Expression Omnibus, Edgar et al. (2002), `www.ncbi.nlm.nih.gov/geo`.

Apéndice A

Anexo Digital

En este anexo se describen los archivos que se encuentran incorporados en el CD que acompaña al documento impreso de tesis y no se encuentran libremente disponibles, como material suplementario de alguna de las publicaciones desarrolladas en el transcurso del doctorado.

A.1. Consistencia e integridad de anotación

En el capítulo 3 correspondiente a los diferentes *aportes realizados al análisis ontológico-funcional* dentro del marco de la presente tesis, se desarrollaron dos módulos para asistir el problema de consistencia e integridad de anotación presentado en la sección 3.2, mediante la conectividad a diferentes bases de datos: uno para comunicarse con **Uniprot** (Apweiler et al., 2004) y otro para consultar a **Entrez Gene** (Maglott et al., 2011) a través de e-utiles (National Center for Biotechnology Information, 2010).

A.1.1. uniprot.R

Consiste de un archivo de código fuente (*script* del inglés) llamado "`uniprot.R`", el cual se encuentra escrito en lenguaje R (R Core Team, 2013). Este script utiliza la librería `RCurl` (Lang, 2013a) para acceder de forma programática a la interfaz web de UniProt.

Este script permite cargar un objeto `uniprot` en R, que posee diferentes atributos como por ejemplo: cual es la dirección web `Base` de la interfaz (`www.uniprot.org`), qué herramientas (`Tool`) están disponibles (anotación, consulta y convertidor de identificadores), cual es el formato (`Format`) de los reportes (.txt, .tab, etc.), las columnas (`Columns`) seleccionadas por defecto, etc. A través del objeto `uniprot`, es posible generar consultas desde R para luego obtener los resultados de una búsqueda (`Query`) o de la conversión de IDs (`Mapping`), como se describe en la sección 3.2.1. Adicionalmente, en el script se encuentra documentado la forma de invocación de las diferentes rutinas, para asistir al usuario en su utilización.

A.1.2. eutiles.R

El sitio web de NCBI, `www.ncbi.nlm.nih.gov`, posee una interfaz de servicios web para su acceso de forma programática llamado **E-utiles** (National Center for Biotechnology Information, 2010). Este servicio web permite realizar prácticamente, todas las acciones que puede realizar un usuario desde la página `www.ncbi.nlm.nih.gov/sites/gquery?itool=toolbar`, las cuales pueden agruparse en tres categorías:

ELink: permite **vincular/convertir** los IDs presentes en una base de datos a otra.

ESummary: permite obtener la **información de resumen** (nombre, símbolo, estado actual, etc.) de los IDs solicitados, para una base de datos en particular.

ESearch: permite realizar una **búsqueda**, utilizando diferentes campos (nombre, símbolos, organismo, etc.) sobre las diferentes bases de datos, dependiendo de la información que posea el usuario.

En esta tesis se desarrolló un script llamado "`eutiles.R`" escrito en lenguaje R (R Core Team, 2013) y requiere de una cuenta académica solicitada a `eutilities@ncbi.nlm.nih.gov` para su utilización. El script emplea la librería **RCurl** (Lang, 2013a) para acceder de forma programática a la interfaz de **E-utiles**. Las consultas se realizan utilizando algunas de las tres funcionalidades disponibles, es decir, **ELink**, **ESummary** y **ESearch**. No obstante, los resultados de cada una de estas funciones se encuentran en formato XML (de las siglas en inglés, eXtensible Markup Language),

los cuales son adaptados utilizando la librería **XML** (Lang, 2013b), para su posterior utilización en R como se mostró en la sección 3.2.3.

A.2. Datos de ejemplo para control de calidad en microarreglos

En el capítulo 3 en el contexto de la *exploración multivariada y control de calidad* presentado en la sección 3.3, se utilizaron los datos de que aún *no se encuentran publicados* del Laboratorio de Terapia Molecular y Celular de la Fundación Instituto Leloir dirigido por el Dr. Osvaldo Podhajcer, a quien debo agradecer por permitir el uso de los mismos en la presente tesis.

En este experimento se utilizaron *microarreglos* de dos colores (sección 2.3.1) para explorar los perfiles de expresión génico. Los niveles de expresión se midieron en diferentes puntos de tiempo, bajo diversas concentraciones de proteínas incluidas en el medio de cultivo de líneas celulares de melanoma independientes. Este experimento posee una estructura ANOVA a dos vías: donde el factor A representa el "*tiempo*" con $a = 3$ niveles $\{0{,}5;\ 4;\ 12\ horas\}$, el factor B representa la "*concentración*" con $b = 3$ niveles $\{0;\ 1;\ 10\ unidades\}$ para $r = 3$ replicas biológicas, dando un total de 27 muestras. En el archivo "`example2.RData`" se encuentran almacenado los dos objetos R correspondiente a la matriz de diseño (`design`) y la matriz de expresión (`M`) como se describe en la sección 3.3.2.

