VNiVERSiDAD D SALAMANCA

DESARROLLO Y ANÁLISIS BIOINFORMÁTICO DE MAPAS ÓMICOS DE INTERACCIÓN DE PROTEÍNAS Y DE COEXPRESIÓN DE GENES: REDES FUNCIONALES DERIVADAS

TESIS DOCTORAL

Carlos Prieto Sánchez

Febrero 2009

D. JAVIER DE LAS RIVAS SANZ, INVESTIGADOR CIENTÍFICO DEL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (C.S.I.C)

CERTIFICA:

Que ha sido el director del trabajo titulado "Desarrollo y análisis bioinformático de mapas ómicos de interacción de proteínas y de coexpresión de genes: redes funcionales derivadas" presentado por el doctorando D. CARLOS PRIETO SÁNCHEZ y reúne, a su juicio, originalidad y contenidos suficientes para ser presentado ante el correspondiente tribunal y optar al Título de Doctor por la Universidad de Salamanca.

Y para que así conste, a los efectos oportunos, expide el presente certificado en Salamanca a 30 de Diciembre de 2008.

Fdo. Dr. Javier De Las Rivas Sanz
Director de la Tesis Doctoral

D. ALBERTO ORFAO DE MATOS, PROFESOR TITULAR DEL DEPARTAMENTO DE MEDICINA DE LA UNIVERSIDAD DE SALAMANCA

CERTIFICA:

Que ha sido el tutor del trabajo titulado "Desarrollo y análisis bioinformático de mapas ómicos de interacción de proteínas y de coexpresión de genes: redes funcionales derivadas" presentado por el doctorando D. CARLOS PRIETO SÁNCHEZ y reúne, a su juicio, originalidad y contenidos suficientes para ser presentado ante el correspondiente tribunal y optar al Título de Doctor por la Universidad de Salamanca.

Y para que así conste, a los efectos oportunos, expide el presente certificado en Salamanca a 30 de Diciembre de 2008.

Fdo. Dr. Alberto Orfao De Matos
Tutor de la Tesis Doctoral

Esta memoria ha sido realizada siendo **Carlos Prieto Sánchez** beneficiario de una beca FPI de la Junta de Castilla y León (JCyL) para la realización tesis doctorales (2005-2009).

La investigación ha sido financiada por los siguientes proyectos:

- Estudio de redes funcionales de genes y proteínas por métodos bioinformáticos a partir de datos genómicos de expresión y de interacción: enfoque hacia nodos críticos y desregulados en cáncer. Fondo de Investigación Sanitario (FIS) PIO61153. 2007–2009. IP, Javier De Las Rivas.
- Identificación por métodos bioinformáticos de la red de genes que caracterizan un clasificador predictor de hemopatías malignas a partir de datos de microarrays de expresión genómica. Junta de Castilla y León (JCyL) CSI03A06. 2007–2009. IP, Javier De Las Rivas.
- Análisis estadístico y biológico por métodos bioinformáticos de resultados de chips genómicos de hemopatías malignas. Fondo de Investigación Sanitario (FIS) PIO30920. 2003–2006. IP, Javier De Las Rivas.

Parte del trabajo aquí reseñado ha sido publicado en los siguientes artículos científicos:

- Prieto C, Risueño A, Fontanillo C y De las Rivas J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *Plos One 3(12)*: e3911.
- Orchard S, *et al.* (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology 25(8)*: 894-8.
- Hernández-Toro J, Prieto C y De Las Rivas J (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics 23(18)*: 2495-7.
- Prieto C y De Las Rivas J (2006). APID: Agile Protein Interaction DataAnzalizer. *Nucleic Acids Res. 34 (Web Server issue)*: W298-302.
- Prieto C, Rivas-Lopez M J, Sánchez-Santos J M, Lopez-Fidalgo J y De Las Rivas J (2006). Algorithm to find gene expression profiles of de-regulation and identify families of disease-altered genes. *Bioinformatics 22(9)*: 1103-10.

ÍNDICE

INTRODUCCIÓN

En 1953, cuando Watson y Crick descubrieron la estructura de doble hélice del *DNA* (Watson y Crick, 1953), no se podían imaginar el formidable volumen de información biológica que se iría generando a partir de ese momento. Tan solo 50 años después, en abril de 2003, se anunció la finalización de la primera secuenciación y descifrado completo del genoma humano [1], lo que dió origen a la nueva y actual etapa de investigación en biología llamada era postgenómica. A su vez, en esta era postgenómica están emergiendo las denominadas ciencias "ómicas" (genómica, proteómica, transcriptómica, metabolómica, etc), apoyadas por el desarrollo de técnicas biomoleculares capaces de obtener cantidades masivas de información biológica, que pretenden abarcar el estudio global de todos los genes, las proteínas o las biomoléculas que trabajan en las células realizando los distintos procesos y funciones biológicas.

De forma paralela, la informática y las ciencias de la computación han sufrido una gran revolución desde que en 1943 se construyera ENIAC, el primer ordenador digital, hasta nuestros días, en los que los ordenadores se han convertido en un elemento de uso cotidiano en la sociedad. En el ámbito de la investigación científica, la informática es ya una herramienta clave para cualquier estudio y para el avance en cualquier área de conocimiento. Entre estas áreas dependientes de la computación están sin duda las ciencias biológicas "ómicas", que necesitan de la informática para manejar las grandes cantidades de datos que generan.

La convergencia de la tecnología y desarrollo informático-computacional con las áreas citadas de investigación biológica-biomolecular de carácter global ("ómico"), ha dado lugar a la aparición de una nueva disciplina académica y científica, la bioinformática. Según la librería nacional de medicina estadounidense (NLM) [2] la bioinformática es "la recopilación, clasificación, almacenamiento y análisis de información biológica y bioquímica mediante el uso de ordenadores". Pero, junto a esta definición sencilla de una nueva ciencia, hay que destacar que en el caso de la biología y la informática se dá una convergencia muy importante intrínseca a la naturaleza de ambas ciencias, que es el estar basadas en lenguajes y códigos como elementos substanciales constitutivos a

partir de los cuales se construyen:

- a partir del código genético y la secuencia de los genes y proteínas se construye toda la biología
- a partir del código fuente y los lenguajes de programación se construye toda la informática

Este paralelismo entre biología e informática hace seguramente que su confluencia en un nuevo marco de conocimiento sea mucho más poderoso de lo que actualmente vislumbramos, y augura una gran vitalidad para el desarrollo científico de la bioinformática. De hecho, la bioinformática ya se ha ido ampliando al ir surgiendo dentro de ella diversas sub-áreas de investigación adaptadas a los conocimientos y técnicas concretas sobre las que se centran. Algunos ejemplos son, la bioinformática estructural: análisis y comparación de secuencias, predicción de estructura de proteínas, clasificación de estructuras 3D; la genómica funcional: búsqueda y anotación de genes en genomas, análisis de expresión genómica, estudios de regulación de genes, genómica comparativa y evolutiva; la biología de sistemas: modelado de sistemas y procesos biológicos, análisis de redes biomoleculares complejas, etc.

El trabajo de investigación desarrollado en esta Memoria de Tesis Doctoral, ha sido realizado en diversas sub-áreas de la investigación bioinformática. En los dos primeros capítulos se ha trabajado con información proveniente de estudios proteómicos, con los objetivos de desarrollar herramientas que faciliten el análisis global de mapas y redes de interacción de proteínas y de diseñar estrategias que mejoren la calidad de los datos de interacción. Los dos últimos capítulos, en cambio, han usado datos e información generada por estudios transcriptómicos. El capítulo tercero describe el desarrollo de un método robusto de búsqueda de perfiles de expresión correlacionados, que ha permitido construir redes de coexpresión génica que son analizadas en su estructura y en el tipo de funciones biológicas que muestran. Por último, el capítulo cuarto estudia la desregulación de la expresión genómica producida en estados funcionales alterados (como estados de enfermedad), por medio del diseño de un algoritmo de búsqueda de grupos de genes con el perfil de expresión altamente cambiante y desregulado.

OBJETIVOS

Respecto a **redes de interacción de proteína**s:

1º.- Diseñar e implantar una **base de datos** de **interacciones proteína-proteína** (PPI), que integre y unifique la información sobre las interacciones demostradas experimentalmente procedente de diversos repositorios públicos internacionales.

2º.- Diseñar y desarrollar una **plataforma web** que permita acceder fácilmente a la información contenida y organizada en la base de datos de interacciones entre proteínas.

3º.- Implementar herramientas dinámicas de **visualización de redes** de interacción de proteínas, que faciliten la extracción de conocimiento de las redes derivadas de nuestra base de datos unificada.

4º.- Estudiar estrategias y **parámetros de validación** de interacciones proteína-proteína, que midan la fiabilidad de las interacciones y consigan filtrar falsos positivos para obtener redes de interacción más ciertas.

Respecto a **redes de coexpresión de genes**:

5º.- Desarrollar y aplicar un método estadístico-computacional robusto a datos de expresión genómica derivados de *microarrays* de oligonucleótidos de alta densidad, que sea capaz de encontrar **perfiles de expresión correlacionados** validados, de los que se pueda inferir coexpresión génica estable.

6º.- Aplicar el método de búsqueda de coexpresión desarrollado para inferir relaciones binarias entre genes y crear una **red de coexpresión** de **genes humanos** fiable, con la que se podrá asignar funciones a genes no anotados y encontrar factores de transcripción comunes a módulos muy conectados de la red.

7º.- Programar un algoritmo aplicable a datos de expresión genómica derivados de *microarrays* de oligonucleótidos de alta densidad, que encuentre grupos de genes con **perfiles de expresión con alta variabilidad** respecto a los controles; que serán propios de situaciones alteradas frecuentes en estados patológicos.

Capítulo 1

Mapas ómicos de interacción de proteínas

1.1 INTRODUCCIÓN

El conocimiento de los procesos biológicos que se producen en la célula, está ligado a estudiar cómo las proteínas u otras biomoléculas se asocian en determinadas condiciones para realizarlos. Esta claro que las proteínas por si solas no pueden realizar las funciones de la célula, necesitan formar máquinas moleculares para hacer procesos biológicos. De hecho, el análisis de datos de interacción está mostrando cómo el sistema global que forman las biomoléculas y las interacciones, es más grande y más complejo que la suma de sus partes (Aloy y Russell, 2006). Tanto la genómica como la proteómica están permitiendo identificar listas globales de biomoléculas presentes en las células, que participan de forma coordinada y conjunta en los procesos celulares. También se está definiendo una gran cantidad de relaciones moleculares

gracias a las nuevas tecnologías experimentales, que pueden detectar miles de interacciones en un solo ensayo, y están ayudando enormemente a ampliar el conocimiento que se tiene sobre las redes de interacción biomolecular dentro de los sistemas celulares biológicos.

Entre todos los tipos de redes biológicas, los mapas de interacción proteína-proteína son unos de los conjuntos de datos más grandes y diversos que hay definidos hasta la fecha. Los primeros mapas globales de interacción entre proteínas fueron generados usando la técnica de dos híbridos (Ito, *et al.*, 2001; Uetz, *et al.*, 2000) aplicada a levadura (*Saccharomyces cerevisiae*). Posteriormente se ha seguido aplicando esta técnica en otros organismos y se han desarrollado nuevos métodos de detección de interacciones de alto rendimiento (*high throughput*). Sin embargo, los datos obtenidos por estas técnicas se tienen que usar con precaución, ya que han sido varias las publicaciones que han advertido sobre la gran cantidad de falsos positivos que generan estos ensayos (Bader y Hogue, 2002; von Mering, *et al.*, 2002). Es importante tratar el problema de los falsos positivos, ya que el empleo de datos ruidosos puede llevar a conclusiones erróneas. Por esto numerosas investigaciones están desarrollando estrategias computacionales (Bader, 2003; Deane, *et al.*, 2002; Iossifov, *et al.*, 2004) y experimentales (Yu, *et al.*, 2008) para aumentar la fiabilidad de los datos de interacción y se está trabajando en generar parámetros que midan y calibren la veracidad de las interacciones.

El acúmulo de interacciones que se ha producido por los métodos experimentales de alto rendimiento, hace que cada vez sea más importante organizar las interacciones en bases de datos, para que esta información pueda ser recuperada y analizada con facilidad. Las bases de datos de interacción de proteínas se han convertido en uno de los recursos más valorados para la consulta y el análisis sistemático de redes moleculares, que están siendo usadas tanto en investigaciones bioinformáticas como biológicas.

El descubrimiento e identificación de todas las máquinas moleculares que realizan los procesos biológicos celulares, mediante la deducción experimental y los análisis computacionales de redes de biomoléculas interactuantes, es uno de los mayores retos de la era "ómica", que necesita de herramientas y de avances bioinformáticos para abordar la complejidad de los sistemas biológicos estudiados de modo global.

1.1.1 Métodos experimentales de detección de interacciones proteína-proteína

En los últimos años, con el nacimiento de la proteómica, se han desarrollado técnicas experimentales capaces de inferir masivamente interacciones entre proteínas. Estas técnicas han sido ampliamente usadas para descubrir nuevas interacciones y han ayudado de forma significativa a aumentar el conocimiento sobre los interactomas de especies determinadas. Los métodos experimentales de alto rendimiento (*high-throughput*) más relevantes, clasificados por el tipo de interacción que buscan, son de dos tipos principales:

(**i**) Orientados a la búsqueda de complejos de proteínas:

- Co-Inmunoprecipitacion (**Co-IP**)
- Purificación por inmovilización y arrastre (***PullDown***)
- Purificación por afinidad en tándem (**TAP**)

(**ii**) Orientados a la búsqueda de parejas de proteínas interactuantes (interacciones binarias):

- Matrices de proteínas
- Sistemas de dos híbridos (**2H**)

Los métodos de Co-Inmunoprecipitación (**Co-IP**) son muy populares en el análisis de interacciones de proteínas. Un ejemplo de estos métodos esta representado en la **figura 1.1** y se puede resumir en 3 pasos: **1**) Se lisa la célula y se añade un anticuerpo

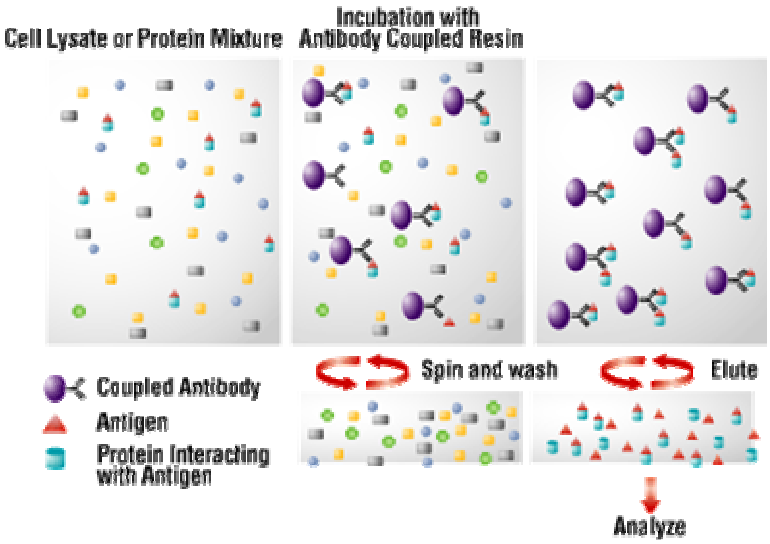


Figura 1.1: Co-Inmunoprecipitación.
Resumen esquemático del proceso normalmente usado de Co-IP (fuente http://thunder.biosci.umbc .edu) [3].

específico de la proteína de interés. **2**) Se precipita por centrifugación el complejo formado por la proteína y el anticuerpo, de forma que, si hay alguna molécula que se ha unido a la proteína de interés, ésta también precipitará. **3**) Se identifica la proteína precipitada por medio de un *western blot* o secuenciando la banda de la proteína purificada [4].

Los ensayos de purificación por inmovilización y arrastre (***PullDown***) son métodos *in vitro* usados para definir interacciones físicas entre dos o más proteínas. Un ejemplo está representado en la **figura 1.2**. Hay muchas variantes de este tipo de ensayo pero en general es necesario tener una proteína (cebo, *bait*) purificada y etiquetada, la cual será inmovilizada para capturar posibles interactores (presas, *prey*). Finalmente el complejo que se ha formado será analizado para identificar las proteínas que interaccionan con la proteína cebo [5].

Los métodos de purificación por afinidad en tándem (**TAP**) (**figura 1.3**) están basados en una etiqueta que es genéticamente añadida a la proteína cebo (*bait*) de la que queremos conocer sus interactores. La etiqueta más usada consiste en un péptido de unión a calmodulina (*CBP*), una secuencia de corte para la proteasa *TEV* y una proteína A. Este etiquetado permitirá aislar y purificar complejos formados por la

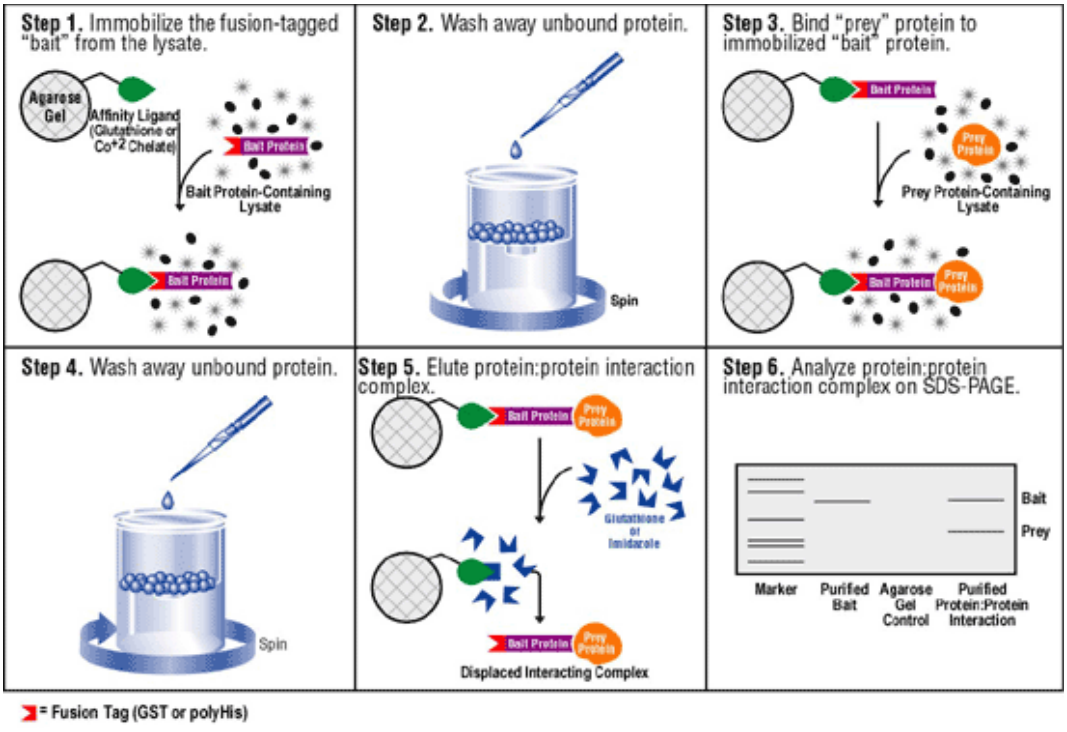


Figura 1.2: Purificación por inmovilización y arrastre
Ejemplo de una aproximación PullDown usando la proteína de fusión GST (fuente http://www.piercenet.com/) [5].

proteína y otros interactores. El complejo se aislará por cromatografía usando una matriz de afinidad *IgG*, ya que ésta se une fuertemente a la proteína A (que es parte de la etiqueta de la proteína cebo). El complejo se puede liberar de la matriz usando proteasa *TEV*, que rompe su secuencia específica de corte, que había sido introducida previamente en la etiqueta. Para deshacerse después de la proteasa, se incuba la proteína con esferas recubiertas de calmodulina que se unen al *CBP* de la proteína etiquetada. Finalmente se procede a la identificación de las proteínas que forman el complejo por medio de espectometría de masas u otra técnica de identificación (Puig, *et al.*, 2001).

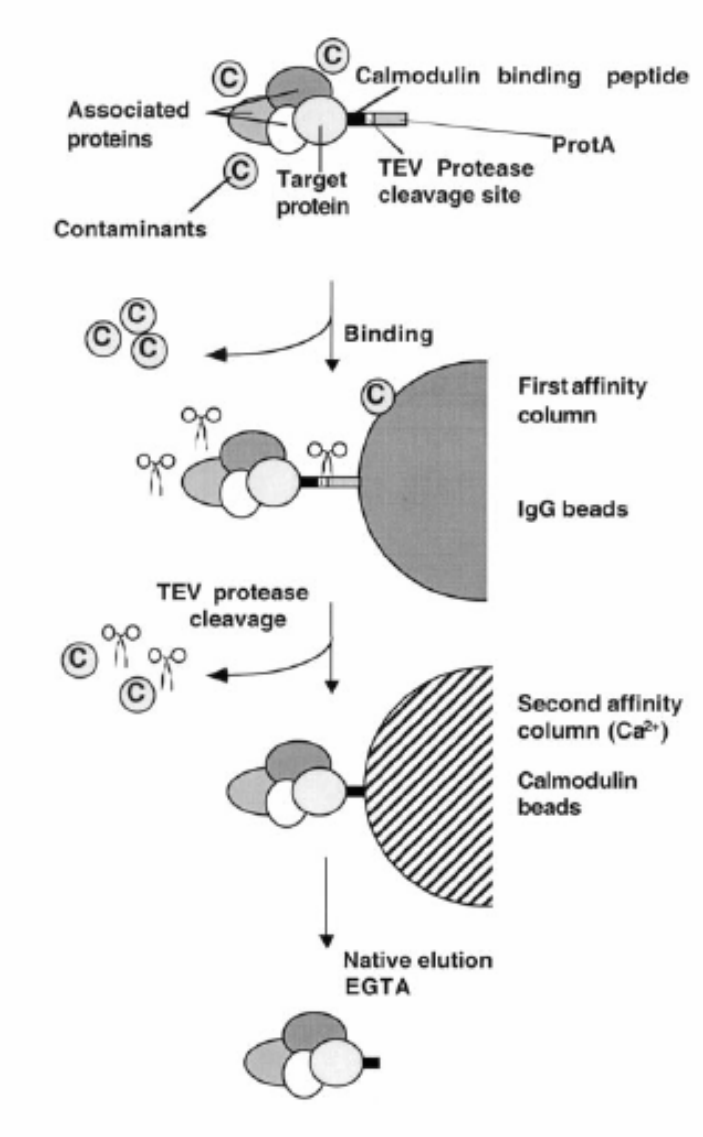


Figura 1.3: Purificación por afinidad en tándem.
Visión general del método biomolecular de *Tandem Affynity Purification* (*TAP*) (fuente Puig *et. al.*) (Puig, *et al.*, 2001).

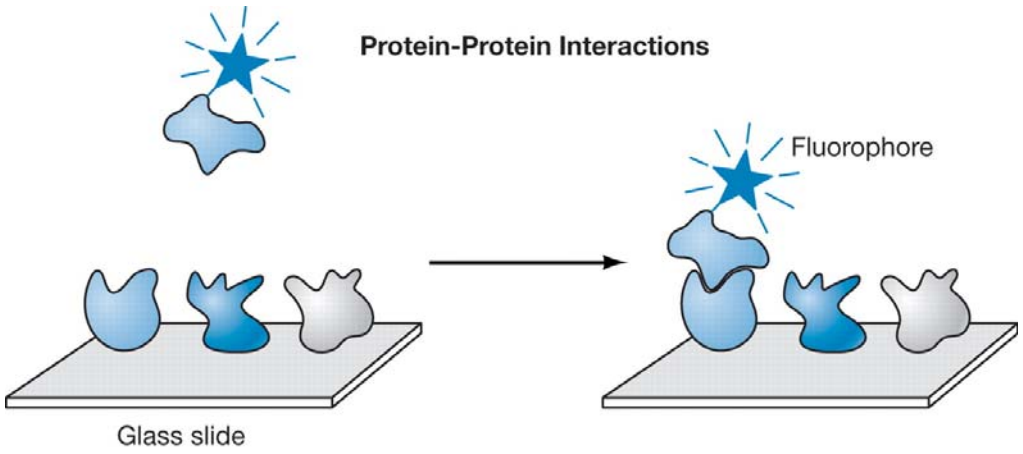


Figura 1.4: Matriz de proteínas.
Esquema sencillo de una aproximación de matriz de proteínas para detectar interacciones proteína-proteína (fuente http://cshprotocols.cshlp.org) [6].

Una aproximación más directa de búsqueda de interacción proteica son las matrices de proteínas. Estos dispositivos suelen ser superficies de cristal a los que se han fijado proteínas en localizaciones controladas, formando una matriz microscópica. Para construirlo se elige un conjunto de proteínas que es sobreexpresado, purificado y distribuido en la matriz fijándolas por distintas estrategias químicas. Posteriormente, esta matriz se incuba con proteínas marcadas con fluorescencia y, tras un lavado, las interacciones estables son identificadas escaneando el cristal en busca de puntos fluorescentes (ver **figura 1.4**). Los diseños específicos para la construcción de las matrices de proteínas son muy variados y están actualmente en evolución [6].

Durante los últimos años el sistema de dos híbridos (**2H**) ha sido muy útil para detectar a gran escala interacciones entre proteínas; ha posibilitado el análisis de proteomas completos y ha colaborado de forma significativa al conocimiento de las complejas redes de interacción de proteínas en diversos organismos. El sistema de dos híbridos está basado en la naturaleza modular de ciertos activadores transcripcionales en organismos eucariotas. Estos activadores tienen dominios separables que deben estar juntos para producir actividad transcripcional, véase por ejemplo *GAL4* (Keegan, *et al.*, 1986) y *GCN4* (Hope y Struhl, 1986). Los dominios separables son de dos tipos:

- el dominio de unión a *DNA* (***BD***), encargado de colocar el factor de transcripción en una secuencia específica de *DNA*, presente en la región promotora que precede a los genes que están regulados por ese factor de transcripción.
- el dominio de activación de la transcripción (***AD***), que establece contactos con componentes de la maquinaria de transcripción, permitiendo el inicio de la transcripción.

Como se ve en la **figura 1.5a** para los experimentos de dos híbridos se usan dos plásmidos de levadura, uno que codifica el dominio *DB* fusionado con una proteína cebo (*bait*) X, y otro que codifica el dominio *AD* fusionado con la proteína presa (*prey*) Y. Si las proteínas X e Y interaccionan, el complejo formado debe ser capaz de unirse a la región promotora de un gen delator, que contiene el sitio de reconocimiento correspondiente al *DB*, permitiendo al dominio *AD* activar la transcripción de dicho gen. En caso contrario el gen delator permanecerá silenciado (Shoemaker y Panchenko, 2007).

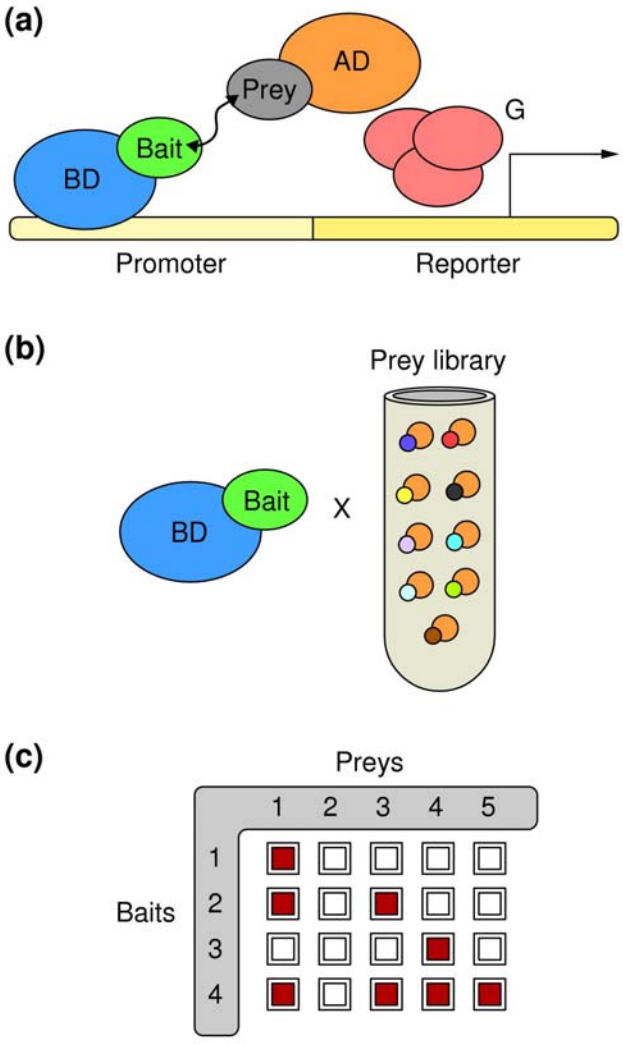


Figura 1.5: Dos híbridos.
Representación de una librería y una matriz de una aproximación dos híbridos (**2H**). **a**) Modelo del sistema dos híbridos. La proteína cebo (*bait*) se une al dominio de unión a *DNA BD* y la proteína presa (*prey*) se une al dominio de activación de transcripción *AD*. Cuando la proteína cebo y la presa (*prey*) interactúen, se unirán los dominios *AD* y *BD* y se inducirá la transcripción del gen delator. **b**) Librería de un sistema de alto rendimiento dos híbridos. Se forma con una colección de presas para una proteína cebo de interés. **c**) Matriz creada para hacer un ensayo dos híbridos a gran escala. Las interacciones se deducen por la expresión del gen delator, que está marcado en rojo en el dibujo. (fuente Giorgini F. y Muchowski PJ.) (Giorgini y Muchowski, 2005).

1.1.2 Bases de datos de interacción de proteínas

La información es un conjunto organizado de datos procesados, que constituyen un mensaje sobre un determinado ente o fenómeno. Es necesario que la información esté organizada para que pueda ser manejada y compartida. Los datos de interacción biomolecular no son distintos del resto de datos. La fuente de estos datos son las publicaciones científicas, que deben ser recopilados y almacenados de forma organizada para formar información, que pueda ser compartida y manejada para construir conocimiento. Con estos propósitos surgieron las bases de datos de interacción, que están recopilando y organizando información de interacciones biomoleculares y han construido plataformas Web para facilitar el acceso a dicha información.

Las bases de datos de interacción se pueden clasificar en 3 tipos en función de su contenido: (**i**) las que contienen interacciones comprobadas experimentalmente; (**ii**) las que contienen interacciones deducidas por métodos de predicción *in silico*; y (**iii**) las que almacenan los dos tipos de interacciones. Las bases de datos de interacciones con evidencia experimental tienen dos mecanismos para conseguir la información. Puede ser recopilada por personas que leen artículos científicos en busca de interacciones o puede ser facilitada por los investigadores que han realizado los ensayos que las validan experimentalmente. La mayoría de las bases de datos tienen personal (*curators*) que busca interacciones en artículos científicos y las anotan manualmente, aunque ahora se esta intentando que sean los autores de los artículos los que faciliten esta información a las bases de datos (Ceol, *et al.*, 2008). Para ello se ha descrito la información mínima que debe incluir un experimento que reporta una interacción biomolecular en un estándar llamado *MIMIx* (Orchard, *et al.*, 2007); de modo que los autores, cuando reporten que en un artículo han sido validadas un conjunto de interacciones, deberán facilitar una serie de datos mínimos que describen el ensayo realizado (Orchard, *et al.*, 2007). Además, el consorcio *IMEx* [7] de bases de datos está empezando a colaborar con revistas científicas para que faciliten los datos de interacción que publican.

A continuación se describen las principales bases de datos que han recopilado interacciones descritas en bibliografía, y que han conseguido almacenar el mayor número interacciones comprobadas experimentalmente:

- **BIND** (Alfarano, *et al.*, 2005) [8]: Es una base de datos de asociaciones biomoleculares que están clasificadas en 3 categorías: interacciones moleculares binarias, complejos moleculares y rutas moleculares. La mayoría de las interacciones que tiene son entre proteínas pero también almacena interacciones con ácidos nucleicos y moléculas pequeñas. La función de las proteínas se resalta usando *ontoglyphs*, que es un conjunto de símbolos que representan términos *GO* (Ashburner, *et al.*, 2000). La información estructural y la relacionada con propiedades de unión de proteínas, se representa también con símbolos llamados *proteoglyphs*. Además proporciona una validación de las interacciones basada en *textmining* de publicaciones, en la existencia de interacciones homólogas, en anotaciones funcionales *GO* relacionadas, en la posible interacción de dominios y en el perfil fenotípico. Para el acceso automatizado a los datos cuenta con un *interface SOAP*. Esta base de datos, que empezó en el año 2002, actualmente ya no está mantenida, ya que fue abandonada como proyecto científico en 2006.
- **BioGRID** (Breitkreutz, *et al.*, 2008) [9]: Es una base de datos que contiene tanto interacciones proteicas como lo que denominan "interacciones" genéticas. Esta relacionada con la base de datos *SGD* (Cherry, *et al.*, 1998), y por ello tiene mucha información sobre interacciones de *Saccharomyces Cerevisiae*. Como visor de interacciones usa *Osprey* (Breitkreutz, *et al.*, 2003), que es una herramienta Java de visualización y análisis de redes.
- **DIP** (Salwinski, *et al.*, 2004) [10]: Además de tener interacciones directas y complejos, almacena también interacciones funcionales. Esta conectada con *Cytoscape* (Shannon, *et al.*, 2003) de modo que desde esta plataforma se pueden visualizar las redes de interacción de *DIP* a través de un vínculo en su página Web. Como servicios adicionales da la posibilidad de evaluar conjuntos de interacciones experimentales o predichas por el usuario por medio de métodos de verificación basados en parálogos (*PVM*), en perfiles de expresión (*EPR*) o en dominios interactuantes (*DPV*). Además tiene varios proyectos relacionados, como es *LiveDIP* (Xenarios, *et al.*, 2002) que almacena interacciones entre proteínas describiendo el estado fisiológico y las transiciones de estado de las proteínas. Otro proyecto es *DLRP* (Graeber y Eisenberg, 2001) que es una base de datos de receptores y ligandos que transducen señales. Por último, entre los proyectos satélite de *DIP* está *ProLinks* (Bowers, *et al.*, 2004), que es una plataforma de predicción *in silico* de interacciones.

- **HPRD** (Keshava Prasad, *et al.*, 2009) [11]: Es una base de datos centrada en proteínas humanas. Las proteínas están clasificadas por categorías en función del tipo, los dominios, los motivos y la localización que tengan. Además tiene una amplia anotación del proteoma humano. Sin embargo, las interacciones están pobremente anotadas, sobre todo en la descripción del tipo de experimento usado para validarlas, ya que solo diferencia si el experimento ha sido realizado *in vivo*, *in vitro* o por una técnica dos híbridos. Tiene varios proyectos que están relacionados con ella, como son la *Proteinpedia* (Kandasamy, *et al.*, 2009) que permite a los investigadores añadir información a cerca de proteínas, *PhosphoMotif Finder* [12] que tiene sustratos kinasa/fosfatasa así como motivos de unión encontrados en bibliografía y *NetPath* [13] que es una base de datos de rutas moleculares de transducción de señales.
- **IntAct** (Kerrien, *et al.*, 2007) [14]: Esta base de datos incluye una descripción detallada de cómo y en qué condiciones se produce la interacción, de los dominios que interactúan, del método experimental que valida dicha interacción y de la publicación que la describe. Su información está también presente en *UniProt*, donde se han añadido para cada entrada las interacciones almacenadas en *IntAct*. Su plataforma Web es de código abierto, así que cualquiera puede descargar el código fuente, colaborar en el proyecto e incluso instalarse la plataforma para mostrar las interacciones propias. Se puede acceder a la base de datos por medio de búsquedas simples o avanzadas de forma ágil y sencilla, de modo que en un solo clic se presenta una descripción básica de las interacciones de la proteína buscada. Además *IntAct* esta desarrollando herramientas basadas en su base de datos con propósitos diversos, por ejemplo se puede hacer una predicción de dianas para experimentos *pull-down*. También tiene herramientas para visualizar la red de interacción, para explorar información funcional sobre la red y para buscar caminos mínimos en la red de proteínas.
- **MINT** (Chatr-aryamontri, *et al.*, 2007) [15]: Ofrece una interfaz de fácil acceso para consultar la base de datos de interacción. La calidad de las interacciones está cuantificada con un parámetro de calidad (*score*), basado en el número de artículos científicos que validan la interacción y en el tipo de los métodos experimentales que se han empleado (según sean de alto rendimiento o detecten interacciones no directas entre proteínas). Tiene un visor de redes que permite filtrar las interacciones en función de su *score*, así como exportar los datos en un fichero

tabulado o en formato *XML*. Un proyecto derivado de esta base de datos es *HomoMINT* (Persico, *et al.*, 2005) que incluye solamente interacciones entre proteínas humanas. Estas interacciones pueden haberse derivado de literatura o pueden haber sido inferidas por proteínas no humanas que interaccionan y tienen ortólogos en humano. Además tiene asociada una base de datos de interacción de proteínas víricas y otra de interacciones entre dominios peptídicos.

1.1.3 Método de construcción de bases de datos de interacción

Para conseguir que una base de datos sea capaz de almacenar y organizar la información de forma eficiente, es necesario realizar un diseño previo. El diseño de base de datos se puede dividir en tres partes: (**i**) diseño **conceptual**, (**ii**) diseño **lógico** y (**iii**) diseño **físico**. El diseño conceptual consiste en construir un esquema que sea capaz de almacenar toda la información necesaria para definir el ente o fenómeno que se quiere representar en base de datos. El esquema conceptual se transforma en el diseño lógico a un modelo de base de datos (modelo relacional, jerárquico, orientado a objetos o en red) que cumpla todas las especificaciones del esquema conceptual. Por último, en el diseño físico, se especifica la implementación concreta de la base de datos, definiendo las estructuras de almacenamiento y los métodos de acceso que garanticen un acceso eficiente a los datos (Teorey, 1999).

El diseño conceptual de las bases de datos de interacción, buscará un esquema que contenga la información necesaria para definir una interacción biomolecular. Cada base de datos ha seguido un esquema conceptual diferente, sin embargo, como se comentó en el apartado anterior, hay una iniciativa de la que somos partícipes el grupo de bioinformática del CIC, que especifica la información mínima para definir experimentos que reportan interacciones moleculares llamada *MIMIx*. Esta iniciativa, además de posibilitar que los autores de los artículos envíen sus propias interacciones, es un modelo de datos estándar para la representación e intercambio de datos de interacción de proteínas. Los datos se dividen en 3 entidades: (**i**) las moléculas interactuantes, (**ii**) el experimento realizado y (**iii**) la interacción reportada. Sobre las moléculas interactuantes se hace especial hincapié en que se describan con un "identificativo no ambiguo" de alguna base de datos pública, como por ejemplo para proteínas *UniProt* (Bairoch, *et al.*, 2005) o *RefSeq* (Pruitt, *et al.*, 2007), para genes *Ensembl* (Hubbard, *et al.*, 2009) o *Entrez Gene* (Maglott, *et al.*, 2007) o para entidades

químicas *PubChem* [16] o *ChEBI* (Degtyarenko, *et al.*, 2008). Además se debe definir "la forma canónica" de la molécula (si está en la forma natural de la célula o ha sido alterada), así como el "rol biológico" (p.e. enzima o diana de la enzima) y el "rol experimental" (p.e. cebo o presa) de la molécula en la interacción. Para describir el experimento se tiene que especificar el método de detección de la interacción y el organismo en el que se ha realizado el experimento (en caso de no haberse hecho *in vitro*). Además, si el experimento lo requiere, se debe especificar el método con el que se han detectado los participantes. Por último, es necesario describir la interacción especificando las moléculas que han participado en ella y el valor de confianza que tienen en caso de haber hecho una evaluación de la calidad de la interacción.

Las bases de datos han seguido esquemas conceptuales muy heterogéneos, de modo que cada base de datos ha considerado informaciones diferentes para definir interacciones entre proteínas. Sin embargo, en la mayoría de los esquemas está presente la información mínima que especifica el estándar *MIMIx*.

A la hora de transformar el esquema conceptual a un esquema lógico, las bases de datos de interacción han elegido un modelo relacional. Un ejemplo de un esquema lógico de interacción entre proteínas está representado en la **figura 1.6** mediante un

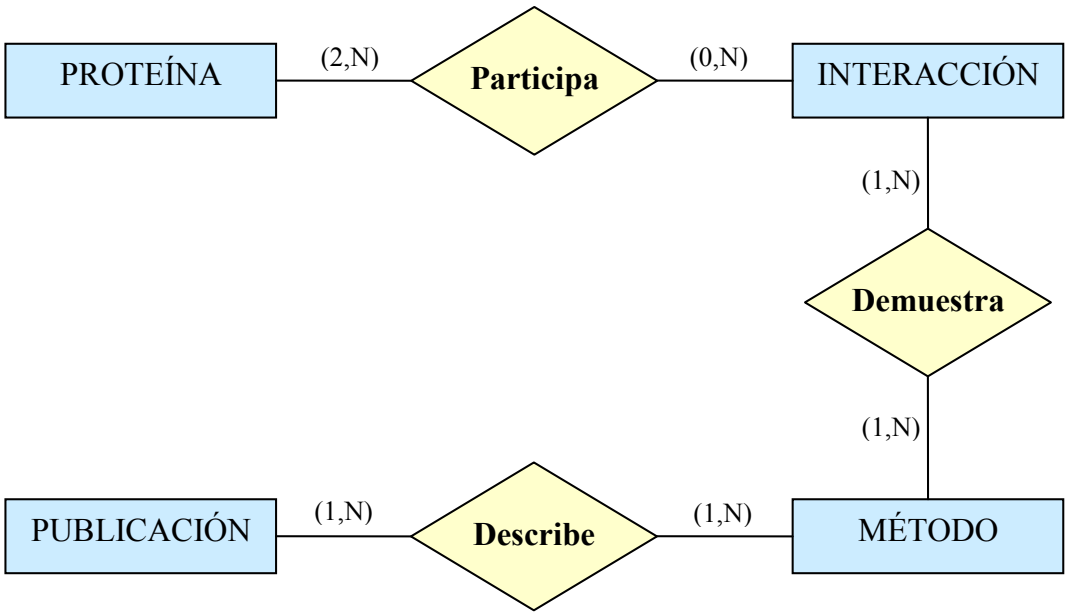


Figura 1.6: Diagrama entidad-relación de la base de datos de APID.
Ejemplo de un diseño físico para representar interacciones de proteínas.

diagrama entidad relación. Finalmente se hace el diseño físico antes de implementar la base de datos, en el que se concretará la construcción de la base de datos con el objetivo de optimizar el tiempo de acceso a base de datos y minimizar el tamaño de almacenamiento.

1.1.4 Intercambio de datos de interacciones moleculares.

La información además de estar organizada en base de datos, es necesario que se pueda compartir. Ya se ha hablado de cómo los autores de artículos y las revistas científicas pueden compartir las interacciones con las bases de datos a través de *MIMIx*; pero también es importante, que las bases de datos de interacciones compartan sus datos con el mundo científico, para permitir su análisis y la deducción de nuevo conocimiento. El formato de intercambio de datos más extendido entre las bases de datos de interacción es el lenguaje de marcas ampliable, más conocido como *XML* (*Extensible Markup Language*), que se ha convertido en un estándar para el intercambio de información estructurada. La información se estructura en partes bien definidas que se componen a su vez por otras partes, esto es, *XML* estructura la información en un árbol jerárquico con piezas de información en cada uno de sus nodos. Inicialmente, para exportar los datos de interacción, las bases de datos tenían su propio formato *XML*, de modo que la información estaba estructurada de diferente forma en el fichero exportado, pero actualmente se está trabajado para lograr un formato común a todas estas bases de datos. Ésta es una de las razones por las que en 2002 se creó una iniciativa de estándares en proteómica llamada *Proteomics Standards Initiative* (*PSI-MI*) [17]; que tiene como objetivo definir un conjunto de estándares para la representación de datos en proteómica, que facilite la comparación, el intercambio y la verificación de los datos. Entre los grupos de trabajo que están definiendo los estándares hay uno relativo a interacciones moleculares (*MI*) que tiene dos objetivos principales: **1**) mejorar la anotación y la representación de los datos de interacción biomolecular, **2**) mejorar la accesibilidad de las interacciones moleculares a la comunidad de usuarios, definiendo estándares comunes de datos para que estos puedan ser fácilmente compartidos y combinados.

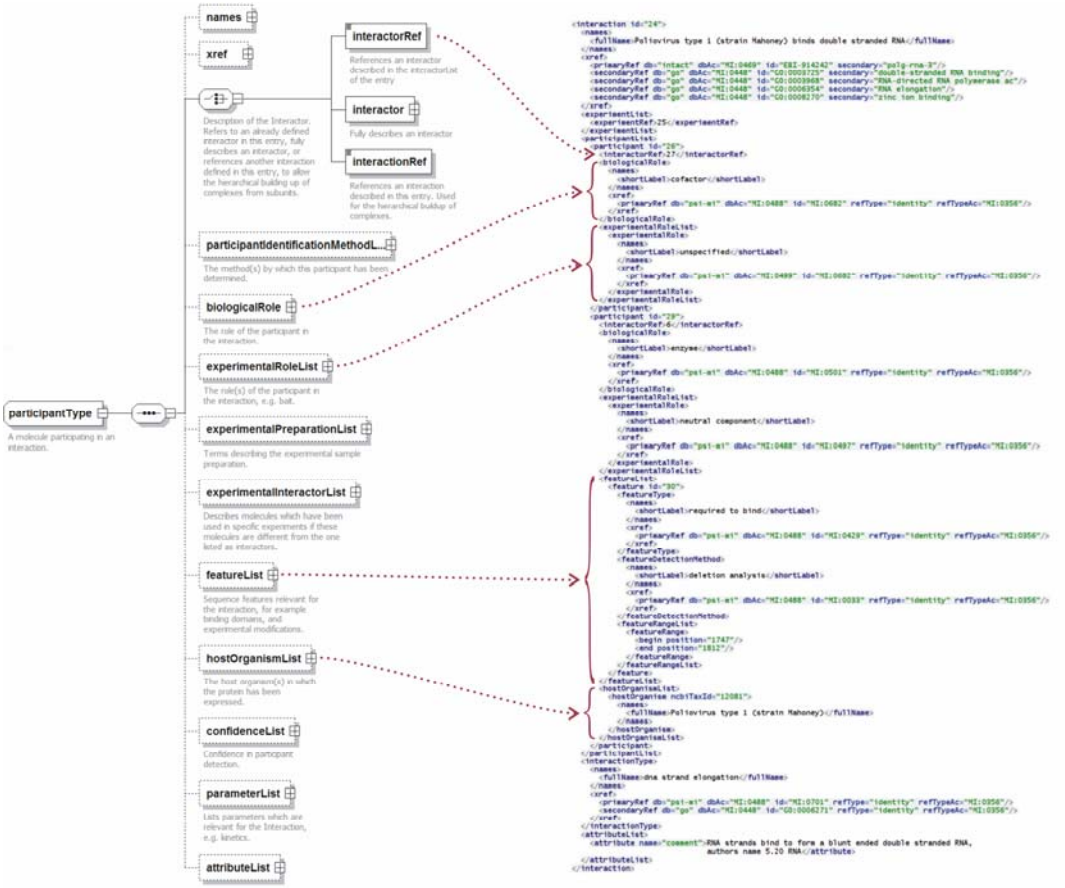


Figura 1.7: PSI-MI XML. Representación gráfica de la sección del esquema
PSI-MI XML 2.5 que describe a un participante, con el código XML correspondiente (fuente Kerrien *et al.*) (Kerrien, *et al.*, 2007).