A.3. Reportes del efecto protector del aceite de pescado en IRA

En el capítulo 4 en el contexto de la aplicación de los diferentes aportes desarrollados en esta tesis, para la búsqueda del *efecto protector del aceite de pescado en IRA* presentado en la sección 4.2, se anexan los diferentes archivos adicionales a los publicados por Denninghoff et al. (2014):

Venn.tiff: Imagen con el diagrama de Venn de la figura 4.11.

Genes.xls: Hoja de cálculo que posee la información de la totalidad de genes de los microarreglos utilizados en el análisis. En ella se encuentran las siguientes columnas:

- AffyID: el identificador correspondiente a Affymetrix® para cada sonda.
- EntrezId, GeneSymbol y GeneDescription: corresponden a la información de los registros de la base de datos de Entrez Gene para cada gen, donde se encuentra el identificador (EntrezId), cual es el símbolo o nemónico (GeneSymbol) y el nombre del gen (GeneDescription).
- log2FC.XXvsYY: posee la diferencia de expresión medida en escala log_2 entre los tratamientos de riñon XX y YY, donde $XX, YY \in \{CDVO, CSVO, CDMO, CSMO\}$.
- p.adj.XXvsYY: es el False Discovery Rate (FDR) (Benjamini y Hochberg, 1995) de los tratamientos de riñon XX y YY, donde $XX, YY \in \{CDVO, CSVO, CDMO, CSMO\}$.
- Comparison: columna que indica a que subconjunto del diagrama de Venn (Venn.tiff) pertenece el gen.

Contraste Ontológico: una carpeta que incluye los reportes de los grupos A1-A3 de la tabla 4.3, es decir, CDAM vs CDAP, CDAM vs CSAM y CDAM vs CSAP. En cada una de ellas se encuentran las siguientes carpetas:

- In: carpeta que posee los tres archivos necesarios para realizar un MRCM, es decir, BRII, BRIII y los genes diferenciales como se describe en la sección 3.5.
- Out: carpeta que posee los resultados del "Functional Annotation Chart" de DAVID obtenido con RDAVIDWebService como se describe en la sección 3.4.
- Reporte HTML: una carpeta con el nombre del contraste que posee el reporte del Contraste Ontológico propiamente dicho. Dentro de esta carpeta se encuentra un archivo llamado `index.html` que abre el reporte para su exploración, para lo cual se recomienda utilizar el navegador Chrome®.

Bibliografía

Abdi, H. y Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Abel, M. H., Wootton, A. N., Wilkins, V., Huhtaniemi, I., Knight, P. G., y Charlton, H. M. (2000). The effect of a null mutation in the follicle-stimulating hormone receptor gene on mouse reproduction. *Endocrinology*, 141(5):1795–1803.

Affymetrix, I. (2004). *GeneChips Expression Analysis: Data Analysis Fundamentals, Part no. 701190, Rev. 4*.

Al-Shahrour, F., Díaz-Uriarte, R., y Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580.

Alibés, A., Yankilevich, P., Cañada, A., y Díaz-Uriarte, R. (2007). IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics*, 8:9.

Alvarez, M. J. (2006). *Rol de la proteína de matriz extracelular SPARC en la progresión tumoral del melanoma humano*. PhD thesis, Instituto de Investigaciones Bioquímica, Universidad de Buenos Aires, Argentina.

Alvarez, M. J., Prada, F., Salvatierra, E., Bravo, A. I., Lutzky, V. P., Carbone, C., Pitossi, F. J., Chuluyan, H. E., y Podhajcer, O. L. (2005). Secreted protein acidic and rich in cysteine produced by human melanoma cells modulates polymorphonuclear leukocyte recruitment and antitumor cytotoxic capacity. *Cancer Research*, 65(12):5123–5132.

Anwar, A., Norris, D. A., y Fujita, M. (2011). Ubiquitin proteasomal pathway mediated degradation of p53 in melanoma. *Archives of Biochemistry and Biophysics*, 508(2):198–203.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(suppl 1):D115–D119.

Archer, K. J. y Reese, S. E. (2010). Detection call algorithms for high-throughput gene expression microarray data. *Briefings in Bioinformatics*, 11(2):244–252.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., y Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Au, E., Richter, M. W., Vincent, A. J., Tetzlaff, W., Aebersold, R., Sage, E. H., y Roskams, A. J. (2007). SPARC from olfactory ensheathing cells stimulates schwann cells to promote neurite outgrowth and enhances spinal cord repair. *The Journal of Neuroscience*, 27(27):7208–7221.

Bagshaw, S. M. (2006). The long-term outcome after acute renal failure. *Current Opinion in Critical Care*, 12(6):561–566.

Barnes, P. J. (2009). Histone deacetylase-2 and airway disease. *Therapeutic Advances in Respiratory Disease*, 3(5):235–243.

Barrios-de Tomasi, J., Nayudu, P., Brehm, R., Heistermann, M., Zariñán, T., y Ulloa-Aguirre, A. (2006). Effects of human pituitary FSH isoforms on mouse follicles in vitro. *Reproductive Biomedicine Online*, 12(4):428–441.

Batista, G. E. y Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533.

Bauer-Mehren, A., Furlong, L. I., y Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, 5(1).

Beissbarth, T. y Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

Bellei, B., Maresca, V., Flori, E., Pitisci, A., Larue, L., y Picardo, M. (2010). p38 regulates pigmentation via proteasomal degradation of tyrosinase. *Journal of Biological Chemistry*, 285(10):7288–7299.

Bellomo, R. (2006). The epidemiology of acute renal failure: 1975 versus 2005. *Current Opinion in Critical Care*, 12(6):557–560.

Benjamini, Y. y Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, p:289–300.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., y Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bolton, E. E., Wang, Y., Thiessen, P. A., y Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, 4:217–241.

Bourgon, R., Gentleman, R., y Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences USA*, 107(21):9546–9551.

Brenner, B. M. (2004). *The Kidney*. W. B. Saunders.

Bussey, K. J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay, W., y Weinstein, J. N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology*, 4(4):R27.

Castillo-Davis, C. I. y Hartl, D. L. (2003). GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892.

Charpentier, T. H., Thompson, L. E., Liriano, M. A., Varney, K. M., Wilder, P. T., Pozharski, E., Toth, E. A., y Weber, D. J. (2010). The effects of CapZ peptide (TRTK-12) binding to S100B-Ca2+ as examined by NMR and X-ray crystallography. *Journal of Molecular Biology*, 396(5):1227–1243.

Chavey, C., Boucher, J., Monthouël-Kartmann, M.-N., Sage, E. H., Castan-Laurell, I., Valet, P., Tartare-Deckert, S., y Obberghen, E. (2006). Regulation of secreted protein acidic and rich in cysteine during adipose conversion and adipose tissue hyperplasia. *Obesity*, 14(11):1890–1897.

Chun, S. Y., Eisenhauer, K., Minami, S., Billig, H., Perlas, E., y Hsueh, A. (1996). Hormonal regulation of apoptosis in early antral follicles: follicle-stimulating hormone as a major survival factor. *Endocrinology*, 137(4):1447–1456.

Cohen, J. et al. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Courrèges, M. C., Caruso, C., Klein, J., y Monserrat, A. J. (2002). Protective effect of menhaden oil on renal necrosis occurring in weanling rats fed a methyl-deficient diet. *Nutrition Research*, 22(9):1077–1089.

Day, R. y Lisovich, A. (2010). *DAVIDQuery: Retrieval from the DAVID bioinformatics data resource into R.* R package version 1.20.0.

De Haan, J., Wehrens, R., Bauerschmidt, S., Piek, E., Van Schaik, R., y Buydens, L. (2007). Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics*, 23(2):184–190.

Denninghoff, V., Ossani, G., Uceda, A., Rugnone, M., Fernández, E., Fresno, C., González, G., Díaz, M. L., Avagnina, A., Elsner, B., y Monserrat, A. (2014). Molecular pathology of acute kidney injury in a choline-deficient model and fish oil protective effect. *European Journal of Nutrition*, 53(3):897–906.

Dennis Jr, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., y Lempicki, R. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(5):P3.

Dieudonné, S. C., Kerr, J. M., Xu, T., Sommer, B., DeRubeis, A. R., Kuznetsov, S. A., Kim, I.-S., Gehron Robey, P., y Young, M. F. (2000). Differential display of human marrow stromal cells reveals unique mrna expression patterns in response to dexamethasone. *Journal of Cellular Biochemistry*, 76(2):231–243.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., y Tainsky, M. A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31(13):3775–3781.

Edgar, R., Domrachev, M., y Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Erlich, H. A. (1989). Polymerase chain reaction. *Journal of Clinical Immunology*, 9(6):437–447.

Eroglu, C. (2009). The role of astrocyte-secreted matricellular proteins in central nervous system development and function. *Journal of Cell Communication and Signaling*, 3(3-4):167–176.