Para conseguir el segundo objetivo se ha definido el formato *XML PSI-MI* (Hermjakob, *et al.*, 2004), que especifica un estándar de intercambio de datos de interacciones moleculares. En la **figura 1.7** se describe la estructura debe cumplir un fichero *XML* para seguir la especificación propuesta por *PSI-MI*. Este formato ha sido adoptado por la mayoría de las bases de datos de interacción. Además también se está usando como fichero de entrada para visualizadores de redes y para hacer públicos los datos de interacción de proteínas.

Un formato de intercambio de datos por si solo no garantiza su compatibilidad. También es necesario que los atributos de los datos estén bien definidos por medio de documentación y vocabularios controlados. El formato PSI-MI está documentado detalladamente en la página Web de PSI [18], donde se describe cada atributo y la información que debe contener. Para estandarizar los contenidos de los atributos se han creado los vocabularios controlados u ontologías. El uso de vocabularios

controlados hace que los valores de los atributos no sean texto libre, sino que cada posible atributo debe estar definido en el vocabulario. De este modo, se consigue que haya estandarización y univocidad en los términos empleados, siendo ésto muy importante para facilitar las búsquedas y garantizar la compatibilidad en el intercambio de datos. El vocabulario controlado se organiza en una estructura arborescente que tiene en los primeros niveles a los términos más generales del vocabulario, de los que descienden ramas con términos más específicos de significación cada vez más concreta. El primer nivel de PSI-MI divide el árbol en 5 vocabularios controlados: tipo de interacción, tipo de características de secuencia, detección de características, detección de participante y detección de interacción. Partiendo de las categorías descritas en el primer nivel, los niveles sucesivos definen términos cada vez más específicos que se usan para definir interacciones moleculares. Esta estructura jerárquica tiene ventajas tanto para anotar como para consultar los datos, ya que la anotación se puede hacer al nivel de detalle deseado, y la búsqueda puede encontrar los objetos anotados a un término determinado y a todos sus descendientes. Los vocabularios controlados se distribuyen en el mismo formato que *GO* y están unidos a *Open Biomedical Ontologies* (*OBO*) (Smith, *et al.*, 2007) usando el prefijo "MI" para los identificativos *OBO*.

Los esquemas y vocabularios definidos por *PSI* han conseguido que el formato de los datos (es decir, la sintaxis) y su significación (la semántica) sean coherentes y estén bien definidos. Con todo ello se facilita la búsqueda, la distribución, la compatibilidad y la integración de los datos de interacción.

1.2 MÉTODOS Y RESULTADOS

1.2.1 APID: Agile Protein Interaction DataAnalizer

En la introducción se ha visto que los datos de interacción de proteínas se almacenan en artículos científicos y hay bases de datos que se dedican a extraer e integrar esta información. Sin embargo, cuando iniciamos este trabajo comprobamos que cada base de datos tiene sus propios protocolos de búsqueda, extracción, anotación y almacenamiento de interacciones, que hacen que las bases de datos exploren distintos artículos científicos, anoten las interacciones de forma diferente y almacenen la información de distinto modo. De hecho, observamos que la información que tienen las bases de datos es muy heterogénea, tanto por su contenido como por su anotación, y hay poco solapamiento entre ellas. El poco solapamiento se representa en la **figura 1.8** mediante un diagrama de *Venn* que muestra la intersección que hay entre tres bases de datos de interacción (*BIND*, *DIP*, e *IntAct*), indicando el número de interacciones que tienen en común y el número de interacciones que son únicas de cada una de ellas. Es interesante resaltar que el 62% del total de las interacciones proteína-proteína incluidas en *BIND*, *DIP* e *InAct* esta solo presente en una de esas bases de datos, esto es, está solo incluida en uno de estos recursos de interacción.

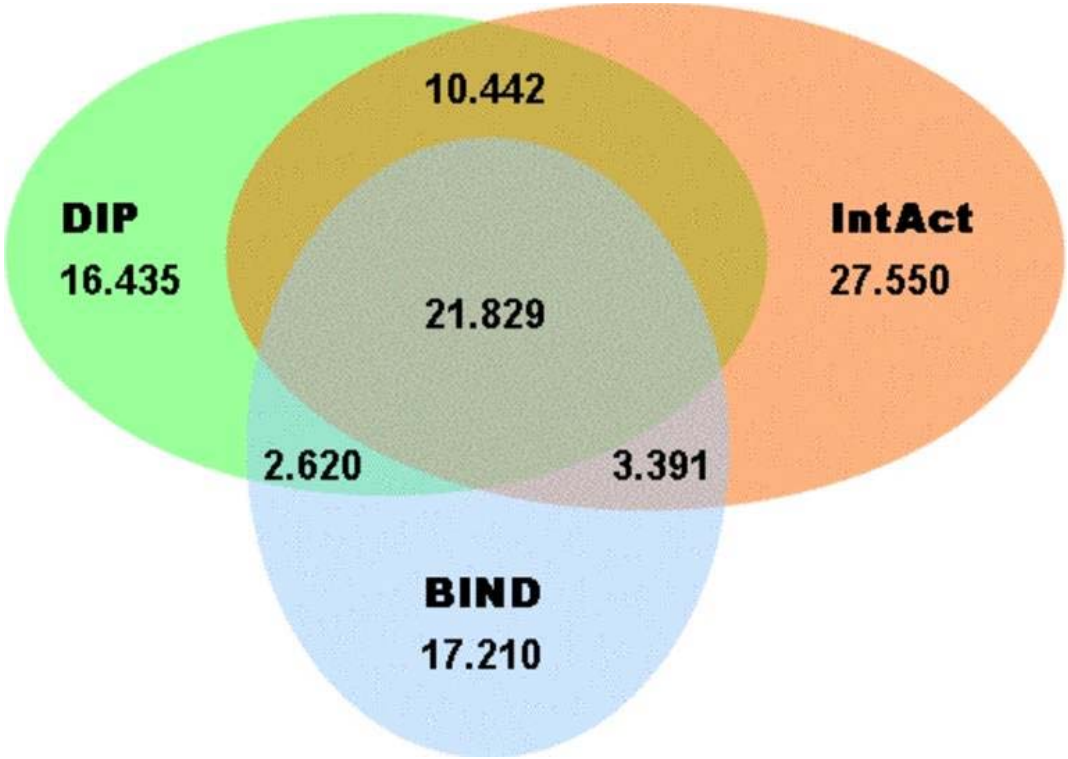


Figura 1.8: Diagrama de *Venn*.
Representación del solapamiento e intersección de las bases de datos *BIND*, *DIP* e IntAct.

Además, no hay que olvidar que las técnicas proteómicas masivas de las que provienen la mayoría de las interacciones proteína-proteína, son ruidosas y producen un elevado número de falsos positivos (Bader y Hogue, 2002; von Mering, *et al.*, 2002). Esto hace que las redes de interacción sean inexactas e incompletas, pese a tener anotadas un número de interacciones superior a estimaciones hechas por diferentes estudios (por ejemplo para levadura se estiman entre 12000 y 40000 interacciones (Uetz y Finley, 2005)). Por otro lado, tampoco existen conjuntos de referencia adecuados suficientemente validados.

En este marco, concluimos que la información de las bases de datos de interacción biomolecular es complementaria y puede ser unificada para incrementar y mejorar nuestro conocimiento sobre los interactomas. Además, la unificación puede permitir aumentar el número de evidencias experimentales que demuestran las interacciones y contrastar la información de distintas bases de datos. De este modo, abordamos una integración de los datos de interacción para mejorar la cobertura, la calidad y el conocimiento de las complejas redes de interacción biomolecular.

Con el objetivo inicial de hacer una integración global de los principales repositorios de interacción proteína-proteína y de definir indicadores de la fiabilidad de las interacciones, hemos desarrollado una herramienta bioinformática interactiva de acceso Web, que integra y unifica en una plataforma común los principales interactomas conocidos. A esta herramienta la hemos denominado *APID* (*Agile Protein Interaction DataAnalizer*) (Prieto y De Las Rivas, 2006).

1.2.2 Diseño e implementación de APID

APID ha sido diseñado con el objetivo de ser una plataforma tan simple como sea posible, que ofrezca la información necesaria para garantizar un acceso sencillo y riguroso a todos los conjuntos de datos que integra. El diseño sigue una metodología de ingeniería de software denominada *agile* (Cockburn, 2001), que favorece el desarrollo del software usando métodos ligeros y adaptables. De este modo, los métodos *agile* persiguen hacer un diseño evolutivo que pueda asimilar los cambios y permitan que éstos ocurran a lo largo de todo el ciclo de vida del producto. Los cambios deben ser controlados y fáciles de implementar, y los diseñadores deben tener una actitud receptiva a dichos cambios. *APID* ha seguido esta estrategia para hacer una integración dinámica de las bases de datos de interacción de proteínas, tratando de

adaptarse con facilidad a los cambios surgidos por la integración de nuevos recursos o por la especificación de nuevas funcionalidades. Esta metodología de ingeniería de software es muy adecuada para ser usada en proyectos de desarrollo de software científico, ya que el producto puede ir evolucionando de acuerdo a los avances científicos que se hagan en su ámbito.

Todo el trabajo de *APID* ha sido desarrollado en el lenguaje de programación *Java* [19] y se ha seguido una arquitectura *J2EE* [20] para construir la interfaz Web y el a*pplet* [21] de visualización de redes que se describirá más adelante. Para el manejo de los ficheros fuente, se han desarrollado programas *SAX* [22] y *DOM* [23] que extraen información de los ficheros *XML*, y programas *JDBC* [24] para insertar los datos procesados en un servidor de base de datos relacional *MySQL* [25].

Uno de los principales obstáculos que hay que salvar al unificar los datos, es el uso de múltiples y diversos tipos de "identificadores" (*IDs*) de proteínas en las distintas bases de datos de interacción, que muchas veces causan incoherencia y disyunción en los datos. Para solventarlo usamos las secuencias de aminoácidos de las proteínas como código único y de mayor significación biológica. Si la secuencia es proporcionada por la base de datos de interacción, ésta nos permite usar el algoritmo *BLAST2* (Altschul, *et al.*, 1997) contra las secuencias de *UniProt* para buscar homologías. Cuando la secuencia tiene una coincidencia perfecta con alguna de *UniProt*, la proteína es reconocida y anotada a un código *UniProt* único. Si las secuencias no son facilitadas, se usan referencias cruzadas de base de datos para conseguir el código *UniProt*. Una vez que los códigos de proteína tienen un formato coherente y uniforme, se tiene que unificar la información referente a las interacciones. Esta unificación se hizo en base a tres identificativos: (**i**) de proteína, (**ii**) de publicación y (**iii**) de método experimental. La identificación del método o métodos experimentales que validan la interacción se hizo siguiendo el *OBO* de *PSI-MI* y buscando un consenso o acuerdo entre las distintas bases de datos que han anotado el artículo donde se describe. De esta forma, se construyó un protocolo capaz de almacenar y unificar bases de datos de interacción en una estructura uniforme, manteniendo la integridad de los datos y corrigiendo posibles fallos presentes en los ficheros fuente. Este protocolo es representado de modo esquemático en la **figura 1.9**.

Siguiendo esta estrategia, la unificación de los datos se hizo basándose en 3 identificativos concretos (*Ids*): (**i**) *UniProt ID*, que permitió una identificación

específica de cada proteína y un vínculo directo a información adicional a cerca de la proteína. (**ii**) *PubMed ID* (*PMID*) [26], para relacionar las interacciones con el artículo que las describe a través de una referencia *PubMed*, así como para relacionar los métodos experimentales de detección de interacción con el artículo en el que se describen. (**iii**) *PSI-MI ID*, para unificar los métodos experimentales usados en diferentes publicaciones a una terminología común desarrollada por *PSI-MI*. Estos indentificativos constituyen el núcleo de información de *APID* y consiguen que sea una herramienta de ágil de acceso y búsqueda de interacciones.

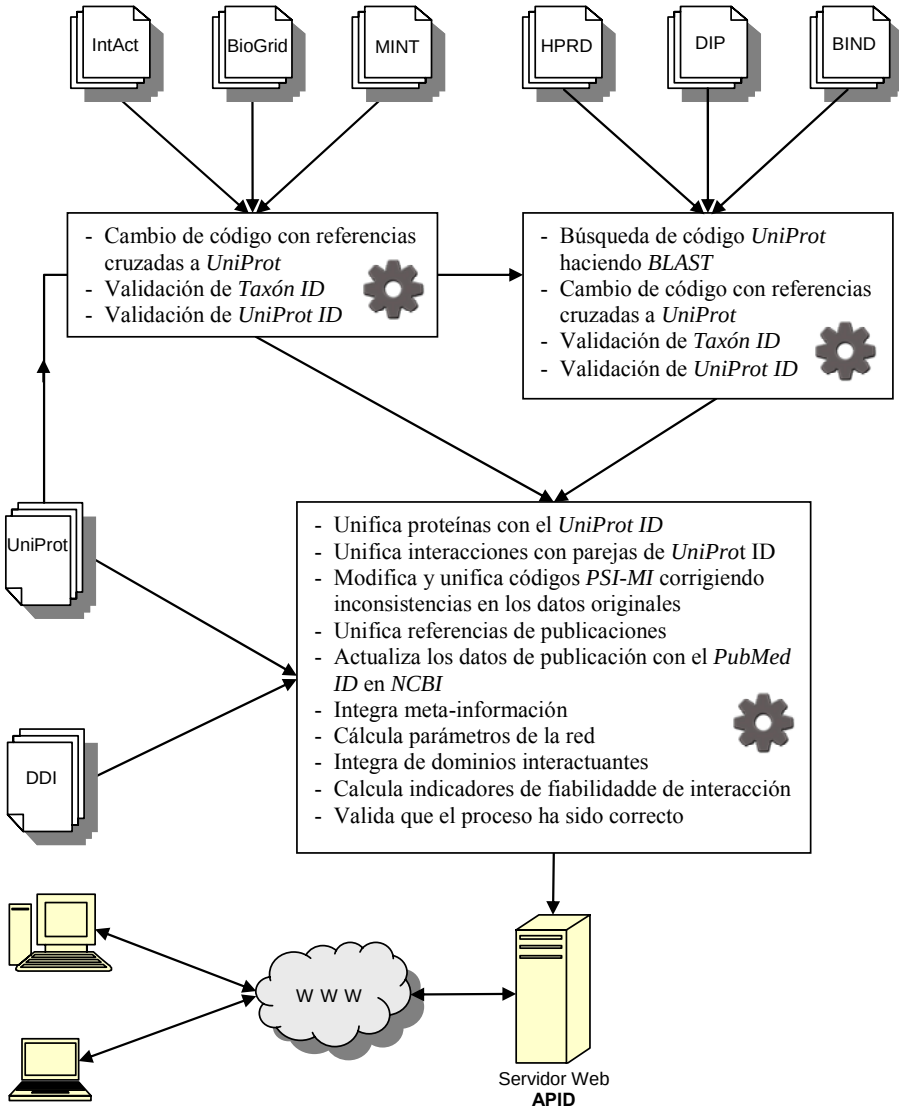


Figura 1.9: Diagrama de construcción de *APID*.
Muestra el flujo de información y los procesos que se realizan para construir la base de datos de *APID*.

1.2.3 Diseño de la base de datos dentro de *APID*

Como se explicó en la introducción, el diseño de la base de datos comienza definiendo un esquema conceptual con la información que debe contener la base de datos. El objetivo de APID es unificar la información proveniente de otras bases de datos de interacción, es por esto que el esquema conceptual está influenciado por lo que se va a importar. La información que cada base de datos recoge sobre las interacciones moleculares es muy diversa y no sigue, hasta la fecha, ningún estándar que especifique un esquema conceptual único para todas. Sin embargo, sí hay un conjunto de datos que son necesarios para poder definir una interacción unívocamente y que están presentes en todas las bases de datos. Estos datos son capaces de definir a todos los actores de una interacción y con ellos se podrá acceder a las bases de datos originales para consultar información adicional de la interacción. Para describir una interacción se identificaron 4 entidades principales que forman el esquema conceptual de *APID*: (**i**) proteína, (**ii**) método experimental, (**iii**) publicación, (**iv**) interacción. En base a estas entidades se definen atributos y relaciones para poder representar las interacciones proteicas.

A partir de las especificaciones del esquema conceptual, se define el esquema lógico. El esquema definido debe adaptarse con facilidad a los cambios que vayan surgiendo, ya que se sigue una metodología de diseño *agile*, abierta a la creación de nuevas entidades y atributos. Como se va a usar un sistema gestor de base de datos *MySQL*, se hace un modelo relacional de la base de datos. En la **figura 1.10** está representado el núcleo de este modelo. Las tablas principales de la base de datos son:

- PROTEIN: almacena información referente a las proteínas (identificativos *UniProt*, descripción) y parámetros que describen el entorno de la proteína en la red de interacción (coeficiente de *clustering*, conectividad). Está relacionada con la tabla TAXON que define la especie de la proteína, y con otras tablas (no representadas en la figura) que tienen información funcional, estructural y descriptiva de la proteína.
- METHOD: en esta tabla se definen atributos descriptivos de métodos experimentales que se usan para validar interacciones entre proteínas. La descripción y los identificativos de estos métodos se han obtenido de los vocabularios controlados definidos por *PSI*. Cada método tendrá un identificativo *OBO* único, un identificativo *PubMed* que hace referencia a donde se ha publicado

el método, una descripción corta y una definición proporcionada por *PSI*.

- PUBLICATION: aquí se almacena información sobre los artículos científicos que definen interacciones entre proteínas. Se identifican con un código *PubMed* único y se describen en forma de cita bibliográfica.
- INTERACTION: está relacionada con la tabla de proteínas por medio de dos identificativos que definen una interacción entre ellas. Además almacena información referente a la fiabilidad de la interacción (número de métodos que la validan y existencia de dominios que interaccionan en estructuras 3D). También incluye referencias a las bases de datos que han anotado esa interacción.
- INTMETPUB: relaciona las tablas de interacciones, métodos y publicaciones. De modo que una interacción estará demostrada por un método que ha sido descrito en una publicación. Además incluye referencia a la base de datos que ha anotado esta relación.

Finalmente, en el diseño físico se definen índices y estructuras de almacenamiento con el objetivo de disminuir el tiempo de acceso a los datos por la aplicación. Además se hace una desnormalización de la base de datos para simplificar las consultas que se van a hacer con mas frecuencia cuando se acceda a la información.

A modo de conclusión se puede indicar que el diseño de base de datos de *APID*, integra la información que tienen en común las bases de datos fuente para describir interacciones y añade nueva información útil para el usuario sobre las proteínas interactuantes, sobre la calidad de las interacciones y parámetros de la red de

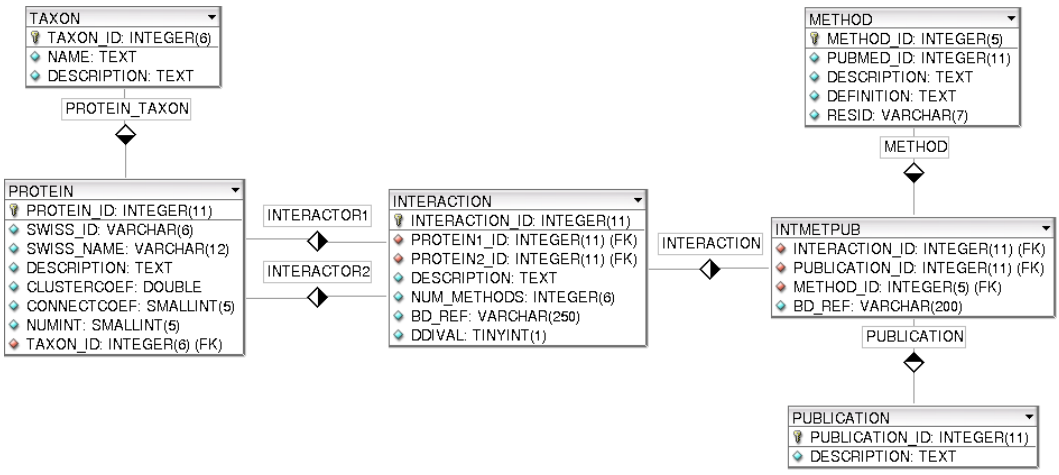


Figura 1.10: Modelo relacional de la base de datos de *APID*.
Diagrama del núcleo de atributos, entidades y relaciones del modelo de dicha base de datos.

interacción. Además se ha adaptado para minimizar el tiempo de acceso de la aplicación Web a la información.

1.2.4 Diseño de la aplicación Web de APID

APID está diseñado para ser una aplicación Web de acceso abierto, donde se han unificado interacciones proteína-proteína validadas experimentalmente, en una única plataforma accesible por Internet. Permite una exploración eficaz de la red de interacción e incluye parámetros que pesan la fiabilidad de las interacciones y ofrecen una visión del entorno de las proteínas en la red de interacción.

Respecto a las proteínas (como "nodos" de la red) se definen los siguientes parámetros:

- **Conectividad**: índica el número de proteínas que interaccionan directamente con la proteína buscada.
- **Coeficiente de *clustering***: indica cómo de interconectadas están las proteínas que interaccionan con la proteína buscada.
- **Entorno *GO***: muestra la información funcional que tienen las proteínas con las que está asociada la proteína buscada (esta información se define por medio de términos *GO*).
- **Enriquecimiento del entorno *GO***: muestra los términos *GO* más frecuentes del entorno de la proteína buscada que no están anotados para ella.

Respecto a las interacciones (como "enlaces" o "relaciones" de la red) se definen los parámetros:

- **Número de métodos**: número de métodos experimentales, publicados en artículos distintos, que validan una interacción entre proteínas.
- **Solapamiento *GO***: muestra los términos *GO* que tienen en común una pareja de proteínas interactuantes.
- **Dominios interactuantes**: muestra los dominios *Pfam* (Finn, *et al.*, 2008) que han sido co-cristalizados en una estructura tridimensional y que por ello se deduce que tienden a interactuar.

Actualmente *APID* integra la información de 6 bases de datos de interacción: *BIND*, *BioGrid*, *DIP*, *HPRD*, *IntAct* y *MINT*. De estas fuentes de información solo se han incluido las interacciones proteína-proteína (no entre proteínas y otros ligandos como

DNA u otras moléculas) que han sido validadas experimentalmente y descritas en un artículo científico indexado en *PubMed*. Además, no se han incluido las interacciones que tienen proteínas implicadas que no se han podido identificar en *UniProt* ya sea por tener un identificativo erróneo o ambiguo, o bien por no encontrar resultados al hacer *BLAST2* con su secuencia sobre *UniProt*.

Todas las interacciones tienen vínculos a las bases de datos donde han sido anotadas, de modo que se podrá acceder a información adicional en la base de datos original. También están vinculadas con los artículos científicos en los que se han descrito los ensayos que las validan. Por otro lado, las proteínas están vinculadas con la entrada de *UniProt* correspondiente y con otras bases de datos relacionadas (como *InterPro* (Hunter, *et al.*, 2009), *Pfam*, *Gene Ontology*, *Ensembl*, *NCBI Gene*) que describen diversas características de las proteínas.