Falcon, S. y Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., y Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine*, 355(6):560–569.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., y Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. The MIT Press.

Fernández, E. A., Girotti, M. R., del Olmo, J. A. L., Llera, A. S., Podhajcer, O. L., Cantet, R. J. C., y Balzarini, M. (2008). Improving 2D-DIGE protein expression analysis by two-stage linear mixed models: assessing experimental effects in a melanoma cell study. *Bioinformatics*, 24(23):2706–2712.

Fewster, M. E. y Hall, M. O. (1967). The renal phospholipid composition of choline-deficient rats. *Lipids*, 2(3):239–243.

FlyBase Consortium (1994). FlyBase - the Drosophila database. *Nucleic Acids Research*, 22(17):3456–3458.

Fresno, C., , Balzarini, M. G., y Fernández, E. A. (2014). lmdme: Linear models on designed multivariate experiments in R. *Journal of Statistical Software*, 56(7):1–16.

Fresno, C. y Fernández, E. A. (2013a). *lmdme: Linear Model decomposition for Designed Multivariate Experiments*. R package version 1.3.1.

Fresno, C. y Fernández, E. A. (2013b). RDAVIDWebService: a versatile R interface to DAVID. *Bioinformatics*, 29:2810–2811.

Fresno, C. y Fernández, E. A. (2013c). *RDAVIDWebService: An R Package for retrieving data from DAVID into R objects using Web Services API.* R package version 0.99.1.

Fresno, C. y Fernández, E. A. (2013d). *stemHypoxia: differentiation of human embryonic stem cells under hypoxia gene expression dataset by Prado-Lopez et al. (2010).* R package version 0.99.3.

Fresno, C., Llera, A. S., Girotti, M. R., Valacco, M. P., López, J. A., Podhajcer, O. L., Balzarini, M. G., Prada, F., y Fernández, E. A. (2011). Contraste Ontológico: una herramienta para el análisis de experimentos de proteómica/genómica funcional. In *Anales de las 40º Jornadas Argentinas de Informática (JAIIO), Sociedad Argentina de Informática (SADIO), Congreso Argentino de Informática en Salud (CAIS).*

Fresno, C., Llera, A. S., Girotti, M. R., Valacco, M. P., López, J. A., Podhajcer, O. L., Balzarini, M. G., Prada, F., y Fernández, E. A. (2012). The multi-reference contrast method: Facilitating set enrichment analysis. *Computers in Biology and Medicine*, 42(2):188–194.

Gautier, L., Cope, L., Bolstad, B. M., y Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315.

GE, H. (2008). *DeCyder 2D Software. User Manual. Version 7.0.*

Geladi, P. y Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., y Dudoit, S. (2005). *Bioinformatics and computational biology solutions using R and Bioconductor*, volume 746718470. Springer New York.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., y Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., Sarkar, D., y Hansen, K. D. (2013). *Rgraphviz: Provides plotting capabilities for R graph objects.* R package version 2.4.1.

Girotti, M. R., Fernández, M., López, J. A., Camafeita, E., Fernández, E. A., Albar, J. P., Benedetti, L. G., Valacco, M. P., Brekken, R. A., Podhajcer, O. L., y Llera, A. S. (2011). SPARC promotes cathepsin B-mediated melanoma invasiveness through a collagen i/$\alpha 2\beta 1$ integrin axis. *Journal of Investigative Dermatology*, 131(12):2438–2447.

Gordon, A. (2010). *Classification, 2nd Edition.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Graybill, F. (2000). *Theory and application of the linear model*. Duxbury classic series. Duxbury.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5):625–640.

Hackstadt, A. J. y Hess, A. M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10:11.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., y Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–325.

Halestrap, A. P. y Meredith, D. (2004). The SLC16 gene family-from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond. *Pflugers Arch*, 447(5):619–628.

Han, J., Kamber, M., y Pei, J. (2011). *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hanson, B. A. (2012). *ChemoSpec: Exploratory Chemometrics for Spectroscopy*. R package version 1.51-2.

Harrington, P. d. B., Vieira, N., Espinoza, J., Nien, J., Romero, R., y Yergey, A. (2005). Analysis of Variance–Principal Component Analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta*, 544(1):118–127.

Hedegaard, J., Arce, C., Bicciato, S., Bonnet, A., Buitenhuis, B., Collado-Romero, M., Conley, L. N., Sancristobal, M., Ferrari, F., Garrido, J. J., Groenen, M. A. M., Hornshøj, H., Hulsegge, I., Jiang, L., Jiménez-Marín, A., Kommadath, A., Lagarrigue, S., Leunissen, J. A. M., Liaubet, L., Neerincx, P. B. T., Nie, H., van der Poel, J., Prickett, D., Ramirez-Boo, M., Rebel, J. M. J., Robert-Granié, C., Skarman, A., Smits, M. A., Sørensen, P., Tosser-Klopp, G., y Watson, M. (2009). Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. *BMC Proceedings*, 3 Suppl 4:S5.

Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E., y Garrels, J. I. (1999). The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Research*, 27(1):69–73.

Hohenadl, C., Mann, K., Mayer, U., Timpl, R., Paulsson, M., y Aeschlimann, D. (1995). Two adjacent N-terminal glutamines of BM-40 (osteonectin, SPARC) act as amine acceptor sites in transglutaminase-catalyzed modification. *Journal of Biological Chemistry*, 270(40):23415–23420.

Hornik, K. (2012). The comprehensive R archive network. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):394–398.

Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C., y Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70.

Hu, Z., Fan, C., Livasy, C., He, X., Oh, D. S., Ewend, M. G., Carey, L. A., Subramanian, S., West, R., Ikpatt, F., et al. (2009). A compact VEGF signature associated with distant metastases and poor outcomes. *BMC Medicine*, 7(1):9.

Huang, D. W., Sherman, B. T., y Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Huang, D. W., Sherman, B. T., y Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., y Lempicki, R. A. (2007). DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35(Web Server issue):W169–W175.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., y DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., y Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., y Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806.

Joshi-Tope, G., Gillespie, M., Vastrik, I., DÉustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432.

Kanehisa, M. y Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kelly, K. A., Allport, J. R., Amy, M. Y., Sinh, S., Sage, E. H., Gerszten, R. E., y Weissleder, R. (2007). SPARC is a VCAM-1 counter-ligand that mediates leukocyte transmigration. *Journal of Leukocyte Biology*, 81(3):748–756.

Khatri, P. y Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595.

Khatri, P., Draghici, S., Ostermeier, G. C., y Krawetz, S. A. (2002). Profiling gene expression using onto-express. *Genomics*, 79(2):266–270.

Kumar, T. R., Wang, Y., Lu, N., y Matzuk, M. M. (1997). Follicle stimulating hormone is required for ovarian follicle maturation but not male fertility. *Nature Genetics*, 15(2):201–204.

Lang, D. T. (2012). *SSOAP: Client-side SOAP access for S*. R package version 0.9-1.

Lang, D. T. (2013a). *RCurl: General network (HTTP/FTP/...) client interface for R*. R package version 1.95-4.1.

Lang, D. T. (2013b). *XML: Tools for parsing and generating XML within R and S-Plus*. R package version 3.98-1.1.

Loreti, N. (2012). *Implicancia del grado de sialilación y complejidad de los oligosacáridos de FSH en la regulación de la función ovárica: estudios clínicos y experimentales*. PhD thesis, Instituto de Investigaciones Bioquímica, Universidad de Buenos Aires, Argentina.

Loreti, N., Fresno, C., Barrera, D., Andreone, L., Albarran, S. L., Fernandez, E. A., Larrea, F., y Campo, S. (2013). The glycan structure in recombinant human FSH affects endocrine activity and global gene expression in human granulosa cells. *Molecular and Cellular Endocrinology*, 366(1):68–80.

Loreti, N., Fresno, C., Fernández, E., y Campo, S. (2011). The integration of expression, experimental and ontology information allow us to identify specific effects induced by FSH carbohydrate moiety. In *2do Congreso Argentino de Bioinformática y Biología Computacional*.

Lotia, S., Montojo, J., Dong, Y., Bader, G. D., y Pico, A. R. (2013). Cytoscape App Store. *Bioinformatics*, 29(10):1350–1351.

Luna, C., Li, G., Qiu, J., Epstein, D. L., y Gonzalez, P. (2009). Role of miR-29b on the regulation of the extracellular matrix in human trabecular meshwork cells under chronic oxidative stress. *Molecular Vision*, 15:2488.

Lusa, L., McShane, L. M., Reid, J. F., De Cecco, L., Ambrogi, F., Biganzoli, E., Gariboldi, M., y Pierotti, M. A. (2007). Challenges in projecting clustering results across gene expression–profiling datasets. *Journal of the National Cancer Institute*, 99(22):1715–1723.

López, M., Mallorquín, P., y Vega, M. (2005). *Aplicaciones de los microarrays y biochips en salud humana: Informe de vigilancia tecnológica*. Genoma España.

Madeja, Z., Master, A., Michalik, M., y Sroka, J. (2001a). Contact-mediated acceleration of migration of melanoma B16 cells depends on extracellular calcium ions. *Folia Biologica-Krakow*, 49(3/4):113–124.