El inicio de la navegación Web comienza en la búsqueda ("*search*"). La búsqueda en *APID* pretende ser lo más sencilla posible para el usuario. Permite introducir un texto libre que se buscará entre todos los identificativos que almacena la base de datos. Además tiene una búsqueda avanzada con la que se puede especificar el identificativo por el que se quiere buscar y combinar varios parámetros de búsqueda para refinar los resultados. La búsqueda devuelve las proteínas que tienen coincidencia con el texto buscado, indicando el ńumero de interacciones que tienen en la base de datos y mostrando información adicional que describe a la proteína. A partir de esta información se puede acceder a las interacciones de la proteína buscada. Las interacciones se describen con una pareja de identificativos de proteínas que se ha visto que interaccionan. Se puede filtrar por parámetros indicativos de su fiabilidad y acceder directamente a las bases de datos fuente donde se han anotado. Además, por cada interacción se indica el número de métodos experimentales que la han reportado. A través de un vínculo se puede ver una descripción de estos métodos y de las publicaciones científicas que los han reportado.

1.2.5 Medida de la fiabilidad de las interacciones proteína-proteína

En la introducción se habló sobre el ruido que introducen las técnicas proteómicas, pero ésta no es la única causa de la presencia de falsos positivos en los datos de interacción. Hay que tener también en cuenta posibles fallos que hayan ocurrido en el proceso de anotación de las interacciones, así como los diferentes tipos de

interacciones que se pueden dar entre las proteínas, que en muchos casos no están diferenciadas en las bases de datos. Según el tipo de ensayo que las haya validado las interacciones han podido ser demostradas *in vivo* o *in vitro*, de modo que a pesar de existir ensayos *in vitro* es posible que la interacción no se produzca realmente en la célula. Además, las interacciones pueden ser simplemente "funcionales", pero no interacciones físicas directas o físicas indirectas. Que una interacción sea funcional quiere decir simplemente que dos proteínas trabajan en el mismo proceso biológico. Las interacciones físicas indirectas hacen referencia a que las dos proteínas forman parte del mismo complejo pero no tienen porqué estar en contacto. Por último las interacciones físicas directas, que son el objetivo principal de las bases de datos de interacción, implican que las dos proteínas estén en contacto para formar la interacción. La dificultad de diferenciar este tipo de interacciones, hace que las interacciones directas se mezclen con los demás tipos de interacción provocando falsos positivos.

Uno de los objetivos de *APID*, es definir indicadores de fiabilidad de las interacciones, que se puedan aplicar en la base de datos unificada, y permitan filtrar las interacciones en busca de datos más fiables. Además de la aplicabilidad de este objetivo en *APID*, la validación de las interacciones proteína-proteína definidas en bases de datos, es una de las principales áreas de investigación en las que se ha trabajado en esta tesis. Por un lado, se han estudiado métodos de validación de interacción y por otro, mediante la aplicación de estos métodos, se han definido conjuntos de interacciones fiables que pueden ser usados como conjuntos (*sets*) de referencia en otras investigaciones.

El estudio de la cuantificación y mejora de la fiabilidad de las interacciones está siendo abordado por diversas aproximaciones bioinformáticas (Futschik, *et al.*, 2007; Ramirez, *et al.*, 2007) y nuevas técnicas experimentales (Cusick, *et al.*, 2005; Stelzl y Wanker, 2006; Yu, *et al.*, 2008). Uno de estos estudios lo hizo *Von Merring et al*. (von Mering, *et al.*, 2002), en el que comparaba la fiabilidad de métodos experimentales y predictivos de interacción de proteínas. El estudio comprobó la heterogeneidad existente entre los distintos métodos y determinó que la mayor fiabilidad se obtenía cuando se combinaban varios métodos que validan las interacciones. Esta medida de fiabilidad se puede aplicar en *APID*, considerando que las interacciones que han sido demostradas en varios ensayos son más fiables. Con esta idea el número de ensayos que valida una interacción, está claramente indicado en cada pareja de proteínas que

interactúa, y se pueden filtrar las interacciones visualizadas en función del número de métodos que las validan.

Otro tipo de aproximaciones para validar interacciones están basados en bioinformática integrativa. Estas estrategias integran información relacionada con las proteínas, con el objetivo de diseñar un método que consiga aumentar la fiabilidad de los conjuntos de interacciones. En *APID* se ha seguido este tipo de aproximaciones integrando información funcional y estructural. Como información funcional se añadió información proveniente de *Gene Ontology* (*GO*), para anotar las funciones, los procesos biológicos y la localización celular de cada proteína. Con esta información y basándose en la suposición de que las proteínas con términos *GO* iguales tienen más probabilidad de interactuar (por estar implicados en procesos biológicos similares o estar en la misma localización celular), se ha resaltado en *APID* la coincidencia de términos *GO* en las interacciones de proteínas. Sin embargo, no se puede establecer una relación directa entre el hecho de que dos proteínas tengan términos *GO* relacionados y que interactúen, ya que una cosa no conlleva la otra. De todos modos, es importante resaltar la coincidencia de términos *GO* en la información de la interacción, como otro parámetro que puede mejorar la confianza sobre dicha interacción.

Respecto a la información estructural que se ha integrado, se ha hecho un esfuerzo importante en recopilar dominios deducidos como posibles interactuantes en base a estructuras tridimensionales. Esta información será útil para validar datos de interacción, ya que se ha visto que muchas interacciones físicas entre proteínas ocurren por medio de dominios que tienden a interactuar (Aloy y Russell, 2006). Este

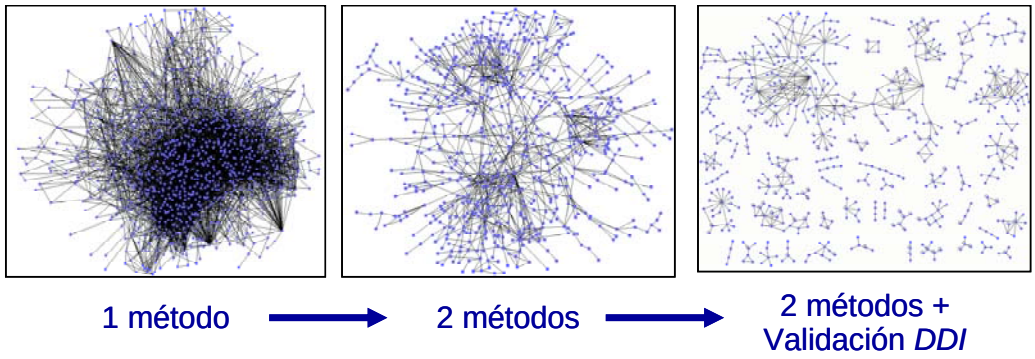


Figura 1.11: Ejemplo de aplicación de métodos de validación.
Muestra tres redes de interacción: **Viñeta 1**) Red de interacción humana obtenida de APID. **Viñeta 2**) Resultado de filtrar (en la red de la viñeta 1) las interacciones que han sido demostradas solo por un método. **Viñeta 3**) Resultado de filtrar (en la red de la viñeta 2) las interacciones que no tienen dominios interactuantes conocidos.

estudio está descrito detalladamente en el siguiente apartado y se ha integrado en *APID* como una medida de fiabilidad de las interacciones, de modo que se puede filtrar interacciones según tengan o no dominios interactuantes.

Un ejemplo de validación de interacciones proteína-proteína está representado en la **figura 1.11**. Se parte de una red proveniente de *APID* (**viñeta 1**) de la que se filtran las interacciones que han sido demostradas en un solo ensayo (**viñeta 2**) y las que no tienen dominios interactuantes (**viñeta 3**). Se observa como se va clarificando la red de interacción a medida que ésta va siendo más fiable.

La búsqueda de parámetros o métodos que midan la fiabilidad de las interacciones, es un área de investigación que sigue abierta en *APID*. Dada la importancia que tiene el evitar la presencia de falsos positivos en los datos de interacción, una de las principales líneas futuras de trabajo estará relacionada con este objetivo.

1.2.6 Integración de interacciones estructurales entre dominios para mejorar los datos de interacción

El conocimiento de estructuras tridimensionales de proteínas esta creciendo rápidamente (ver **figura 1.12**), de modo que se empieza a tener una cobertura significativa de proteínas con estructura conocida en los proteomas. Usando la información estructural que hay almacenada en el repositorio *Protein Data Bank*

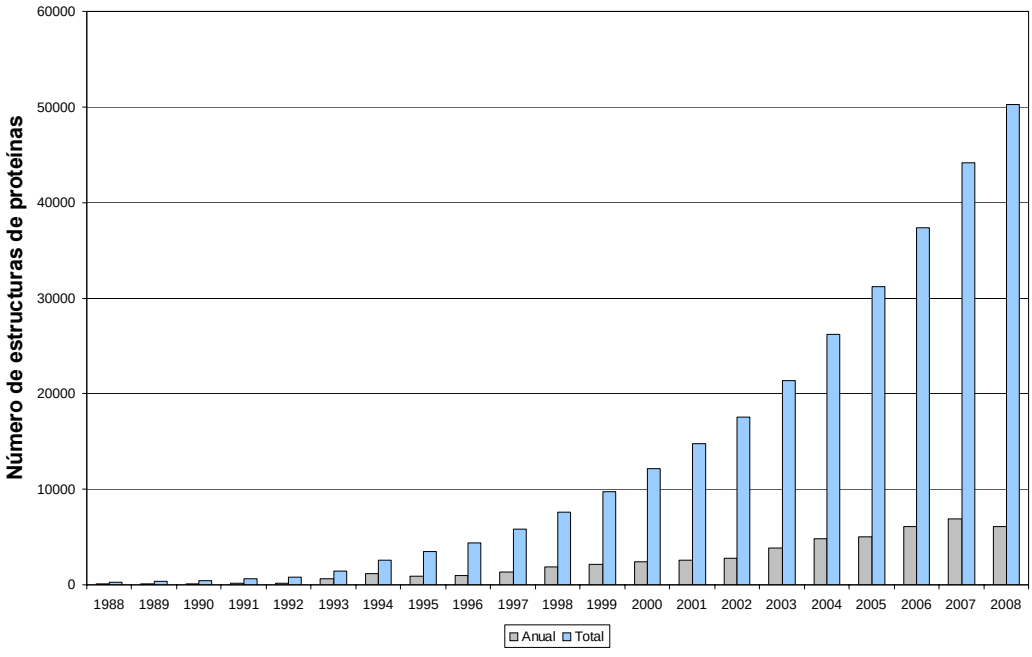


Figura 1.12: Crecimiento de PDB.
Diagrama de barras en el que se representa el crecimiento del número de estructuras 3D de proteínas resueltas por año.

(*PDB*), varios estudios bioinformáticos han intentado describir interacciones entre dominios mediadas por interfaces estructurales. Estos estudios han explorado la información estructural almacenada en *PDB* para definir familias o dominios estructurales que interaccionan. Para ello han desarrollado diversas aproximaciones que suelen estar basadas en: (**i**) identificar dominios con regiones cercanas en el espacio, (**ii**) buscar interfaces o patrones de unión entre dominios o (**iii**) definir interfaces de interacción conservados.

Como se comentó en el apartado anterior la integración de datos de interacción dominio-dominio (*DDI*) con datos de interacción de proteínas, puede ser muy útil para mejorar la fiabilidad de los datos de interacción, ya que las interacciones físicas directas se dan mediante *DDIs*. Esta integración no es trivial ya que los datos de *DDI* han sido generados por métodos diversos muchas veces predictivos, que dan lugar a *sets* muy diferentes. Por tanto es necesario hacer una comparativa de los *sets* de *DDI* para comprobar si se pueden usar para mejorar la fiabilidad de las interacciones de proteínas.

Se integraron 6 recursos de interacción entre dominios estructurales que han sido publicados: *3DID* (Stein, *et al.*, 2009), *CBM* (Shoemaker, *et al.*, 2006), *iPFAM* (Finn, *et al.*, 2005), *PIBASE* (Davis y Sali, 2005), *PSIBASE* (Gong, *et al.*, 2005), *SNAPPI* (Jefferson, *et al.*, 2007). La integración se hizo transformando todos los identificadores de dominios a códigos *Pfam* (es decir en *Pfam* IDs). Todos estos recursos están basados en datos de estructuras 3D, pero han aplicado sus propios métodos para definir *DDI*. En la **tabla 1.1**, se muestra el número de *DDI* que tienen los distintos recursos (en azul en la diagonal) y el número de *DDI* que tienen en común unos recursos con otros (en negro). En este estudio comparado se han evitado las interacciones entre dominios iguales (es decir, lo que serían homodímeros u homo-oligómeros). Como se puede observar hay una gran heterogeneidad entre los datos

	3did	CBM	iPFAM	PIBASE	PSIbase	SNAPPI
3did	**2.385**	562	1.559	1.039	1.047	974
CBM	562	**5.335**	553	908	794	460
iPFAM	1.559	553	**2.027**	1.063	1.055	1.078
PIBASE	1.039	908	1.063	**9.742**	6.471	789
PSIbase	1.047	794	1.055	6.471	**6.978**	820
SNAPPI	974	460	1.078	789	820	**1.134**

Tabla 1.1: Número de *DDI*.
Número de *DDI* que tiene cada recurso bioinformático (en **azul**) y el número de *DDI* que tienen en común entre ellos (en **negro**).

obtenidos por diferentes métodos, de hecho el porcentaje de similitud por ejemplo para *iPFAM* o *CBM* ronda el 10% frente a las otras recopilaciones de *DDIs*. Además se observa que el número total de dominios interactuantes difiere mucho en función del método que se ha usado, de modo que el estudio más amplio (*PIBASE*) encuentra 9742 parejas de dominios interactuanes mientras que *SNAPPI* tan solo tiene 1134 parejas.

Para estudiar en qué medida pueden ayudar estos recursos a validar interacciones de proteínas, se definieron 3 *sets* de referencia de interacción proteína-proteína: (**i**) Conjunto de todas las interacciones que almacena *APID* (241204 interacciones entre 51873 proteínas); (**ii**) Subconjunto de las interacciones de *APID* que han sido validadas por 2 o más ensayos (37918 interacciones entre 15266 proteínas); (**iii**) Subconjunto de *APID* en el que las proteínas están anotadas al mismo identificativo *KEGGs* (Kanehisa, 2002) de ruta molecular (12315 interacciones entre 2290 proteínas).

Sobre estos conjuntos de referencia se midió la cobertura que tienen los recursos de *DDI*. Los resultados de este análisis están en la **tabla 1.2**. Como se observa, la mayor cobertura se da en el *set* de referencia construido con *KEGGs*, que ronda el 20% para todos los recursos de dominios interactuantes excepto para *CBM* que solo tiene un 4,7%. El hecho de que el *set* de *KEGGs* tenga mayor cobertura que el resto, es posible que sea por la existencia de máquinas moleculares o complejos que han sido cristalizados. Con el objetivo de aumentar la cobertura se ha hecho una unión de los recursos de *DDI*. De este modo, se incrementa la máxima cobertura en el *set* de referencia de *KEGG* a un 30% (ver columna Unión en la **tabla 1.2**). Si la unión de los recursos puede validar significativamente interacciones de proteínas, es más beneficioso utilizar esta unión para aumentar el número de interacciones que se

		3did	CBM	iPFAM	PIBASE	PSIbase	SNAPPI	Unión
APID todos	Validados	7.061	2.500	7.050	9.354	8.260	5.245	12.584
	Positivos	164.357	164.357	164.357	164.357	164.357	164.357	164.357
	Cobertura	4,30	1,52	4,29	5,69	5,03	3,19	7,66
APID 2 métodos	Validados	2.614	777	2.697	3.035	2.818	1.996	4178
	Positivos	29.095	29.095	29.095	29.095	29.095	29.095	29.095
	Cobertura	8,98	2,67	9,27	10,43	9,69	6,86	14,36
APID =KEGG	Validados	1.097	258	1.204	1.266	1.233	930	1.634
	Positivos	5.485	5.485	5.485	5.485	5.485	5.485	5.485
	Cobertura	20,00	4,70	21,95	23,08	22,48	16,96	29,79

Tabla 1.2: Cobertura de recursos de *DDI* en el interactoma.
Estudio de la cobertura que tiene la validación con *DDI* en los tres *sets* de referencia de interacción que se han definido procedentes de *APID*.

validan. También es interesante la observación de que más de un 60% de los *DDI* del *set* de referencia de *KEGG* están en varios recursos *DDI*.

Se comprobó también que la validación de las interacciones no se realizaba por mero azar. Para ello se compararon los resultados obtenidos por los *sets* de *DDI* con conjuntos aleatorios de parejas de *DDI*. Estos conjuntos aleatorios se construyeron partiendo de los conjuntos con los que se iban a comparar, cambiando unos dominios por otros, de modo que los parámetros y la estructura de la red de interacción de dominios permanecieron inalterados. Se generaron 1000 conjuntos aleatorios por cada recurso y se calculó la media y la desviación estándar del número de positivos que se habían obtenido aleatoriamente. De este modo, se obtuvo un ***Z-Score*** para ver como mejoran los resultados de los sets de *DDI* respecto a los conjuntos aleatorios generados. Cuanto mayor sea el ***Z-Score*** más fiable será el recurso, de modo que si está por encima de 10 se puede considerar que los resultados no son debidos al azar. Los resultados obtenidos fueron positivos (ver **tabla 1.3**), ya que dieron ***Z-Scores*** muy altos, indicando que la validación por medio de dominios interactuantes está muy lejos de producirse por azar. Además el ***Z-score*** obtenido por la unión de los 6 recursos de *DDI* es uno de los más altos, con lo que podemos concluir que la integración de recursos de *DDI* es un método muy válido para mejorar la fiabilidad de las redes de interacción de proteínas.

		3did	CBM	iPFAM	PIBASE	PSIbase	SNAPPI	Unión
APID todos	Verdaderos Positivos	7.061	2.500	7.050	9.354	8.260	5.245	12.584
	Avg Positivos Aleatorios	30,89	86,44	77,91	96,65	89,30	33,30	86,46
	Sd Positivos Aleatorios	28,63	50,23	44,26	66,68	57,13	24,58	54,89
	Z-Score	245,54	48,05	157,54	138,84	143,02	212,02	227,70
APID 2 métodos	Verdaderos Positivos	2.614	777	2.697	3.035	2.818	1.996	4.178
	Avg Positivos Aleatorios	24,85	90,74	74,73	97,26	94,43	31,89	92,89
	Sd Positivos Aleatorios	21,65	49,28	43,58	60,27	57,50	22,67	59,88
	Z-Score	119,57	13,93	60,17	48,74	47,37	86,62	68,22
APID KEGG igual	Verdaderos Positivos	1.097	258	1.204	1.266	1.233	930	1.634
	Avg Positivos Aleatorios	32,00	85,70	77,43	90,19	93,25	28,14	83,42
	Sd Positivos Aleatorios	26,42	46,84	42,09	62,90	57,23	23,05	55,48
	Z-Score	40,31	3,68	26,77	18,69	19,91	39,13	27,95

Tabla 1.3: Validación de interacciones con *DDI* frente a conjuntos aleatorios.
Comparación del método de validación con *DDI* frente a conjuntos aleatorios de *DDI*. Se observa un *Z-score* alto, que indica que la validación está muy lejos de producirse al azar.

1.2.7 Navegación Web

La **figura 1.13** representa de forma esquemática los pasos que hay que realizar para consultar las interacciones de una proteína en la aplicación. De esta forma se muestra un flujo de trabajo y se clarifica la información y las herramientas que *APID* ofrece. La **viñeta 1** muestra la página de búsqueda de proteínas. En la caja de texto se pueden poner nombres, identificativos y descripciones de proteínas. Como ejemplo se buscó la proteína humana *hras* insertando su identificativo *UniProt* '*RASH_HUMAN*'. Esta búsqueda dio un solo resultado que se muestra en la **viñeta 2**, en la que hay una tabla con seis columnas de información sobre la proteína: (**i**) nombre *UniProt*, (**ii**) número de interacciones, (**iii**) identificativo *UniProt*, (**iv**) identificativo de taxón *NCBI*, (**v**) descripción y (**vi**) un vínculo a información adicional sobre la proteína. Haciendo clic sobre '+*info_prot*', se abre una nueva ventana con información detallada sobre la proteína que incluye vínculos a otras bases de datos biomoleculares de referencia, parámetros de red (conectividad y coeficiente de agrupamiento) e información de su entorno funcional basado en *GO* (*GO environment*). Se observa que la conectividad es 84 que se corresponde con el número de proteínas que interaccionan con *hras* de humano. Éste es un número muy grande de interacciones que es posible que incluya muchos falsos positivos. Haciendo clic sobre el número en rojo de la **viñeta 2**, se muestra una nueva página con información de las 84 interacciones que se han reportado para *hras*. La página está representada en la **viñeta 3** y contiene una tabla con 6 columnas de información sobre: (**i**) los nombres de las proteínas interactuantes, (**ii**) el número de métodos y de anotaciones en bases de datos que validan la interacción, (**iii**) el número de dominios estructurales interactuantes que hay entre proteínas, (**iv**) vínculos a las bases de datos donde está descrita la interacción y (**v**) un vínculo a más información sobre la interacción. Haciendo clic sobre '+*info_inter*' se abre una nueva ventana con información adicional sobre la interacción, en la que aparecen marcados los términos *GO* que tienen en común las dos proteínas (en amarillo), así como los dominios *Pfam* que se ha visto que interaccionan (en verde). La página que muestra las interacciones (**viñeta 3**) permite además filtrarlas en función del número de experimentos y la validación estructural. En el ejemplo se han filtrado las interacciones con menos de 4 métodos experimentales y sin una validación estructural, de modo que el número de interacciones se ha reducido a 8. También se puede consultar información sobre los métodos y los dominios interactuantes que

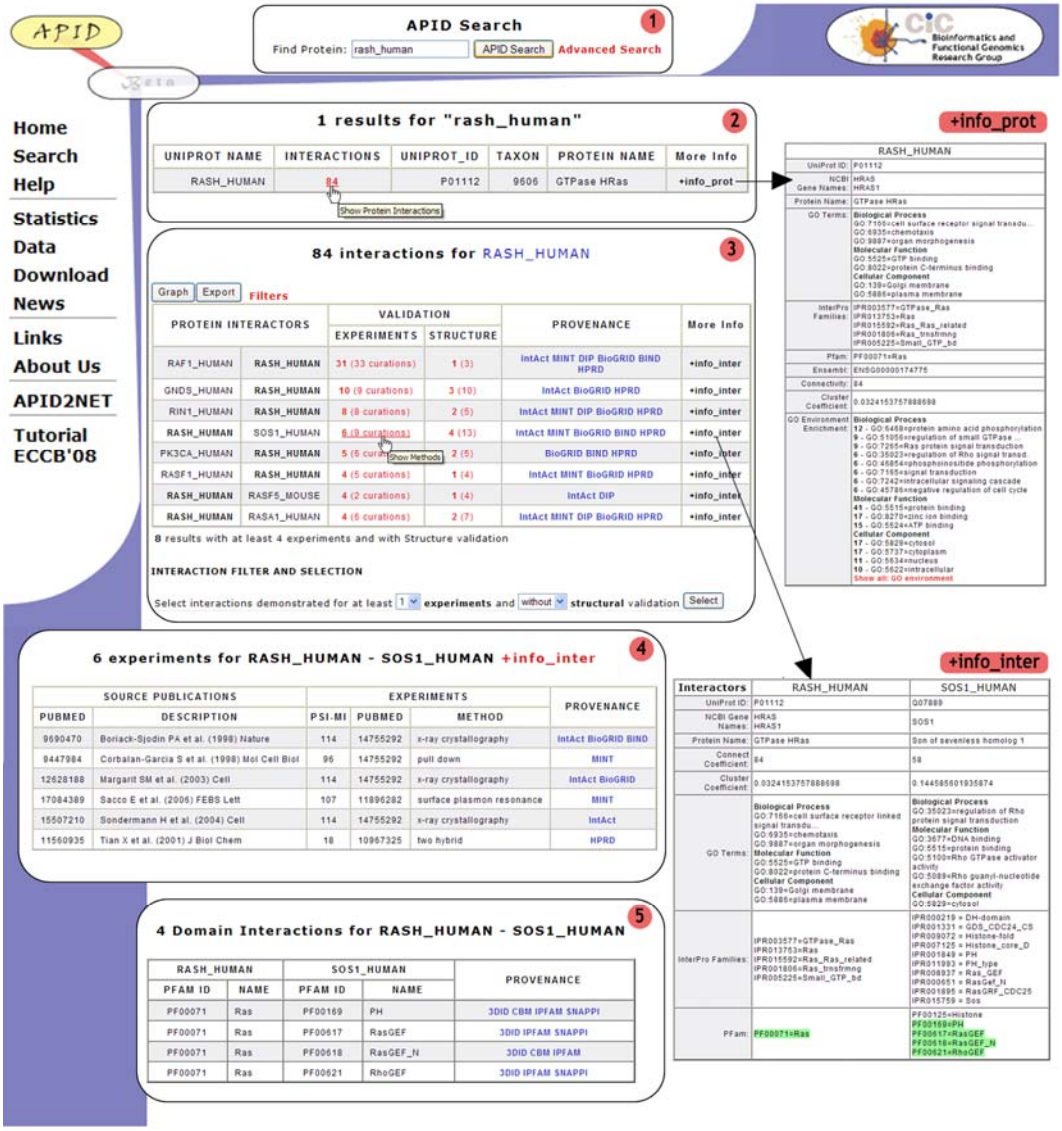


Figura 1.13: Ejemplo de flujo de trabajo en *APID*..
Búsqueda de la proteína 'rash_human' (viñeta 1). Proteína que ha encontrado la búsqueda (viñeta 2) con información adicional que la describe (+info_prot). La proteína encontrada tiene 84 interacciones y se han filtrado las que tienen menos de 4 experimentos y no tienen validación estructural (viñeta 3). Cada interacción tiene un vínculo a información adicional (+info_inter). Los métodos experimentales que validan cada interacción se muestran haciendo clic sobre el número correspondiente (viñeta 4). Las parejas de dominios interactuantes que tienen las proteínas se muestran haciendo clic sobre el número de la columna Structure (viñeta 5). Cada viñeta se corresponde a páginas web consecutivas en APID.

validan las interacciones, haciendo clic sobre el número correspondiente. La **viñeta 5** muestra la información de los métodos experimentales de la interacción entre *hras* y *sos1*, en una tabla con 5 columnas de información sobre: (**i**) el identificativo *PubMed* del artículo que define la interacción, (**ii**) una descripción breve del artículo, (**iii**) el código *PSI-MI* del método experimental, (**iv**) el identificativo *PubMed* del artículo donde se define el método experimental, (**v**) el nombre del método y (**vi**) vínculos a las bases de datos que lo han anotado. Finalmente, la **viñeta 6** muestra una página con información a cerca de parejas dominios que diversos recursos han deducido que

interaccionan, y que están presentes en las proteínas que forman la interacción (*hras* y *sos1*).

1.2.8 Estadísticas

El mejor indicativo de la utilidad de una página web es el número de usuarios que tiene. Desde el comienzo de *APID*, hemos ido recopilando estadísticas detalladas de todas las visitas que ha recibido para poder valorar el número de usuarios y los accesos

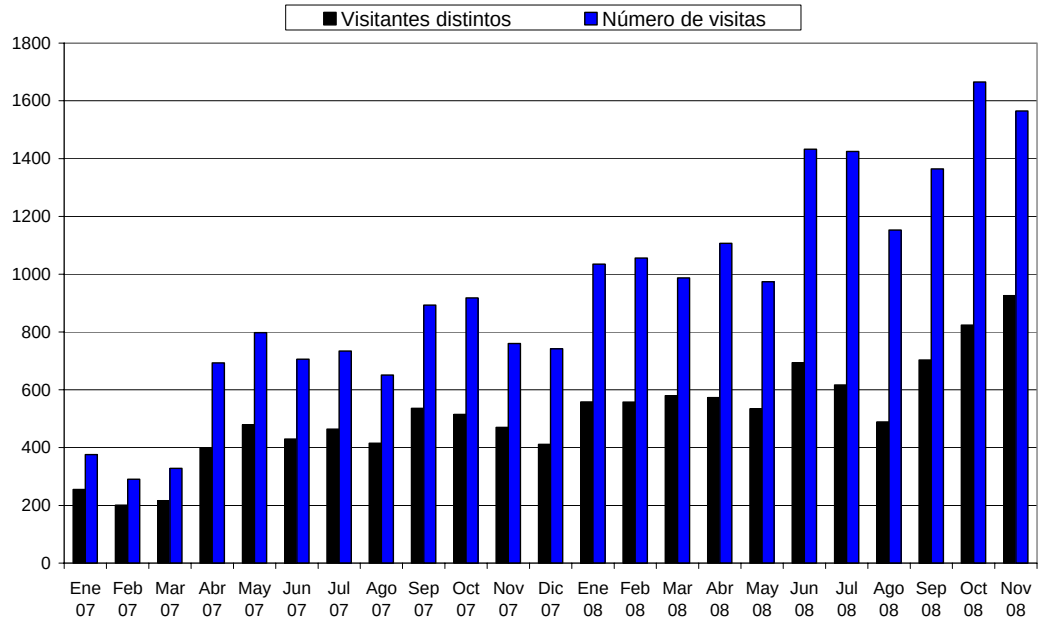


Figura 1.15: Estadísticas de acceso a *APID*.
Diagrama de barras con el número de visitantes distintos (en **negro**) y el número de visitas (en **azul**) que ha recibido *APID* cada mes en los últimos dos años.

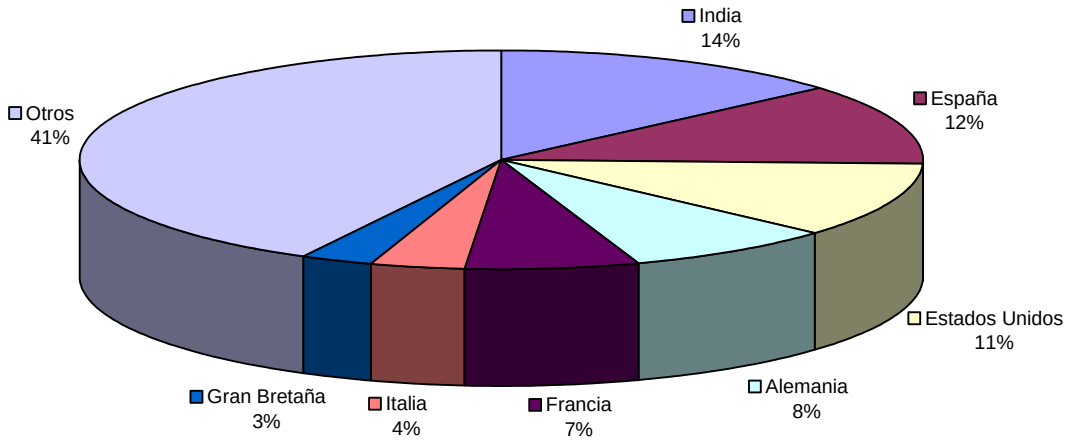


Figura 1.15: Nacionalidad de los usuarios de APID.
Diagrama circular que muestra el porcentaje de páginas visitadas por los usuarios de los países que mas accesos hacen a la aplicación.

que se han hecho. La **figura 1.14** muestra el número de usuarios distintos cada mes y las visitas que han realizado mensualmente a la aplicación en los últimos dos años. Como se observa tanto el número de usuarios como el número de visitas se ha ido incrementando a lo largo de estos años, de tal modo que se han llegado a cuadriplicar en dos años. En la actualidad unos **1000** usuarios distintos hacen alrededor de **1600** visitas cada mes a *APID*. Esta cifra es muy superior a la del mes de Enero de 2007 en el que **255** usuarios hicieron **376** visitas. Respecto a la procedencia de los visitantes, la **figura 1.15** muestra el porcentaje de páginas que se han visitado agrupadas por países. Como se observa, 5 países acumulan más del 50% de las páginas visitadas: India, España, Estados Unidos, Alemania y Francia.

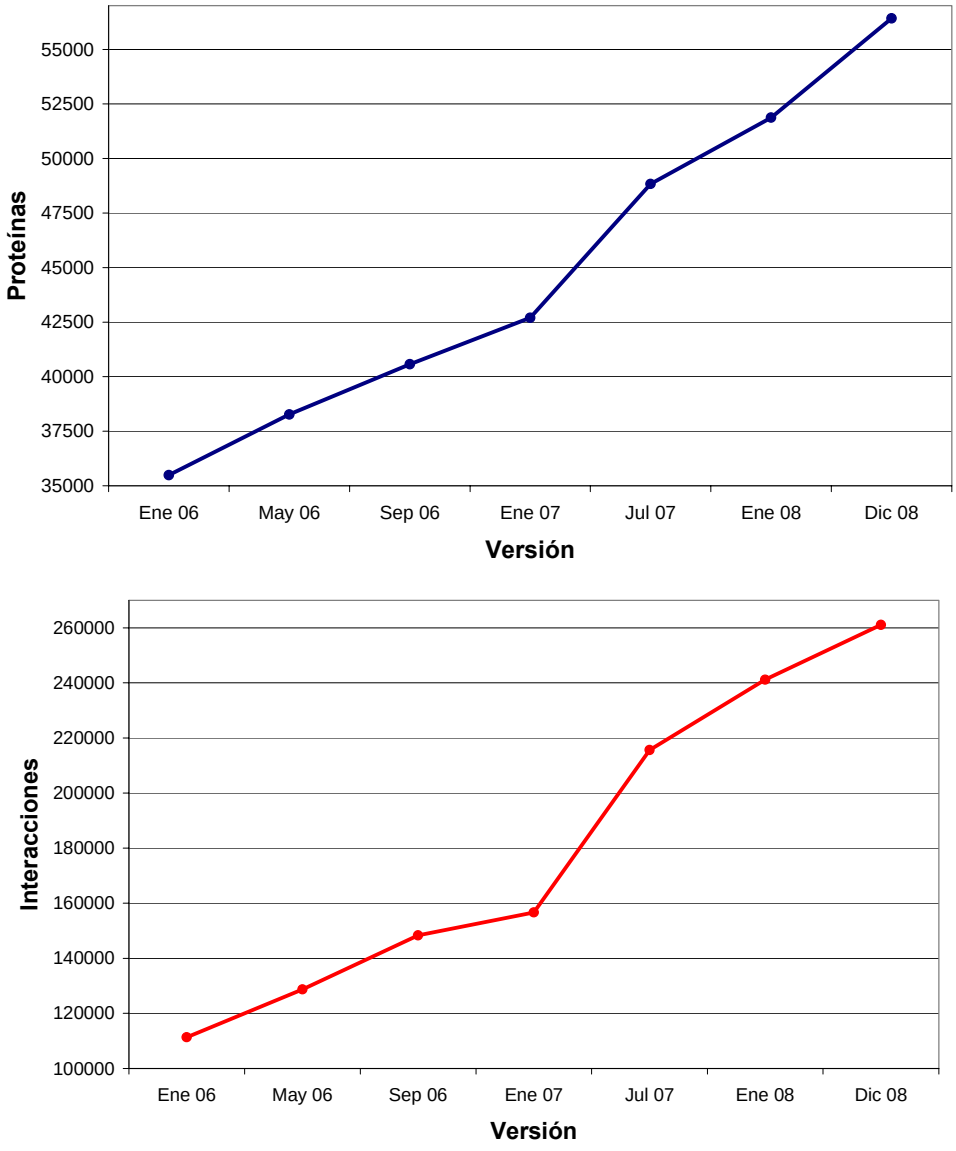


Figura 1.16: Crecimiento de los datos en *APID*
Número de proteínas (**a**) e interacciones (**b**) que han formado las ultimas 7 versiones de *APID*.

A lo largo de estos años no solo ha crecido el número de usuarios que usan *APID*, sino que también ha crecido el número de datos de interacción que almacena. La **figura 1.16** muestra el crecimiento de interacciones (**figura 1.16b**) y proteínas (**figura 1.16a**) que ha experimentado *APID* en sus últimas 7 versiones. Como se observa en tres años se ha duplicado el número de interacciones que tiene almacenadas *APID*. Se ha pasado de **111332** a **261063** interacciones y de **35495** a **56422** proteínas. Todo esto se ha producido por la inclusión de nuevas fuentes de datos y por las periódicas actualizaciones de la base de datos que se han realizado.

Otro dato a tener en cuenta es el número de métodos que validan a las interacciones. Todas las interacciones han sido validadas por al menos un método experimental y es interesante que estén validadas en diferentes ensayos ya que esto permite mejorar la fiabilidad de los datos. Sin embargo, en este aspecto *APID* solo ha conseguido que el **6%** de las interacciones estén validadas por más de un método experimental (ver estadísticas en su web).

Por último en la **figura 1.17**, se ha representado el número de proteínas que hay en *APID* en función de su conectividad. Como se observa la curva sigue una distribución de ley exponencial (*power law distribution*) que suelen seguir las redes con topología libre de escala (scale free). Esta topología es la que clásicamente se ha definido para las redes de interacción proteína-proteína (Barabasi y Oltvay 2004).

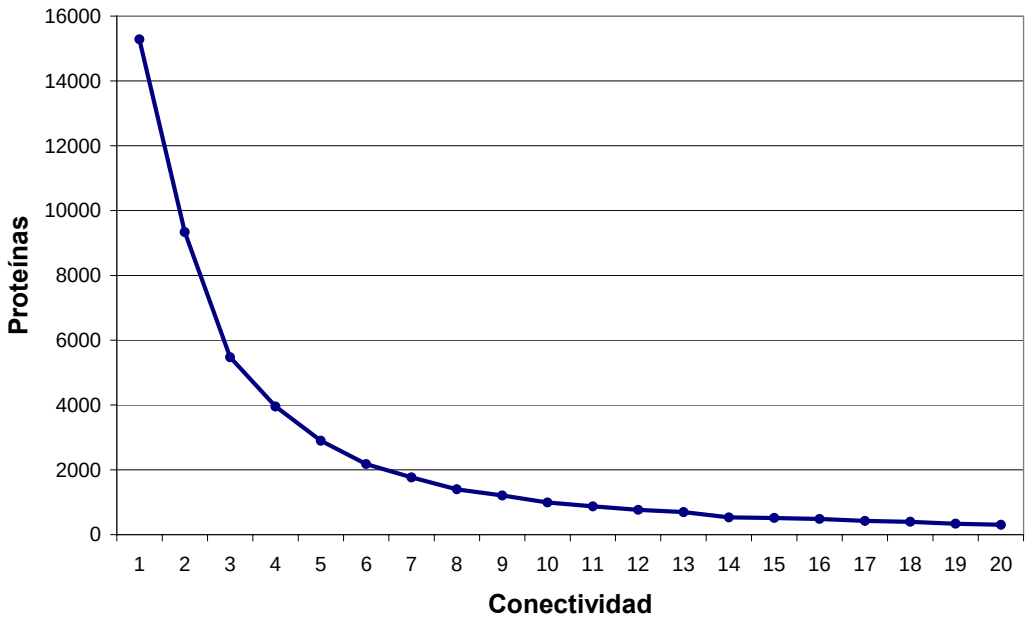


Figura 1.17: Conectividad de las proteínas almacenadas en *APID*.
Gráfico que muestra el número de proteínas que hay en *APID* en función de su conectividad. Se observa como la curva sigue una distribución de ley exponencial.

1.3 CONCLUSIONES

Una de las principales razones que motivaron el trabajo realizado en este capítulo fue la gran heterogeneidad y falta de convergencia observada entre las bases de **datos de interacción entre proteínas**. Esta heterogeneidad se refiere tanto a la información biológica que contienen, como a los formatos que usan para almacenar las interacciones. En los últimos años se han creado estándares para describir e intercambiar datos de interacción biomolecular, que la comunidad científica está empezando a seguir. Sin embargo, estos estándares a veces se aplican de modo no estricto y con diversas interpretaciones, lo cual provoca que todavía sea difícil combinar de modo práctico diversas fuentes de datos. Por este motivo, al iniciar este trabajo se vio necesaria la organización y la recopilación ordenada de todas las interacciones biomoleculares conocidas y descritas entre proteínas, para facilitar el acceso a una **información unificada y no redundante**. Éste es uno de los objetivos conseguidos con *APID*, que recopila y clasifica la información de las principales bases de datos de interacción entre proteínas en una única **plataforma *web*** de acceso libre.

Además en este capítulo se ha resaltado el **elevado número de falsos positivos** que hay en los datos de interacción proteína-proteína derivados de distintas técnicas proteómicas experimentales; lo cual hace necesario estudiar técnicas de **validación**, para conseguir una fiabilidad mayor de los interactomas. La unión e integración de los recursos de interacción, además de aumentar la cobertura de los interactomas conocidos, ha posibilitado el diseño de una serie de parámetros cuantitativos para mejorar la fiabilidad de los datos de interacción. Un parámetro de fiabilidad derivado de la unificación de las interacciones, es en primer lugar el **número de experimentos diferentes publicados** que las validan y el **número de métodos experimentales** distintos utilizado para identificarlas, de modo que se obtienen interacciones mas fiables cuanto mayores son estos números, que hemos denominado de modo genérico: **número de experimentos**. Otro parámetro de validación que hemos utilizado es la información estructural proveniente de métodos capaces de determinar o predecir la **interacción específica dominio-dominio en estructuras de proteínas**. La información estructural es sin duda clave para mejorar la fiabilidad de las redes de interacción, ya que toda interacción molecular tiene una base estructural física. El problema con los datos estructurales es la falta de cobertura, por ello, en este trabajo se han estudiado un conjunto importante de recursos que deducen interacciones dominio-

dominio a partir de dominios estructurales que han sido co-cristalizados. Estos recursos se han integrado y unificado para aumentar su cobertura sobre los interactomas. Además se ha comprobado que este tipo de validación de interacciones está lejos de producirse por mero azar y aporta un método muy útil para clarificar la complejidad del interactoma y filtrar falsos positivos de la red.

Capítulo 2

Visualización y análisis de redes de interacción de proteínas

2.1 INTRODUCCIÓN

2.1.1 Representación de datos e interacciones en redes

La visualización de datos es una disciplina transversal que utiliza el poder de comunicación de las imágenes para explicar las relaciones de significado, causa y dependencia que se pueden encontrar en conjuntos de información asociada a procesos complejos, como son los procesos biológicos [27].

Las técnicas de visualización han sido muy aplicadas a datos de interacción proteína-proteína. La representación más evidente para visualizar los datos de interacción es la

de una red, en la que los nodos representan a las proteínas y los vínculos entre nodos representan a las interacciones. Esta representación nos da una visión global de los mapas de interacción y nos permite explorar fácilmente los datos para clarificar su significado. Sin embargo, la representación de los datos de interacción no es un objetivo sencillo debido a que algunas especies tienen un elevado número de interacciones conocidas, que está creciendo exponencialmente debido a la aplicación de técnicas proteómicas de alto rendimiento. Este elevado número de interacciones provoca que las redes se representen como un ovillo difícil de desentrañar y del que es complicado deducir visualmente nada. Para solucionar este problema es importante que las herramientas de visualización permitan aplicar filtros sobre la red para simplificarla. Estos filtros pueden estar basados, por ejemplo, en información descriptiva de las proteínas (sobre localización, función o estructura) o en información relativa a los complejos y rutas moleculares presentes en la red. Por último, también hay que tener en cuenta la dificultad que conlleva trabajar con varios tipos de interacciones entre proteínas (ver capítulo 1), las cuales, en muchas ocasiones, se representan de forma indistinta ya sea por desconocimiento o por la dificultad de diferenciarlas gráficamente.

La representación de las interacciones proteína-proteína como un mapas bien caracterizados, además de permitir la visualización de los datos en forma de red, ha permitido aplicar teoría de análisis de redes, lo cual ha ayudado a revelar diversas hipótesis sobre los sistemas biológicos. Por ejemplo, se ha visto que las redes biológicas tienen una topología libre de escala (*scale free*) (Barabasi y Oltvai, 2004), ya que hay un reducido número de proteínas que está implicado en muchas interacciones y un gran número que está implicado en pocas (Barabasi y Albert, 1999). Este tipo de redes son más robustas a la eliminación de un nodo de forma aleatoria, ya que la mayoría de los nodos no hacen funciones esenciales de la célula. Sin embargo, son más frágiles cuando suceden alteraciones o eliminaciones de alguno de los nodos muy conectados (llamados *hubs*), ya que esto puede causar un desajuste celular bastante dramático.

De acuerdo con lo expuesto, es importante conseguir representaciones de los datos de interacción para lograr que la información sea más comprensible y facilitar la deducción e inferencia de nuevo conocimiento. Por tanto es interesante el desarrollo de herramientas bioinformáticas que faciliten el acceso a los datos de interacción, por

medio de métodos de visualización, que permitan hacer representaciones selectivas y análisis de los datos.

2.1.2 Herramientas de visualización de redes biológicas

El desarrollo de herramientas de visualización de redes biológicas es un área en la que se ha trabajado mucho en los últimos años. El objetivo de las herramientas de visualización es conseguir representaciones flexibles, adaptables y claras para el usuario. Para ello se intenta que el usuario pueda personalizar las representaciones, mediante la modificación de parámetros de visualización (color, tamaño, forma, texto) y mediante la ejecución de algoritmos de distribución de nodos (*layouts*).

Para mejorar el uso y la disponibilidad de las aplicaciones, es importante tener en cuenta los formatos de importación-exportación de datos que manejan los programas de visualización y análisis de redes. De este modo para importar redes moleculares es interesante facilitar el manejo de ficheros en formatos *PSI-XML* y *XLS* (hojas de cálculo). En cuanto a la exportación, esta tiene que permitir guardar el trabajo que el usuario ha realizado sobre la red, para esto se suelen usar archivos en formato gráfico, vectorial, *XML* o formatos desarrollados por la propia aplicación.

Una característica que vez es más común en las aplicaciones avanzadas de visualización de redes biológicas, es que permitan extender sus funcionalidades por medio de *plugins*. Los *plugins* son programas que se unen a una aplicación ya existente, para añadirle nuevos servicios o funcionalidades. Para que se puedan desarrollar es necesario que se proporcione una interfaz que permita al *plugin* comunicarse con la aplicación y que se faciliten métodos y funciones de programación para su creación e instalación. Esta es una característica muy interesante, ya que permite a investigadores y desarrolladores hacer sus propias aplicaciones sobre una plataforma de representación de redes, de modo que la aplicación aumenta sus funcionalidades y el autor del *plugin* consigue difundir su trabajo con mayor facilidad.

En los últimos años se han desarrollado varias aplicaciones que permiten explorar mapas moleculares y ofrecen diversos métodos para analizar redes de interacción de proteínas. Las más populares y de uso gratuito son:

- **Cytoscape** (Shannon, *et al.*, 2003): es una plataforma abierta (*open source*) muy conocida de representación y análisis de redes biológicas. Permite personalizar

todos los parámetros de visualización y tiene una gran cantidad de formatos de entrada/salida. La mejor característica de esta aplicación es la facilidad que da a la hora de desarrollar *plugins*; así ha conseguido que en sus 5 años de vida se hayan desarrollado más de 50 *plugins* que amplían considerablemente sus capacidades. Los *plugins* están clasificados según su función en 5 categorías: análisis (*analysis*), entrada/salida de redes y atributos (*network and attribute I/O*), inferencia de redes (*network inference*), enriquecimiento funcional (*functional enrichment*) y comunicación/guiones de ejecución (*communication/scripting*). Estos *plugins* hacen que *Cytoscape* sea la aplicación que ofrece más herramientas para trabajar con datos biológicos.