Madeja, Z., Szymkiewicz, I., Żaczek, A., Sroka, J., Miękus, K., y Korohoda, W. (2001b). Contact-activated migration of melanoma B16 and sarcoma XC cells. *Biochemistry and Cell Biology*, 79(4):425–440.

Maere, S., Heymans, K., y Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449.

Maglott, D., Ostell, J., Pruitt, K. D., y Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue):D52–D57.

Marouga, R., David, S., y Hawkins, E. (2005). The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Analytical and Bioanalytical Chemistry*, 382(3):669–678.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., y Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, 5(12):R101.

Martinek, N., Shahab, J., Sodek, J., y Ringuette, M. (2007). Is SPARC an evolutionarily conserved collagen chaperone? *Journal of Dental Research*, 86(4):296–305.

Masseroli, M., Martucci, D., y Pinciroli, F. (2004). GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Research*, 32(suppl 2):W293–W300.

McClintick, J. N. y Edenberg, H. J. (2006). Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, 7:49.

McGinnis, S. y Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server issue):W20–W25.

McGrath-Morrow, S., Rangasamy, T., Cho, C., Sussan, T., Neptune, E., Wise, R., Tuder, R. M., y Biswal, S. (2008). Impaired lung homeostasis in neonatal mice exposed to cigarette smoke. *American Journal of Respiratory Cellular Molecular Biology*, 38(4):393–400.

Mevik, B.-H., Wehrens, R., y Liland, K. H. (2011). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.3-0.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the USA*, 102(38):13550–13555.

Monserrat, A., Ghoshal, A., Hartroft, W., y Porta, E. (1969). Lipoperoxidation in the pathogenesis of renal necrosis in choline-deficient rats. *The American Journal of Pathology*, 55(2):163.

Monserrat, A., Musso, A., Tartas, N., Nicastro, M., Konopka, H., Arienti di García, I., y JC, S. A. (1981). Consumption coagulopathy in acute renal failure induced by hypolipotropic diets. *Nephron*, 28(6):276–284.

Monserrat, A. J., Cutrin, J. C., y Coll, C. (2000). Protective effect of myristic acid on renal necrosis occurring in rats fed a methyl-deficient diet. *Research in Experimental Medicine*, 199(4):195–206.

Monserrat, A. J., Porta, E. A., Ghoshal, A. K., y Hartman, S. (1974). Sequential renal lipid changes in weanling rats fed a choline-deficient diet. *The Journal of Nutrition*, 104(11):1496–1502.

Monserrat, A. J., Romero, M., Lago, N., y Aristi, C. (1995). Protective effect of coconut oil on renal necrosis occurring in rats fed a methyl-deficient diet. *Renal Failure*, 17(5):525–537.

Nadon, R. y Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends in Genetics*, 18(5):265–271.

Nakayama, K. (2010). Growth and progression of melanoma and non-melanoma skin cancers regulated by ubiquitination. *Pigment Cell & Melanoma Research*, 23(3):338–351.

National Center for Biotechnology Information (2010). Entrez Programming Utilities Help. http://www.ncbi.nlm.nih.gov/books/NBK25501/.

Newberne, P., Bresnahan, M., y Kula, N. (1969). Effects of two synthetic antioxidants, vitamin E, and ascorbic acid on the choline-deficient rat. *The Journal of Nutrition*, 97(2):219–231.

Nishi, Y., Yanase, T., Mu, Y.-M., Oba, K., Ichino, I., Saito, M., Nomura, M., Mukasa, C., Okabe, T., Goto, K., et al. (2001). Establishment and characterization of a steroidogenic human granulosa-like tumor cell line, KGN, that expresses functional follicle-stimulating hormone receptor. *Endocrinology*, 142(1):437–445.

Nueda, M., Conesa, A., Westerhuis, J., Hoefsloot, H., Smilde, A., Talón, M., y Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA. *Bioinformatics*, 23(14):1792–1800.

O´Neal, R., Still, W., y Hartroft, W. (1961). Increased lipotropic requirements with renal necrosis induced in rats by high-fat diets. *The Journal of Nutrition*, 75(3):309–318.

Orallo, J., Quintana, M., y Ramírez, C. (2004). *Introducción a la minería de datos*. Fuera de colección Out of series. Pearson Educación.

Ossani, G. (2012). *Deficiencia de colina y patogenia de la insuficiencia renal aguda*. PhD thesis, Department of Pathology, Universidad de Buenos Aires, Argentina.

Ossani, G., Dalghi, M., y Repetto, M. (2007). Oxidative damage lipid peroxidation in the kidney of choline-deficient rats. *Frontiers in Bioscience: a Journal and Virtual Library*, 12:1174.