- **Osprey** (Breitkreutz, *et al.*, 2003): fue una de las primeras herramientas desarrolladas en el área de la bioinformática para visualizar y analizar grandes redes. Está conectado con la base de datos *BioGrid* y permite aplicar filtros en función de la anotación a *GO* de las proteínas o del tipo de ensayos que validan las interacciones. Tiene muchas opciones para personalizar el aspecto de la representación y la disposición de los nodos, pero está muy limitado en opciones de importación de datos.
- **ProViz** (Iragne, *et al.*, 2005): esta aplicación puede representar de forma eficiente redes de gran tamaño gracias a la librería Tulip (David, 2001) que está programada en C++ y se ocupa de la visualización de la red. Como datos de entrada permite el uso de ficheros en formato *PSI-MI*, lo que hace que se puedan importar los datos de las bases de datos de interacción de proteínas más importantes. La exploración de la red permite resaltar las funciones de las proteínas con *GO* y la información de las interacciones a través del vocabulario controlado definido por *PSI*. También permite la integración de *plugins* en la aplicación, aunque no ha tenido mucha aceptación entre la comunidad de desarrolladores bioinformáticos.
- **VisANT** (Hu, *et al.*, 2008): es una plataforma Web de visualización de redes que muestra la información de varias bases de datos de interacción de proteínas, y da la posibilidad de conectarse a su servidor para guardar y recuperar las sesiones de trabajo que se han hecho. No realiza ningún tipo de análisis funcional pero permite cargar datos de expresión para visualizarlos sobre las interacciones.

2.2 MÉTODOS Y RESULTADOS

2.2.1 APIN: Visualización dinámica de redes de interacción de proteínas en APID

APIN (*Agile Protein Interaction Network*) es una herramienta de visualización de redes, que ha sido diseñada con el objetivo de facilitar el acceso a la información que contiene la plataforma web *APID*, mediante la representación de redes de interacción de proteínas. La herramienta está conectada con el servidor de *APID* para hacer búsquedas en su base de datos y para obtener información descriptiva de las proteínas y las interacciones. Se puede ejecutar como una aplicación independiente o desde la página que describe a las interacciones en *APID*, por medio del botón "Graph". Además de representar redes de interacción, *APIN* también pretende hacer una interfaz amigable, en la que el usuario pueda cambiar parámetros de visualización de la red y hacer una exploración dinámica de los datos de interacción de proteínas.

El desarrollo de *APIN* se hizo en el lenguaje de programación Java [19], que tiene la ventaja de ser multiplataforma, de modo que la aplicación funcionará en los sistema operativos más usados sin necesidad de hacer ninguna adaptación. El único requerimiento necesario para que el usuario pueda ejecutar el programa es que tenga instalada una máquina virtual Java, que ya viene por defecto integrada en muchos sistemas operativos. Además los programas Java se pueden integrar fácilmente en una página Web mediante la tecnología *applet* [21], que permite la ejecución del programa en el navegador web del usuario. Otra de las ventajas que tiene Java es la gran cantidad de aplicaciones y librerías de código abierto que hay desarrolladas, que pueden ser reutilizadas para hacer nuevas aplicaciones libres. Este es el caso de *Touchgraph* [28], que es una herramienta de visualización de redes, que proporciona métodos de representación y una interfaz de distribución de nodos basada en fuerzas de atracción-repulsión. El código fuente de esta herramienta ha sido publicado para facilitar nuevos desarrollos y se ha usado como una librería de código en *APIN* para representar las interacciones proteína-proteína en forma de red.

La **figura 2.1** muestra un esquema de implemetación de *APIN*. El esquema tiene representados a los principales componentes de la aplicación. El *applet* está formado por la interfaz de usuario que permite interactuar con la aplicación, y que como se ha

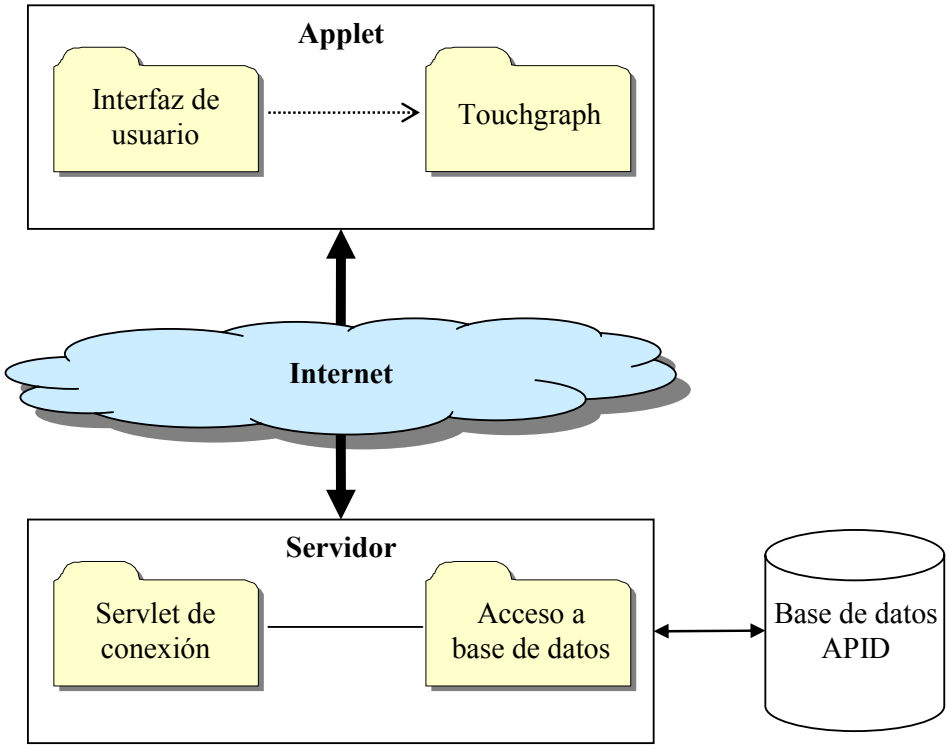


Figura 2.1: Esquema de implementación de *APIN*.
La aplicación está dividida en dos partes, el *applet* y el servidor. El *applet* se encarga de representar las redes de interacción y de ofrecer un interfaz gráfico para que el usuario pueda interactuar con la aplicación. Se constituye por un conjunto de objetos que forman la interfaz de usuario y que usan la librería *Touchgraph* para representar redes. El servidor tiene la función de comunicarse con los *applets*, para enviar los datos de interacción que se van a representar. Tiene dos partes, una que se encarga de hacer consultas a la base de datos de *APID* y otra que realiza la comunicación con los *applets*.

dicho usa la librería *Touchgraph* para hacer las representaciones gráficas. Por otro lado es necesaria la existencia de un servidor que consulte la información de la base de datos de interacción y se la envíe al *applet* para que éste la pueda representar. El servidor está formado por un componente que accede a la información de la base de datos con *JDBC* [24] y otro que maneja las conexiones que se establecen con los *applets*. Se diseñó de esta forma para evitar hacer una conexión directa al sistema gestor de base de datos *MySQL* [25], ya que este tipo de tráfico no suele estar permitido en muchas redes y los cortafuegos suelen vetarlo. Para conseguir que los usuarios puedan recibir información de la base de datos sin interrupciones, se usó el protocolo de transferencia de hipertexto (*HTTP*, *HyperText Transfer Protocol*), de modo que el *applet* hace una petición *HTTP* al servidor, pasándole unos parámetros de búsqueda, y el servidor contesta con los datos que tiene que mostrar el *applet* por medio de una respuesta *HTTP*. A esta estrategia de encapsular el tráfico de otros protocolos en el protocolo *HTTP*, se le denomina *HTTP tunneling*.

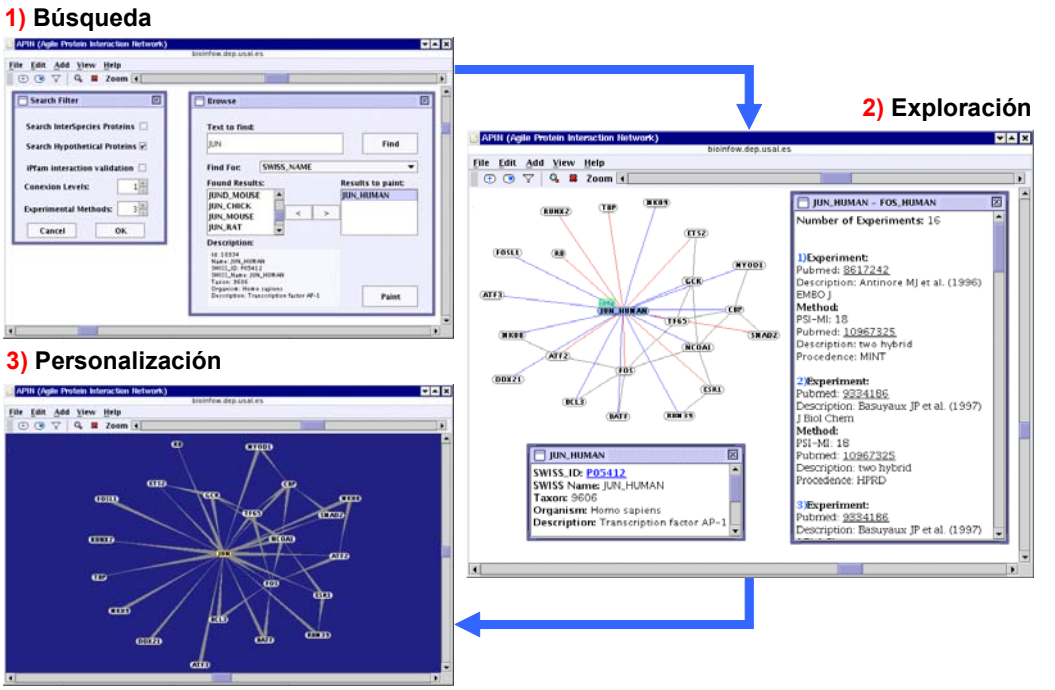


Figura 2.2: Ejemplo de uso de *APIN*.
Se representa un flujo de trabajo en APIN por medio de 3 viñetas: **1**) Búsqueda: muestra la interfaz de búsqueda de proteínas y los filtros que se pueden aplicar. **2**) Exploración: muestra la red resultante de representar las interacciónes de la proteína JUN_HUMAN y un ejemplo de la información que se puede consultar sobre las proteínas y las interacciones. **3**) Personalización: muestra un ejemplo de los posibles cambios que se pueden hacer mediante la modificación de parámetros de visualización.

El empleo de estas tecnologías ha permitido que *APIN* pueda representar las redes de interacción que contiene *APID*, de una forma sencilla, dinámica y transparente para el usuario. Además el usuario, por medio de una interfaz de búsqueda, puede añadir proteínas a la red y modificar parámetros de visualización como son el *zoom*, el tipo de etiqueta del nodo, la clase de vínculos entre nodos y el color de fondo. La **figura 2.2** muestra un ejemplo de uso de *APIN*. En primer lugar se ha hecho una búsqueda de la proteína **JUN_HUMAN** (**viñeta 1**), mediante la interfaz de búsqueda que tiene la aplicación. Esta interfaz permite buscar una proteína o una lista de proteínas de las que se quiere conocer sus interacciones. Además se observa, que se pueden fijar también filtros de búsqueda relativos: (**i**) número de niveles de conexión que se quieren explorar, (**ii**) número mínimo de métodos que validan las interacciones, (**iii**) existencia de dominios interactuantes que validan la interacción, (**iv**) inclusión o no de proteínas hipotéticas en la red y (**v**) inclusión o no de interacciones interespecie. En la segunda ventana se muestra la red resultante de la búsqueda realizada (**viñeta 2**). Sobre esta red se podrá consultar información referente a las proteínas pulsando el botón **+info** y descriptiva de las interacciones pulsando sobre ellas. Además haciendo "doble clic"

sobre una proteína, la aplicación añadirá a la red las nuevas interacciones en las que está implicada dicha proteína en la red. Por último, la aplicación permite cambiar los parámetros de visualización para hacer las representaciones a gusto del usuario (**viñeta 3**).

2.2.2 APID2NET: *Plugin* de análisis de interactomas en *Cytoscape*

En la introducción se habló de *Cytoscape* como una aplicación bioinformática que permite la visualización y el análisis de redes de interacción molecular, y que facilita la integración de *plugins* que le añaden funcionalidad. Cada vez está creciendo más el número de personas que usan *Cytoscape*, lo que hace que una aportación a esta plataforma tenga mucha difusión. También está creciendo la comunidad de desarrolladores de *plugins*, que está consiguiendo consolidar a *Cytoscape*, como una plataforma bioinformática flexible y abierta, con gran número de herramientas para el estudio de redes biológicas. Por todo ello, se puede decir que *Cytoscape* va camino de convertirse en una plataforma de referencia para distribuir métodos aplicables a redes biomoleculares.

Teniendo en cuenta las facilidades que da *Cytoscape* para programar y desarrollar herramientas, decidimos construir un *plugin* llamado **APID2NET**, que permite la exploración y el análisis de subconjuntos de los interactomas que tiene almacenado *APID*. Este *plugin* está completamente integrado en *Cytoscape*, aumentando y mejorando las prestaciones y posibilidades de análisis de redes que tiene *APIN*.

La implementación de la nueva herramienta llamada *APID2NET* ha sido similar a la de *APIN*. Ambos han sido desarrollados en Java y han usado un *HTTP tunneling* para comunicarse con el servidor de datos de *APID*. La mayor diferencia es que *APID2NET* se ha programado como un *plugin* que se integra dentro de *Cytoscape* y *APIN* es un *applet* que funciona de forma independiente. Para integrar *APID2NET* se ha usado la interfaz de creación de *plugins*, que permite llamar a métodos de *Cytoscape*, para representar redes de interacción y cambiar parámetros de visualización de la red.

Las herramientas principales que tiene *APID2NET* organizadas por menús de opciones son:

- '**APID retrieval**': Permite buscar una proteína o una lista de proteínas de las que se quiere conocer sus interacciones. La interfaz es similar a la que se desarrolló para

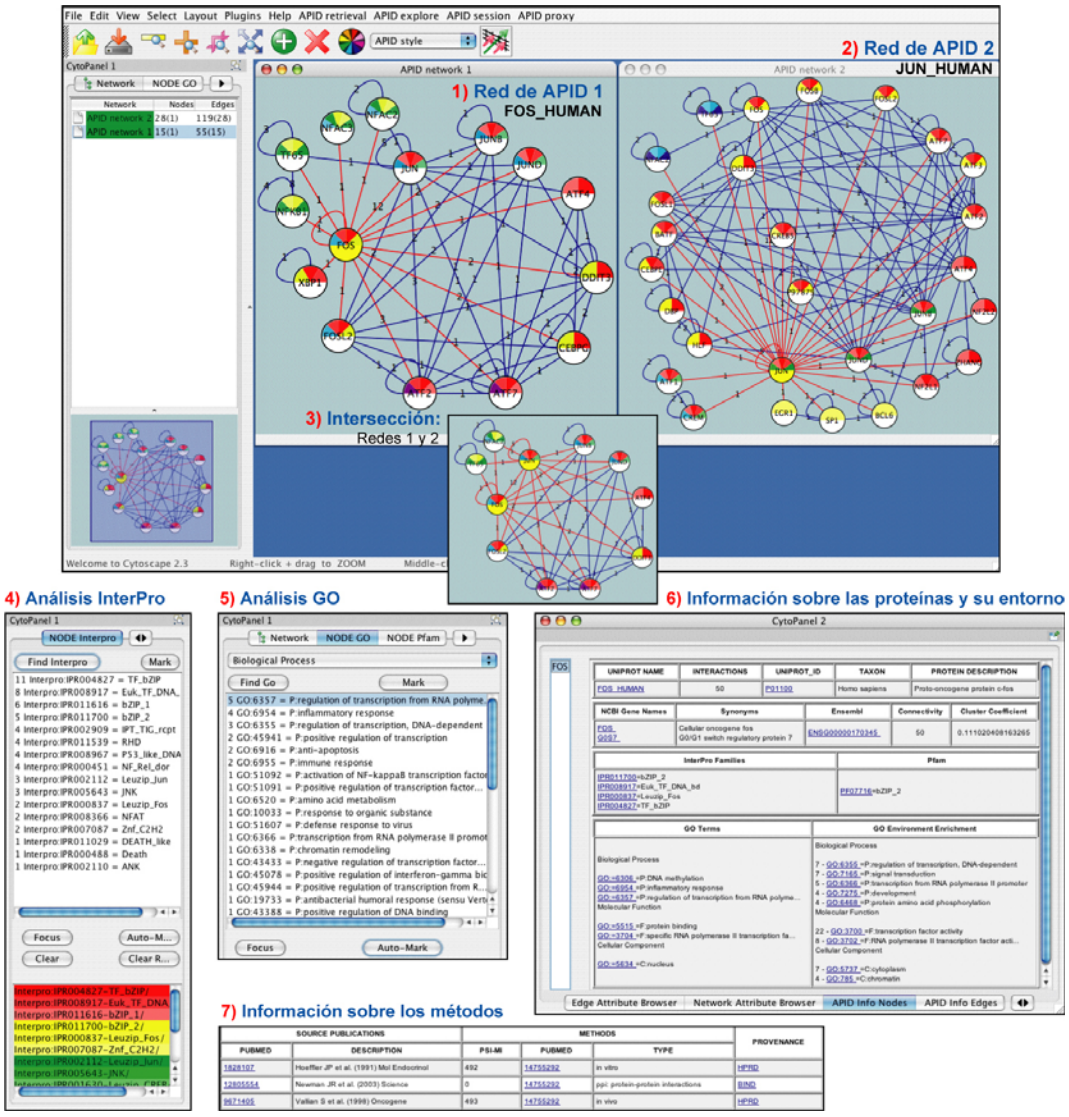


Figura 2.3: Ejemplo de uso de *APID2NET* dentro de *Cytoscape*.
Se muestran las posibilidades que tiene *APID2NET* para analizar redes de interacción. El diagrama esta dividido en 7 viñetas: **1** y **2**) Redes resultantes de representar las interacciones de **FOS_HUMAN** y **JUN_HUMAN**, validadas con dominios interactuantes. **3**) Resultado de hacer la intersección de las redes 1 y 2. **4** y **5**) Ventanas que permiten hacer análisis funcionales con *InterPro* y *GO*. **6** y **7**) Paneles que muestran información sobre las proteínas y las interacciones de la red.

APIN y permite fijar los mismos parámetros de búsqueda comentados en el apartado anterior. Además también se pueden buscar dinámicamente las interacciones de una proteína con la opción del menú conceptual de "expandir red".

- '**APID explore**': Permite analizar la información funcional de las proteínas presentes en la red y los métodos experimentales que se han usado para validar las interacciones (en el 'CytoPanel-1'). El *plugin* permite encontrar y resaltar proteínas de la red en función de su anotación a términos *GO*, a dominios *Pfam* y a dominios y motivos *InterPro*. La aplicación permite además colorear los nodos en función

de la anotación que tengan las proteínas que representan. Las interacciones también se pueden resaltar en función del método experimental que las ha validado. Además (en el 'CytoPanel-2) se puede consultar toda la información sobre las proteínas e interacciones de la red que es traída desde *APID* a *Cytoscape*.

- '**APID session**': Permite al usuario guardar su sesión de trabajo en el propio ordenador, para que pueda ser recuperada posteriormente con todos los cambios o análisis que haya realizado.

Un ejemplo de uso de *APID2NET* está representado en la **figura 2.3**, en ella se muestran las redes resultantes de buscar las interacciones de las proteínas **FOS_HUMAN** y **JUN_HUMAN**, seleccionando la opción de buscar solo las interacciones con validación estructural (**viñetas 1** y **2**). Con estas redes se ha hecho su intersección para ver las interacciones que tienen en común (**viñeta 3**). Las redes están coloreadas en función de los dominios y motivos *Interpro* que tienen las proteínas, esto se hace mediante la opción de análisis *Interpro* "Node Interpro" (**viñeta 4**). También se pueden marcar los nodos en función de la anotación *GO* (**viñeta 5**) y *Pfam* de las proteínas. Por último están las ventanas que muestran información sobre las proteínas y su entorno (**viñeta 6**) y sobre los métodos que validan una interacción (**viñeta 7**).

En la actualidad *APID2NET* es uno de los *plugins* más usados de la plataforma *Cytoscape*. Prueba de ello son las estadísticas sobre el número de descargas de *plugins* que tiene *Cytoscape* en su página web [29]. La **tabla 2.1** muestra un resumen de estas estadísticas (tomadas el 29 diciembre de 2008), en ella se observa que *APID2NET* es el quinto *plugin* más bajado en los últimos 30 días. Además hay que tener en cuenta que el *plugin* se puede bajar también en la página web de *APID*, lo que hace que el número de nuevos usuarios mensuales sea mayor al especificado en la tabla 2.1. Por ejemplo en noviembre de 2008 se hicieron 49 descargas de *APID2NET* desde la página web de *APID*, que habria que añadir al número de instalaciones que se han realizado desde la plataforma Cytoscape.

	Plugin	Instalaciones Totales	Instalaciones los últimos 30 días
1	**BiNGO**	10044	772
2	**jActiveModules**	4115	291
3	**NCBIEntrezGeneUserInterface**	2802	260
4	**AgilentLiteratureSearch**	4404	247
5	**APID2NET**	**2684**	**168**
6	**MCODE**	4266	167
7	**MiMIplugin**	3036	167
8	**GPML-Plugin**	1480	149
9	**CentiScaPe**	473	139
10	**clusterMaker**	548	133
11	**NCBIClient**	297	123
12	**IntActWSClient**	309	119
13	**BioNetBuilder**	673	109
14	**BiomartClient**	250	102
15	**structureViz**	1592	100
16	**dynamicXpr**	1033	100
17	**MetaNodePlugin2**	1835	96
18	**ExpressionCorrelation**	391	95
19	**VistaClaraPlugin**	1118	91
20	**NamedSelection**	1555	88

Tabla 2.1: Estadísticas de instalación de *plugins* en *Cytoscape*
Número de instalaciones totales y de instalaciones realizadas entre el 29 de Noviembre de 2008 y el 29 de Diciembre de 2008, de los 20 *plugins* más instalados en ese intervalo de tiempo.

2.3 CONCLUSIONES

El empleo de **técnicas computacionales de visualización** facilita el análisis y la exploración de redes biomoleculares complejas. De este modo, para el estudio de redes de interacción proteína-proteína vimos necesario construir herramientas bioinformáticas que puedan representar datos de interacción de forma flexible y sencilla para el usuario. Así, se desarrolló y construyó *APIN* que consigue representar de forma gráfica e interactiva las redes de interacciones provenientes de la base de datos de *APID*, y permite personalizar parámetros de visualización de la red, así como hacer una navegación dinámica por los datos de interacción entre proteínas.

Una de las plataformas bioinformáticas internacionales más importantes de representación de redes biomoleculares es *Cytoscape*. Esta plataforma permite la integración de *plugins* para aumentar sus funcionalidades. Basados en este marco de desarrollo internacional y abierto, hemos construido y puesto a punto un nuevo *plugin* para *Cytoscape* llamado *APID2NET* que permite importar los datos de *APID* en *Cytoscape* y hacer análisis funcionales sobre las redes de interacción representadas. Estos análisis incluyen exploración automática sobre las anotaciones de las proteínas como funciones asociadas a *GO*, dominios estructurales derivados de *InterPro*, palabras clave de *UniProt*.

Capítulo 3

Mapas ómicos de coexpresión de genes humanos

3.1 INTRODUCCIÓN

3.1.1 Microarrays de oligonucleótidos de alta densidad para medir expresión génica

Un microarreglo de expresión génica, mas conocido como *microarray*, consiste en un gran número de moléculas de *DNA* ordenadas, que forman una matriz de secuencias, sobre un sustrato sólido. Estos fragmentos de material genético suelen ser secuencias cortas de *cDNA* (*DNA* complementario, sintetizado a partir de *mRNA*), llamadas oligonucleótidos, que han sido inmovilizadas sobre un soporte. A los oligonucleótidos

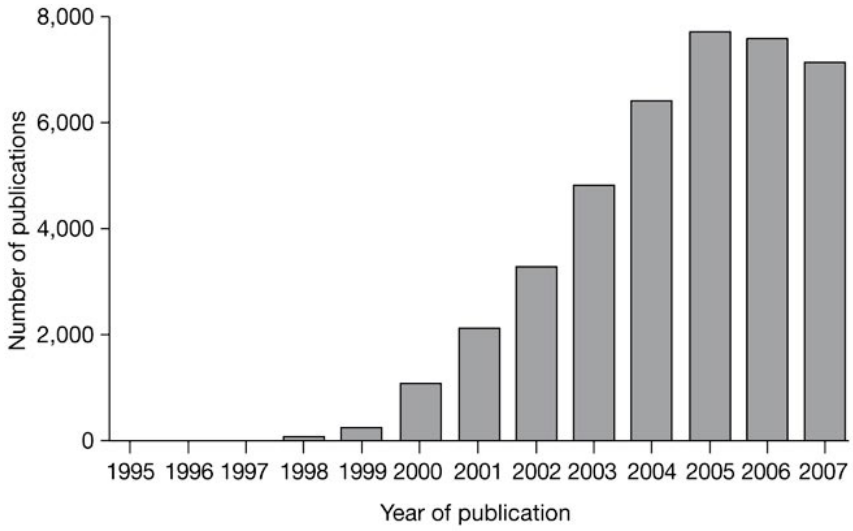


Figura 3.1: Número de publicaciones al año que han utilizado *microarrays*.
Estimación anual del número de artículos científicos publicados que mencionan el uso de *microarrays* como técnica experimental (fuente de Borst et al. (de Borst, *et al.*, 2008)).

inmovilizados se les denomina "sondas" o *probes*. Las muestras que se analizan son fragmentos de *cDNA* o *mRNA* que se marcan (normalmente con etiquetas fluorescentes) y se incuban sobre el *microarray* de sondas, permitiendo la hibridación, es decir el reconocimiento y unión entre secuencias complementarias. Durante la hibridación, las muestras de oligos marcados de *cDNA*, se unen a sus complementarias inmovilizadas en el soporte del chip, permitiendo su identificación y cuantificación. El chip se lee por medio de un escáner que obtiene los valores de intensidad de fluorescencia para cada sonda. Estos datos de intensidad son interpretados y analizados por medio de herramientas bioinformáticas [30].

Los *microarrays* de expresión son una de las técnicas genómicas que más popularidad ha ganado en los últimos años. La **figura 3.1** muestra una estimación anual del número de artículos científicos publicados que mencionan el uso de *microarrays* como técnica experimental. Este número ha tenido un crecimiento exponencial de 1995 a 2005, llegando a un valor cercano a 8000 publicaciones anuales donde parece que se está estabilizando.

Los orígenes de la tecnología de *microarrays* se pueden encontrar en los experimentos llamados "*dot blot*", como *Southern* (Southern, 1975) *y Northern* (Alwine, *et al.*, 1977) *blot*, que aparecieron en bioquímica en los años 70, donde el *DNA* se inmovilizaba sobre membranas y se marcaba habitualmente con una sonda radioactiva (Kafatos, *et al.*, 1979; Saiki, *et al.*, 1989). El desarrollo de métodos de fluorescencia al final de los años 80 (Kaiser, *et al.*, 1989) y de los soportes de vidrio al principio de los 90 (Maskos y Southern, 1992; Maskos y Southern, 1992) fueron avances importantes

para el desarrollo de los modernos *microarrays*. Sin embargo, el mayor cambio llegó cuando se empezó a aplicar la fotolitografía (Fodor, *et al.*, 1991) y la impresión (Schena, *et al.*, 1995), que eran métodos que se habían usado en la industria de semiconductores. Estos métodos junto a la miniaturización y la automatización de procesos, se aplicaron posteriormente en la fabricación de *microarrays*. Los *microarrays* modernos se puede decir que empezaron en 1995, cuando Schena hizo el primer *microarray* con 45 sondas de *cDNA* (Schena, *et al.*, 1995). El progreso tecnológico de los *microarrays* de *cDNA* ha sido muy rápido, en 1996 ya se hicieron publicaciones con 1000 sondas, mientras que al mismo tiempo ya había una compañía pionera en este tipo de técnicas, Affymetrix, que desarrolló una novedosa tecnología de síntesis *in situ* de *arrays* de oligonucleótidos basada en fotolitografía combinada con química de síntesis de *DNA* (Fodor, *et al.*, 1993; Fodor, *et al.*, 1991; Pease, *et al.*, 1994). El uso de síntesis dirigida por luz, para fijar nucleótidos sintetizados a la superficie del chip, permitió la fabricación de los *microarrays* de oligonucleotidos de alta densidad, los cuales, en 1996 ya contenían unas 135.000 sondas.

Actualmente el *GeneChip* de *Affymetrix* es capaz de tener millones de sondas en una superficie de vidrio de 1.28cm^2. El rápido crecimiento de las tecnologías de *microarrays* está muy ligado al desarrollo general que ha habido en métodos de secuenciación automática y en bases de datos biológicas públicas, que han incluido y anotado secuencias de genes y de genomas completos, lo cual ha permitido el diseño y la selección de sondas específicas para la detección de genes.

El proceso de fabricación del *GeneChip* de *Affymetrix* esta representado en la **figura 3.2**. Un *microarray* se construye a partir de un sustrato de vidrio al que se han

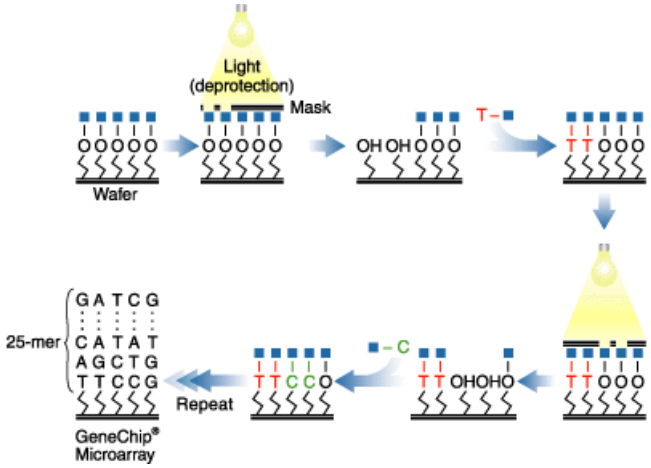


Figura 3.2: Proceso de fabricación del *GeneChip* de *Affymetrix*.
Esquema del proceso de fabricación del *microarray* de nucleótidos de alta densidad *GeneChip* de *Affymetrix* utilizando la técnica de fotolitografía (fuente *Affymetrix* [31]).

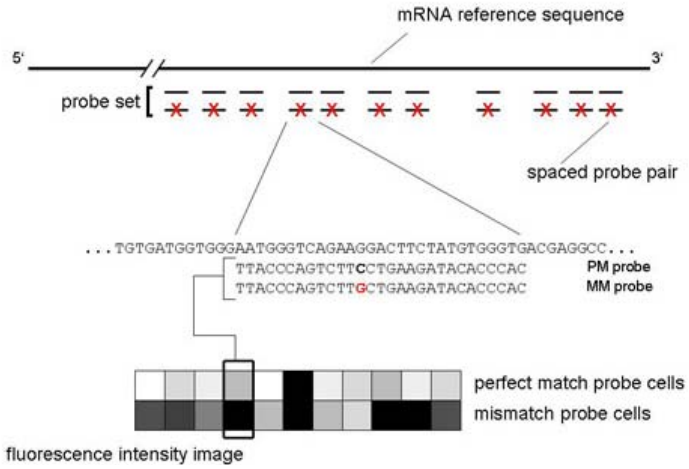


Figura 3.3: Proceso de fabricación del *GeneChip* de *Affymetrix*.
Esquema del proceso de fabricación del *microarray* de nucleótidos de alta densidad *GeneChip* de *Affymetrix* utilizando la técnica de fotolitografía (fuente *Affymetrix*).

adherido moléculas con un grupo eliminable fotoquímicamente. En determinadas áreas de la superficie del *microarray* se hace incidir luz a través de una máscara fotolitográfica, para producir un efecto de fotodesprotección localizada. En ese momento se añade un tipo de nucleótido (A,T,C o G) para que se produzca un acoplamiento químico y por último se hace un lavado (Lipshutz, *et al.*, 1999) [31]. Este proceso se hace 25 veces por cada tipo de nucleótido de modo que cada oligo estará compuesto por 25 bases.

Los oligos se agrupan en conjuntos de sondas correspondientes a un gen (denominados *probesets*) que están formados por entre 11 y 20 parejas de oligos que se corresponden con distintas regiones codificantes de dicho gen. Son parejas porque una de las sondas de la pareja es de homología perfecta (PM, *PerfectMatch*) y la otra tiene un error deliberado o mutación (MM, *MisMatch*) en el nucleótido 13 (ver **figura 3.3**). La función de las sondas *MM* es la de actuar como un control no específico que permite estimar las señales de ruido de fondo y de hibridación cruzada, diferenciando las señales obtenidas por hibridación del gen deseado de otras no específicas. El número de *probesets* depende del tipo de *GeneChip*. Actualmente oscila entre los 1031 *probesets* que tiene RT_U34 a los 61359 *probesets* de U133_X3P.

Una visión general del método experimental está representada en la **figura 3.4**. La técnica parte de los m*RNA* extraídos de las células que se van a estudiar, que se transforman mediante retro-trascripción *in vitro* a *cDNA*. Los *cDNA* se marcan con compuestos fluorescentes mediante transcripción *in vitro* a *cRNA*, se fragmentan y se

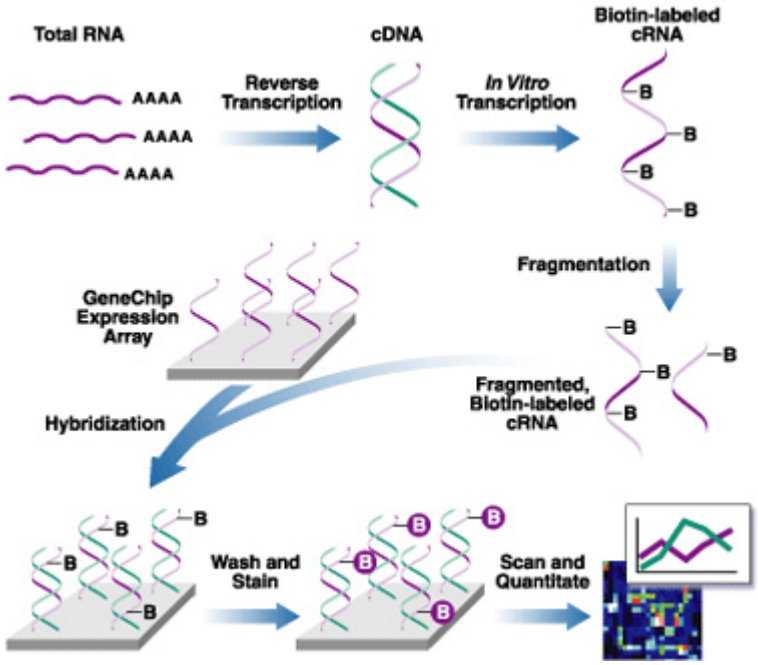


Figura 3.4: Protocolo de hibridación de *GeneChips*.
Protocolo de hibridación de los microarrays de oligonucleótidos *GeneChip* de *Affymetrix* (fuente Affymetrix).

ponen en el *microarray* para comenzar el proceso de hibridación. Este proceso consiste en la unión de los *cDNA* de las muestras, que han sido aislados y marcados, con los oligos de *cDNA* presentes en el *microarray*.

Después de la hibridación se hacen lavados para eliminar todo el *cRNA* que no ha podido hibridar y se procede a obtener la imagen con los niveles de intensidad, para ello se usa un escáner óptico láser que excita la superficie del *microarray* para generar una emisión fluorescente que es recogida por un detector de alta resolución. El resultado es una imagen bidimensional en la que es posible cuantificar la intensidad del proceso de hibridación. A continuación se usa un programa de análisis de imagen para calcular la intensidad de cada sonda del *microarray* y guardar dichas intensidades numéricamente en un fichero de texto.

Los valores numéricos de intensidad obtenidos, deben de ser ajustados para poder ser analizados. Esto se hace con algoritmos de cálculo de señal que suelen realizar 3 pasos diferenciados: **1)** corrección del ruido de fondo (*background*) **2)** normalización (inter- e intra-*microarray*) **3)** sumarización de la señal de las sondas para calcular la intensidad de un conjunto de sondas o *probesets*.

Cuando se trabaja en experimentos que manejan varios *microarrays*, existe una variación inter-chip que debe ser normalizada. Esta variación puede ser debida a

diferencias en la respuesta del escáner, a una diversidad en las muestras, al uso de cantidades diferentes de *RNA* inicial o a otras circunstancias que se hayan dado al hacer el ensayo. La normalización es un proceso que trata de compensar esta variación debida a efectos de la técnica o de los aparatos. La idea principal en la que se basan los métodos de normalización es que cuando se comparan dos o más muestras a nivel global, éstas deben seguir distribuciones similares en las que la mayoría de los *probesets* tienen poca señal (es decir la mayoría de los genes no se expresan o no se alteran en una condición estudiada).

Se han desarrollado numerosos algoritmos para normalizar datos de *microarrays* pero no existe ningún estándar de uso. En general corrigen la variación no biológica de cada *microarray* por separado (intra-chip) y entre distintos *microarrays* (inter-chip), estabilizando la varianza y haciendo más similares las distribuciones. Además en la normalización de *GeneChips* de *Affymetrix* se incluyen sondas control en el chip, que son genes que normalmente se expresan o secuencias de otros organismos que se añaden a la muestra en concentraciones conocidas. Además de mejorar la precisión de los datos, la normalización trata de optimizar la robustez en el cálculo de la señal, que en el caso de la plataforma *Affymetrix*, se hace por medio de la sumarización de las sondas que pertenecen a un *probeset*.

Existen varios métodos de normalización para los *GeneChips* de *Affymetrix*, estos tratan de medir los valores de expresión de los *probesets* en una muestra a partir de los valores de intensidad de las sondas. Los métodos más populares son *RMA* (Irizarry, *et al.*, 2003), *gcRMA* (Wu y Irizarry, 2005), *dChip* (Li y Wong, 2001) y el desarrollado inicialmente por *Affymetrix MAS5* (Liu, *et al.*, 2002). Los valores obtenidos por estos métodos difieren considerablemente y por tanto condicionan a los resultados obtenidos en análisis posteriores. Los dos métodos más usados son sin duda el original *MAS5* y *RMA*.

MAS5 emplea tanto a los *PM* como a los *MM* para hacer la normalización. Por otro lado, en el primer paso de corrección de *background*, calcula el ruido de fondo para cada sonda del *microarray* basado en el valor del 2% de las sondas con menor intensidad, que es ponderado por la distancia a la sonda central para la que se calcula el ruido promedio. Después de restar el ruido de fondo a cada sonda se hace la normalización intra-chip; para ello se usan las sondas *MM*, que en la mayoría de los casos (en los que *PM* > *MM*), se hace restando al *PM* el valor del *MM*. En los casos en

que PM ≤ MM, se ajusta el valor del *MM* antes de restarlo al *PM*. La normalización inter-chip se hace con un simple escalado de los valores de expresión para eliminar la variación entre *microarrays*. Por último, se hace el cálculo de la señal de los *probesets* mediante una media biponderada de *Tukey* (*Tukey biweight*), del valor corregido de las sondas en escala logarítmica en base 2.

RMA es un método que usa solo las sondas *PM* y basa la corrección de *background* en un ajuste a un modelo lineal de señal y ruido discriminables. Después, para la normalización intra e inter-chip escala todos los *microarrays* a una misma distribución empírica de cuantiles a nivel de sonda (Bolstad, *et al.*, 2003). Finalmente calcula el valor de expresión de los *probesets* en escala logarítmica sumarizando las sondas mediante una mediana pulida (*median polish*) (Tukey, 1977).

3.1.2 Ingeniería reversa a datos de expresión génica

La ingeniería reversa es el proceso de analizar un sistema existente, identificando los componentes del sistema, sus funciones (roles) y las relaciones entre ellos. En general trata de descubrir el proceso de construcción de un producto a partir del propio producto. Esta técnica ha sido aplicada en áreas de conocimiento tan variadas como son la informática, la ingeniería industrial y la economía.

En el área de la bioinformática, en los últimos años se está empezando a trabajar con información proveniente de *microarrays* de expresión genómica para hacer ingeniería reversa. Esta información permite hacer un modelo de los procesos celulares subyacentes a la expresión génica global y a los sistemas de regulación de la transcripción de los genes expresados. Este proceso está reflejado en la **figura 3.5** en la que se muestra de modo esquematizado una estrategia general para modelar un sistema biológico con ingeniería reversa aplicada a *microarrays* de expresión.

Los principales retos de esta estrategia parten de los problemas subyacentes a los datos usados: **(i)** existencia de muchos falsos positivos (ruido); **(ii)** gran dimensionalidad; **(iii)** existencia de pocas muestras en comparación al número de genes. Además de esto hay una ausencia de métodos claros para validar el rendimiento de los algoritmos. Por tanto, muchas preguntas quedan en el aire cuando aplicamos ingeniería reversa a datos de *microarrays*, lo cual hace plantearse la fiabilidad de las redes predichas y la utilidad de determinadas aproximaciones a estudios concretos.

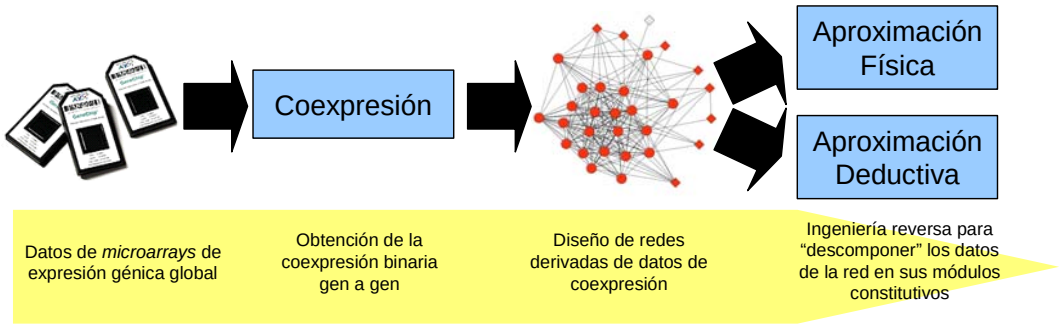


Figura 3.5: Ingeniería reversa aplicada a datos de *microarrays*.
Esquema de aplicación de las aproximaciones de ingeniería reversa con *microarrays* de expresión buscando perfiles de expresión correlacionados.

Los algoritmos de ingeniería reversa se pueden clasificar en dos aproximaciones: las llamadas **físicas**, que tratan de identificar las moléculas que controlan la síntesis de *RNA*; y las llamadas **deductivas**, que buscan un modelo causal de relaciones entre transcritos de *RNA*. Es decir, en el caso de las deductivas no están solo restringidas a describir las relaciones entre factores de trascripción y *DNA* (Gardner y Faith, 2005).

A lo largo de este capítulo abordamos las dos aproximaciones: en primer lugar, se crea con un método nuevo, una red de coexpresión de genes a partir de un *set* heterogéneo de *microarrays* de tejidos sanos; a continuación este *set* se usa para definir relaciones funcionales entre genes y se combina con datos de secuencia genómica para identificar factores de transcripción comunes a grupos de genes muy conectados en dicha red (**Figura 3.5**).

3.1.3 Redes transcripcionales derivadas de datos de coexpresión

Una de las estratégias más comunes en ingeniería reversa a la hora de inferir redes transcripcionales es la búsqueda de coexpresión en datos de *microarrays* de expresión. Se dice que una pareja de genes coexpresa cuando tienen un perfil de expresión correlacionado, es decir, que los valores de expresión siguen la misma tendencia de subida y bajada en las muestras para las que se calcula (**figura 3.6**). A la hora de buscar perfiles de expresión correlacionados, se utilizan indicadores estadísticos que miden el grado de correlación entre dos variables [32]. La medida de la correlación nos permite construir redes de genes, ya que permiten establecer relaciones binarias entre pares de genes, si los valores expresión de estos genes tienen un valor de correlación significativo. Las redes de genes generadas a partir de perfiles de expresión correlacionados, son útiles para desarrollar las aproximaciones de la ingeniería reversa explicadas, de modo que se han creado diversos algoritmos

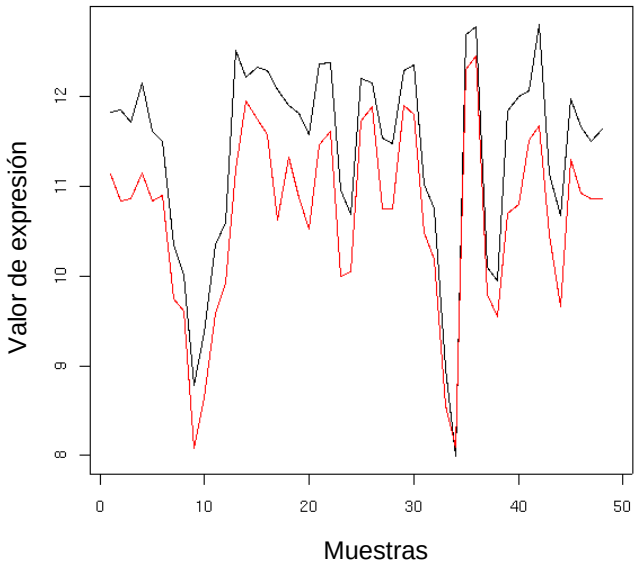


Figura 3.6: Perfiles de expresión de una pareja de genes coexpresados.
Muestra un ejemplo de una pareja de *probesets* que tiene una relación de coexpresión alta. Como se puede apreciar en la figura, los valores de expresión siguen la misma tendencia de subida-bajada a lo largo de las 48 muestras.

(principalmente basados en teoría de grafos o *clustering*), para extraer grupos de genes relacionados y descubrir información transcripcional o funcional a partir de la red.

Además de su aplicación en ingeniería reversa, la coexpresión se ha usado satisfactoriamente en diversos estudios para aportar información útil a la hora de, por ejemplo, definir y relacionar procesos biológicos (Segal, *et al.*, 2003), validar interacciones entre proteínas (von Mering, *et al.*, 2007) o definir la filogenia de un gen (Tirosh, *et al.*, 2006). La gran diversidad de aplicaciones en estudios bioinformáticos de los datos de coexpresión, es debida a varios factores: (**i**) a la información funcional que tiene asociada (Lee, *et al.*, 2004), (**ii**) a la conservación que tienen los clusters de coexpresión en la evolución (Tirosh, *et al.*, 2006), (**iii**) a la correlación de estos clusters con procesos biomoleculares o rutas metabólicas (Magwene y Kim, 2004). Todas éstas características y aplicaciones de los datos de coexpresión hacen que los métodos de búsqueda de coexpresión tengan gran interés en el ámbito de la bioinformática.