Packer, L. M., Pavey, S. J., Boyle, G. M., Stark, M. S., Ayub, A. L., Rizos, H., y Hayward, N. K. (2007). Gene expression profiling in melanoma identifies novel downstream effectors of p14ARF. *International Journal of Cancer*, 121(4):784–790.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill Interamericana de España S.L.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Pfaff, M., Aumailley, M., Specks, U., Knolle, J., Zerwes, H. G., y Timpl, R. (1993). Integrin and Arg-Gly-Asp dependence of cell adhesion to the native and unfolded triple helix of collagen type VI. *Experimental Cell Research*, 206(1):167–176.

Phan, J. H., Moffitt, R. A., Stokes, T. H., Liu, J., Young, A. N., Nie, S., y Wang, M. D. (2009). Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends in Biotechnology*, 27(6):350–358.

Pinheiro, J. y Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., y Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics*, 109(6):678–680.

Prado-Lopez, S., Conesa, A., Armiñán, A., Martínez-Losa, M., Escobedo-Lucea, C., Gandia, C., Tarazona, S., Melguizo, D., Blesa, D., Montaner, D., et al. (2010). Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium. *Stem Cells*, 28(3):407–418.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32:496–501.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Repetto, M. G., Ossani, G., Monserrat, A. J., y Boveris, A. (2010). Oxidative damage: The biochemical mechanism of cellular injury and necrosis in choline deficiency. *Experimental and Molecular Pathology*, 88(1):143–149.

Richards, J. S., Russell, D. L., Ochsner, S., Hsieh, M., Doyle, K. H., Falender, A. E., Lo, Y. K., y Sharma, S. C. (2002). Novel signaling pathways that control ovarian follicular development, ovulation, and luteinization. *Recent Progress in Hormone Research*, 57(1):195–220.

Rivals, I., Personnaz, L., Taing, L., y Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407.

Robker, R. L. y Richards, J. S. (1998). Hormone-induced proliferation and differentiation of granulosa cells: a coordinated balance of the cell cycle regulators cyclin D2 and p27Kip1. *Molecular Endocrinology*, 12(7):924–940.

Sacks-Wilner, R. y Freddo, T. F. (1990). Differences in nuclear pore density among human choroidal melanoma cell types. *Ultrastructural Pathology*, 14(4):311–319.

Sawhney, R. S. (2002). Expression and regulation of SPARC, fibronectin, and collagen IV by dexamethasone in lens epithelial cells. *Cell Biology International*, 26(11):971–983.

Schellings, M. W., Pinto, Y. M., y Heymans, S. (2004). Matricellular proteins in the heart: possible role during stress and remodeling. *Cardiovascular Research*, 64(1):24–31.

Scott, J., Treston, A., Bowman, E., Owens, J., y Cooksley, W. (1984). The regulatory roles of liver and kidney in cobalamin (vitamin B12) metabolism in the rat: the uptake and intracellular binding of cobalamin and the activity of the cobalamin-dependent enzymes in response to varying cobalamin supply. *Clinical Science*, 67(Pt 3):299–306.

Shah, N. y Fedoroff, N. V. (2004). CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, 20(7):1196–1197.

Shawe-Taylor, J. y Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Simon, J., Scheig, R., y Klatskin, G. (1968). Relationship of early lipid changes in kidney and liver to the hemorrhagic renal necrosis of choline-deficient rats. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, 19(5):503.

Smilde, A., Jansen, J., Hoefsloot, H., Lamers, R., Van Der Greef, J., y Timmerman, M. (2005). ANOVA–Simultaneous Component Analysis (ASCA): A new tool for analysing designed metabolomics data. *Bioinformatics*, 21(13):3043–3048.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., y Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3.

Smyth, G. K., Ritchie, M., Silver, J., Wettenhall, J., Thorne, N., Langaas, M., Ferkingstad, E., Davy, M., Pepin, F., Choi, D., McCarthy, D., Wu, D., Oshlack, A., de Graaf, C., Hu, Y., Shi, W., y Phipson., B. (2011). *limma: Linear Models for Microarray Data*. R package version 3.12.1.

Smyth, G. K. y Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4):265–273.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Sosa, M. S., Girotti, M. R., Salvatierra, E., Prada, F., de Olmo, J. A. L., Gallango, S. J., Albar, J. P., Podhajcer, O. L., y Llera, A. S. (2007). Proteomic analysis identified n-cadherin, clusterin, and HSP27 as mediators of SPARC (secreted protein, acidic and rich in cysteines) activity in melanoma cells. *Proteomics*, 7(22):4123–4134.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in

breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272.

Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., y Brody, J. S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the Natural Academy of Sciences of the USA*, 101(27):10143–10148.

Staal, F. J. T., van der Burg, M., Wessels, L. F. A., Barendregt, B. H., Baert, M. R. M., van den Burg, C. M. M., van Huffel, C., Langerak, A. W., van der Velden, V. H. J., Reinders, M. J. T., y van Dongen, J. J. M. (2003). DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324–1332.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., y Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the Natural Academy of Sciences of the USA*, 102(43):15545–15550.

Tarazona, S., Prado-López, S., Dopazo, J., Ferrer, A., y Conesa, A. (2012). Variable selection for multifactorial genomic data. *Chemometrics and Intelligent Laboratory Systems*, 110(1):113–122.

Tarca, A. L., Romero, R., y Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195(2):373–388.

Ulloa-Aguirre, A., Timossi, C., Damián-Matsumura, P., y Dias, J. A. (1999). Role of glycosylation in function of follicle-stimulating hormone. *Endocrine*, 11(3):205–215.

Urbanek, S. (2013). *rJava: Low-level R to Java interface*. R package version 0.9-4.

Vadlamuri, S. V., Sankey, S. S., Nakeff, A., Divine, G., Rempel, S. A., et al. (2003). SPARC affects glioma cell growth differently when grown on brain ECM proteins in vitro under standard versus reduced-serum stress conditions. *Neuro-Oncology*, 5(4):244–254.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Verhoef, P. (2007). Homocysteine? an indicator of a healthy diet? *The American Journal of Clinical Nutrition*, 85(6):1446–1447.

Vincent, A. J., Lau, P. W., y Roskams, A. J. (2008). SPARC is expressed by macroglia and microglia in the developing and mature nervous system. *Developmental Dynamics*, 237(5):1449–1462.

Vitt, U., Kloosterboer, H., Rose, U., Mulders, J., Kiesel, P., Bete, S., y Nayudu, P. (1998). Isoforms of human recombinant follicle-stimulating hormone: comparison of effects on murine follicle development in vitro. *Biology of Reproduction*, 59(4):854–861.

Walpole, R. E., Myers, R. H., y Myers, S. L. (1999). *Probabilidad y estadística para ingenieros*. Pearson Educación.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl 2):W214–W220.

Webb, S. y Dobb, G. (2007). ARF, ATN or AKI? it's now acute kidney injury. *Anaesthesia and Intensive Care*, 35(6):843.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S., Dowsett, M., Ashworth, A., y Reis-Filho, J. S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, 11(4):339–349.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 35(suppl 1):D5–D12.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated.

Wilkinson, L. y Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(2).

Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.-Z., Ledley, R. S., Lewis, K. C., Mewes, H.-W., Orcutt, B. C., et al. (2002). The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Research*, 30(1):35–37.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., y Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15–e15.

Ympa, Y. P., Sakr, Y., Reinhart, K., y Vincent, J.-L. (2005). Has mortality from acute renal failure decreased? a systematic review of the literature. *The American Journal of Medicine*, 118(8):827–832.

Zambrano, E., Barrios-de Tomasi, J., Cárdenas, M., y Ulloa-Aguirre, A. (1996). Studies on the relative in-vitro biological potency of the naturally-occurring isoforms of intrapituitary follicle stimulating hormone. *Molecular Human Reproduction*, 2(8):563–571.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., y Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28.

Zeeberg, B. R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D. W., Reimers, M., Stephens, R. M., Bryant, D., Burt, S. K., Elnekave, E., Hari, D. M.,

Wynn, T. A., Cunningham-Rundles, C., Stewart, D. M., Nelson, D., y Weinstein, J. N. (2005). High-Throughput GoMiner, an índustrial-strengthíntegrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, 6:168.

Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C., y Weinstein, J. N. (2004). Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 5:80.

Zhang, S., Zhang, C., y Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381.

Zhong, S., Storch, K.-F., Lipan, O., Kao, M.-C. J., Weitz, C. J., y Wong, W. H. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Applied Bioinformatics*, 3(4):261–264.

Ziegler, E. y Filer, L. (1997). *Conocimientos actuales sobre nutrición*. Publicaciones Científicas. Organización Panamericana de la Salud, Oficina Sanitaria Panamericana, Oficina Regional de la Organización Mundial de la Salud.

Zuo, J., Tang, C.-j., Li, C., Yuan, C.-a., y Chen, A.-l. (2004). Time series prediction based on gene expression programming. In *Advances in web-age information management*, pages 55–64. Springer.

Zwanenburg, G., Hoefsloot, H., Westerhuis, J., Jansen, J., y Smilde, A. (2011). ANOVA-Principal Component Analysis and ANOVA-Simultaneous Component Analysis: A comparison. *Journal of Chemometrics*, 25(10):561–567.