3.1.4 Estrategias de búsqueda de correlación en perfiles de expresión

Tanto las aproximaciones de ingeniería reversa como los estudios bioinformáticos que han empleado datos de coexpresión, han tenido que hacer una serie de pasos comunes para obtener unos datos de coexpresión adecuados a sus fines. La mayoría de los estudios han seguido una estrategia simple, que está esquematizada en la **figura 3.7**. El

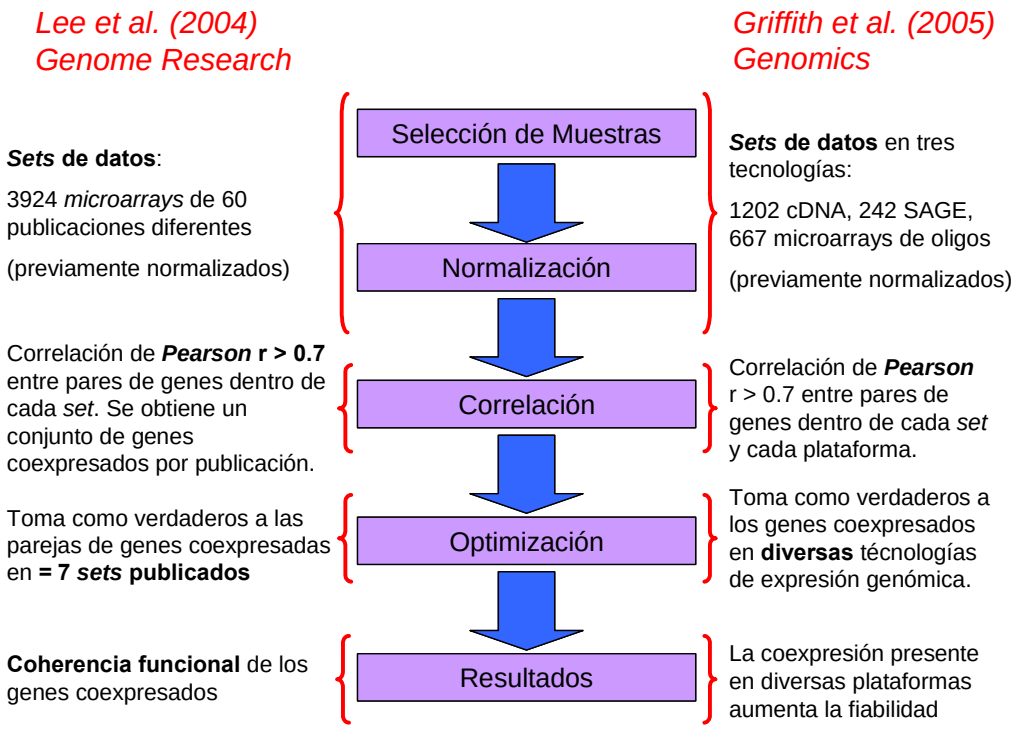


Figura 3.7: Estrategias de búsqueda de perfiles relacionados.
En el flujograma se representan los pasos básicos que se realizan para buscar coexpresión. A los lados están descritas la metodologías usadas por *Lee et al* (izquierda) y *Griffith et al* (derecha), para construir *coexpresomas* fiables.

primer paso como es lógico, es la selección de muestras. Estas pueden provenir de bases de datos de *microarrays* o pueden haber sido generadas específicamente para el estudio propuesto. Este primer paso es muy importante, ya que de él dependen en gran medida los resultados obtenidos. Se debe realizar con la precaución de no incluir muestras que generen datos de coexpresión sesgados hacia resultados no acordes con el estudio propuesto. Una vez seleccionados los *microarrays* de expresión se suele hacer una normalización y cálculo de la señal de expresión. La elección del método de normalización y cálculo de la señal es también muy importante, ya que los resultados son muy variables en función de el método elegido. Sin embargo, según hemos podido comprobar, no se tiene muy en cuenta en la mayoría de los trabajos de investigación, pese a conocerse los beneficios y las vulnerabilidades de los distintos métodos (Lim, *et al.*, 2007).

Tras el cálculo de señal por gen (o por probeset correspondiente a cada gen) se obtienen los datos de expresión preparados para hacer el cálculo de la correlación entre genes; para ello se pueden aplicar varias métricas que miden el grado de similitud entre dos variables. Las más habituales en los estudios de coexpresión son *Pearson* y *Spearman*; sin embargo hay estudios que proponen otros métodos de

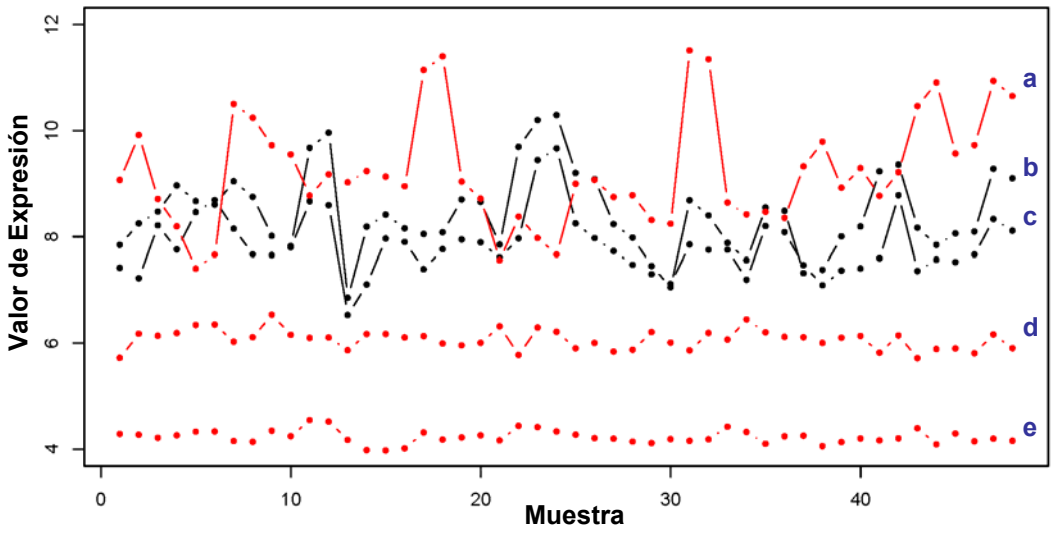


Figura 3.8: Ejemplo de ruido tecnológico en los *microarrays*.
Se representan los valores de expresión de los *probesets* correspondientes a dos genes: uno en negro con 2 *probesets* (b y c) y otro en rojo con 3 *probesets* (a,d y e). Se observa una correlación muy alta entre dos de los *probesets* en negro, mientras que el resto de *probesets* en rojo muestran correlaciones bajas o nulas entre ellos.

búsqueda de coexpresión basados en teoría de la información, en aprendizaje automático, etc. La elección de la métrica influye mucho en los resultados ya que es importante a la hora de establecer un criterio que discierna lo mejor posible a los verdaderos positivos de los falsos positivos y así poder optimizar la precisión del método sin perder mucha cobertura.

El cálculo de la correlación entre parejas de genes permite construir una red de coexpresión, en la que las interacciones se deducen de correlaciones significativas. Sobre esta red se pueden aplicar técnicas basadas en teoría de redes y algoritmos de agrupamiento (clustering) para buscar grupos de genes muy conectados en la red de coexpresión. La generación de estos grupos permite también investigar el entorno de cada gen en la red de coexpresión, para determinar su participación en procesos biológicos y para definir nuevas relaciones funcionales entre conjuntos de genes.

La estrategia general que se ha descrito, tiene varios puntos débiles que es importante tener en cuenta. En primer lugar, los datos de *microarrays* de expresión suelen ser ruidosos y de gran dimensionalidad, de forma que tenemos un número de variables (genes) muy alto en relación al número de muestras (*microarrays*). En la **figura 3.8** hay un ejemplo de ruido tecnológico en datos de *microarrays*. En ella están representados los valores de expresión de los *probesets* que mapean el gen *STOM* (según los ficheros de *Affymetrix*) y se observa que tienen una correlación muy alta entre dos de sus *probesets* (*201060_x_at* **(b)**, *201061_s_at* **(c)**), mientras que el resto

de *probesets* asociados a ese gen muestran correlaciones bajas o nulas entre ellos (201062_at **(e)**, 210824_at **(d)**, 210825_s_at **(a)**). Además de las características de los datos con los que trabaja, la estrategia general de búsqueda de coexpresión no es muy robusta ya que los resultados son muy variables en función del método de correlación y de los parámetros de optimización elegidos. Todas estas dificultades hacen que la fiabilidad de la estrategia general sea baja, esto es, se dan muchos falsos positivos en los resultados. Finalmente, la ausencia de *sets* de referencia para validar los resultados y métodos es otro problema añadido.

Estas dificultades se han intentado superar en diversas investigaciones diseñando estrategias o protocolos robustos de búsqueda de coexpresión, que se han aplicado a grandes conjuntos heterogéneos de *microarrays* de expresión. Una de las aproximaciones más referenciadas es la propuesta por *Lee et al.* (Lee, *et al.*, 2004) resumida en la parte izquierda de la **figura 3.7**. Esta estrategia parte de 60 conjuntos de *microarrays* de expresión ya normalizados, provenientes de distintas publicaciones científicas, que tienen un total de 3924 *microarrays*. Sobre estos datos calcula la correlación de *Pearson* (**r**) de cada par de genes para cada *set* de *microarrays*, de forma que una correlación es considerada positiva si **r>|0.7|**, obteniendo un conjunto de genes coexpresados por cada *set* de *microarrays*. A la hora de identificar a los verdaderos positivos se basa en el número de conjuntos de *microarrays* en los que una pareja de genes ha tenido una coexpresión positiva. De esta forma, define una red de coexpresión fiable de parejas de genes que han coexpresado en al menos 7 *sets* o conjuntos independientes de *microarrays*. Finalmente observan una coherencia funcional en módulos muy conectados de la red.

A la derecha de la **figura 3.7** está descrita la estrategia desarrollada por *Griffith et al.* (Griffith, *et al.*, 2005) que usa 2111 *microarrays* utilizando tres tecnologías diferentes (1202 *cDNA*, 242 SAGE y 667 *microarrays* de oligos) para calcular los coeficientes de correlación. Este trabajo observó que las redes de coexpresión deducidas por diferentes tecnologías son muy heterogéneas y que su intersección produce una ganancia en fiabilidad.

Teniendo en cuenta las principales dificultades que presenta la búsqueda de coexpresión y la gran utilidad de estos datos, es interesante desarrollar una estrategia que optimice la fiabilidad y la robustez de los resultados, abordando los puntos críticos

que se han definido anteriormente, para obtener una mejora respecto a los métodos publicados. A lo largo de este capítulo se describe el método de búsqueda de coexpresión que se ha desarrollado. El método se aplicó sobre un *set* heterogéneo de *microarrays* de tejidos humanos sanos, con el objetivo de obtener una red de coexpresión que muestre las funciones básicas de la célula. Los *microarrays* de expresión fueron seleccionados cuidadosamente para evitar, en la medida de lo posible, el ruido tecnológico y biológico introducido por las muestras. Además el número de muestras empleado fue moderado para poder extender la aplicabilidad del método a otros estudios. La robustez y fiabilidad del método se optimizó basándose en una validación cruzada frente al azar de las relaciones de coexpresión, tomando como *set* de referencia aquellos pares de genes sobre los que se tiene anotación funcional común.

3.2 MÉTODOS Y RESULTADOS

3.2.1 Importancia de la selección de muestras en la inferencia de coexpresión

De todas las aproximaciones de búsqueda de correlación en perfiles de expresión hechas a nivel global, son pocas las que se han hecho con muestras humanas y, cuando esto se ha hecho, se han empleado conjuntos de datos muy heterogéneos que mezclan muestras de personas sanas con muestras de pacientes con alguna patología. Este es el caso de varios de los estudios citados anteriormente (Griffith, *et al.*, 2005; Lee, *et al.*, 2004). La inclusión de datos provenientes de enfermedades (principalmente cáncer) en estos meta-análisis, puede producir un incremento del ruido, así como un sesgo de los resultados hacia funciones celulares afectadas en esas enfermedades. De hecho, se ha comprobado que las células cancerígenas tienen fuertes alteraciones en el genoma que se ven también reflejadas en los datos de expresión (Rhodes, *et al.*, 2002). Es por ello que estos estudios tienen dificultades cuando tratan de inferir el funcionamiento normal del sistema celular humano y son problemáticos a la hora de representar un mapa de coexpresión global fiable.

Considerando la importancia que tiene la selección de las muestras al hacer un análisis de perfiles de expresión, se ha puesto especial cuidado al seleccionar los *microarrays* utilizados para la construcción de redes de coexpresión. Estos *microarrays* se han elegido de un conjunto de *Gene Expression Atlas* (GEO GSE1133) (Su, *et al.*, 2004), formado por 136 *microarrays* de la plataforma HGU133a de *Affymetrix*. Estos *microarrays* tienen 22283 *probesets* que se corresponden con unos 15000 genes humanos. Las muestras con las que se hibridaron estos *microarrays* provienen de 68 tipos de tejidos humanos. De estas 68 muestras se eligieron solo las relativas a órganos, glándulas y tejidos que representaran partes principales del cuerpo humano, evitando las muestras de tipos celulares muy específicos. En total se seleccionaron 24 muestras que provenían de 3 personas, dos hombres y una mujer o dos mujeres y un hombre para las muestras no relacionadas con el sexo, 3 hombres para las muestras provenientes de testículo y próstata y 3 mujeres para las muestras de ovario y útero. Además de cada muestra se usan dos réplicas biológicas, dando como resultado la elección de un conjunto de 48 *microarrays* de 24 partes del cuerpo humano: glándula

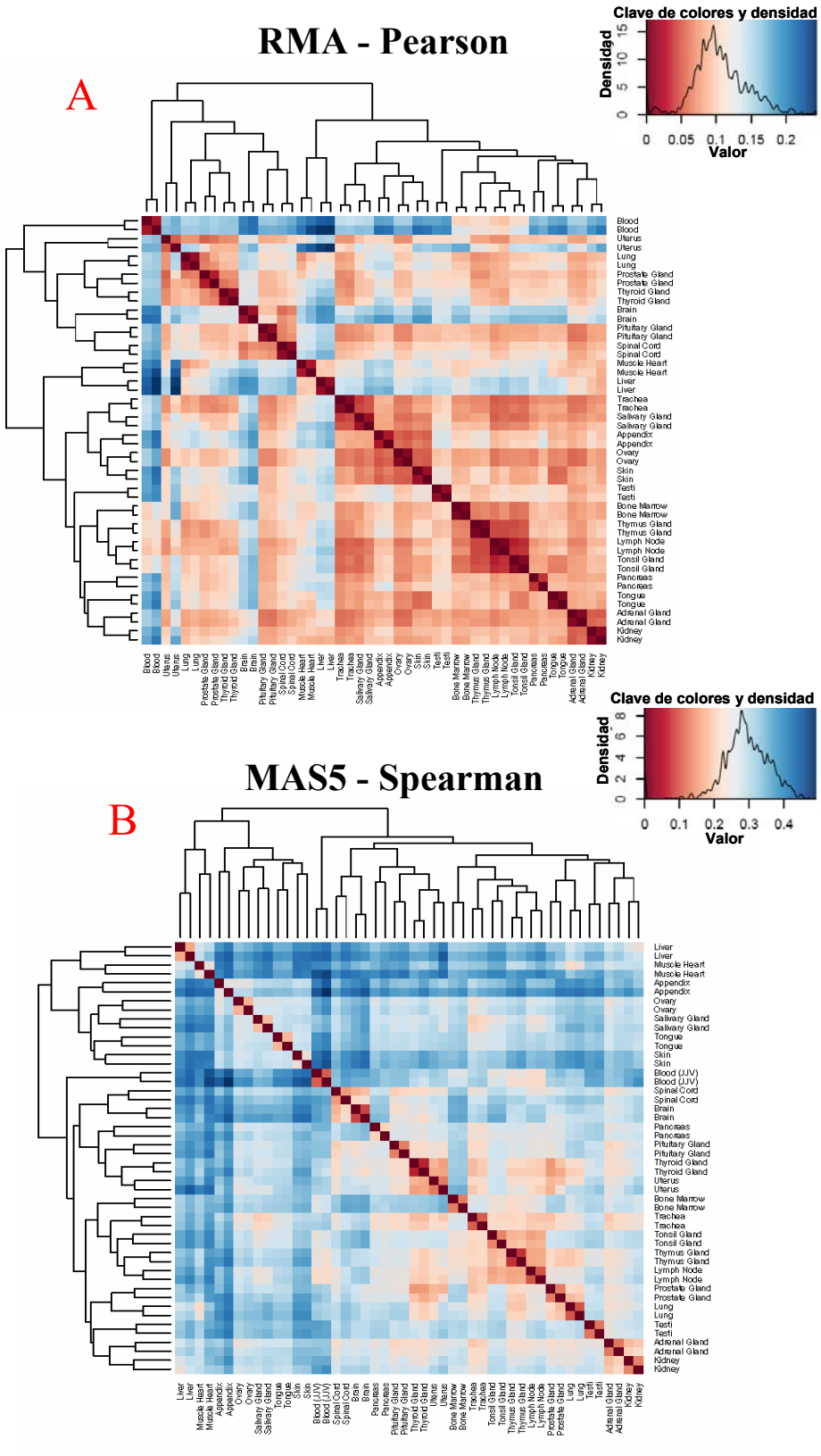


Figura 3.9: Análisis de los datos de entrada.
Heatmaps y *clustering* de 48 muestras de *microarrays* provenientes de 24 tejidos y órganos diferentes, analizadas por dos métodos: A) *MAS5-Spearman*: *MAS5* para el cálculo de señal y *Spearman* para el cálculo de la distancia entre muestras; y B) *RMA-Pearson*: *RMA* para el cálculo de señal y *Pearson* para el cálculo de la distancia entre muestras.

adrenal, apéndice, sangre, médula espinal, cerebro, riñón, hígado, pulmón, ganglio linfático, corazón, ovario, páncreas, glándula pituitaria, próstata, glándula salivar, piel, médula osea, testículo, timo, glándula tiroidea, lengua, amígdala, traquea y útero.

La matriz global de expresión se calculó mediante dos algoritmos, *RMA* (Irizarry, *et al.*, 2003) y *MAS5* (Lim, *et al.*, 2007; Liu, *et al.*, 2002), que incluyen corrección de *background*, normalización y cálculo de señal. Para ver la heterogeneidad de las muestras y la validez de las réplicas se realizó un análisis de *clustering* no supervisado resultando 2 *heatmaps* (**figura 3.9**), en uno se calculó la señal con el algoritmo *MAS5* y la distancia de *Spearman* entre las muestras (*MAS5-Spearman*) y en otro se calculó la señal con *RMA* y se usó *Pearson* como distancia entre muestras (*RMA-Pearson*). Ambos *heatmaps* muestran una clara proximidad entre réplicas, esta proximidad es mayor para el segundo método pero en ninguno de los dos casos hay una separación de las réplicas. Por otro lado, el orden y la *clusterización* de los tejidos en las figuras es muy variante. Esta observación se confirmó con un análisis de *bootstrap* realizado con el algoritmo *pvclust* (Suzuki y Shimodaira, 2006) que permite ver la variabilidad a la hora de formar *clusters* jerárquicos. En la **figura 3.10** se puede ver que las réplicas

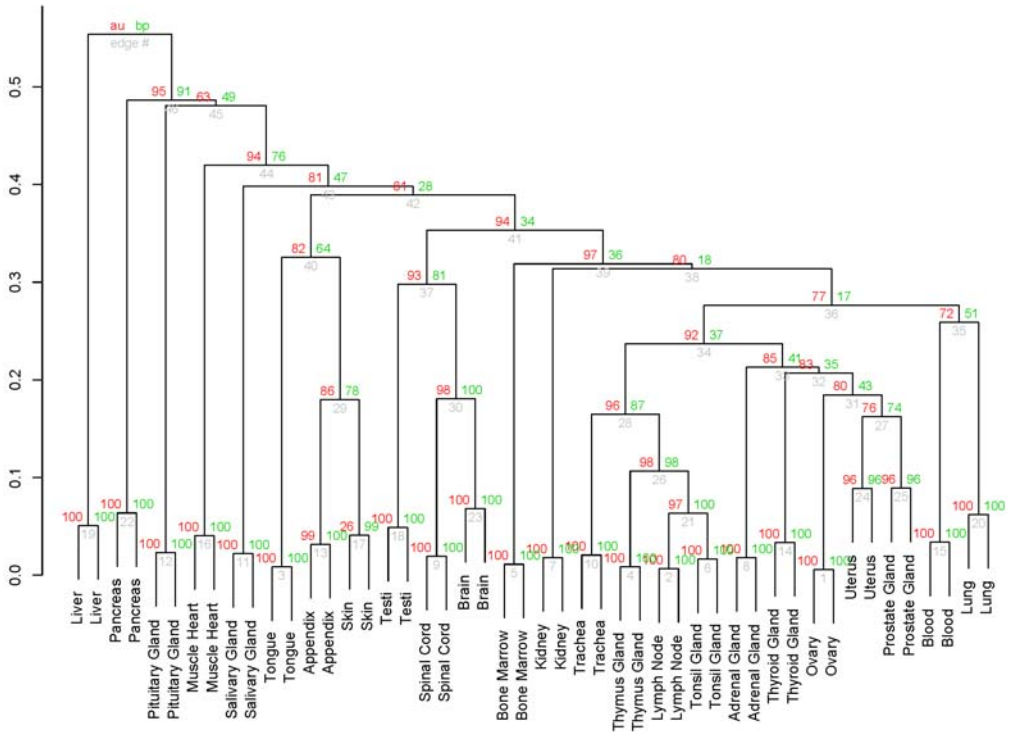


Figura 3.10: Similitud y agrupamiento de los datos de entrada.
Análisis de *bootstrap* con *pvclust* (Suzuki y Shimodaira, 2006) sobre los 48 *microarrays* seleccionados. La probabilidad de *bootstrap* (números en verde) es baja para las ramas no correspondientes a las réplicas (por encima de 0.1). Por tanto se puede deducir que hay gran variabilidad a la hora de formar *clusters* jerárquicos. La única excepción que se observa corresponde al *clúster* formado por: timo, ganglio linfático y amigdala.

biológicas tienen una estabilidad óptima, sin embargo, la *clusterización* entre tejidos solo obtiene dos grupos pequeños más estables: el que incluye timo, ganglio linfático y amígdala (todos ellos órganos relacionados con el sistema linfático) y el que incluye útero y próstata (ambos relacionados con el aparato y proceso reproductor) aunque su proximidad no es grande. Por lo tanto se puede decir que en general las muestras son suficientemente diferentes y distantes en sus perfiles de expresión. En conclusión podemos afirmar que hay una clara separación entre la mayoría de las muestras, lo que hace que sean unas muestras adecuadas para una exploración global del mapa de expresión humano.

3.2.2 Cálculo de correlación con validación cruzada

Los estudios de búsqueda de correlación entre perfiles de expresión, generalmente aplican medidas indicativas de la distancia o correlación entre *probesets*. Entre las más utilizadas están *Pearson* y *Spearman* [32]. Estas medidas de correlación se calculan para cada par de *probesets* del *microarray* de expresión; dado que en este estudio se emplearon 48 *microarrays* de Affymetrix HGU133a que tienen 22283 *probesets*, la matriz global de correlación tiene un total de 248 millones de datos correspondientes a todas las parejas de *probesets* posibles. Para calcular la correlación se utilizó *Spearman* en la matriz de expresión obtenida con *MAS5* y *Pearson* en la matriz de expresión obtenida con *RMA*. Esta decisión se tomó porque *MAS5* genera datos paramétricos y *Spearman* es un test no paramétrico, robusto ante posibles *outliers*. Por otro lado *Pearson* es un test paramétrico y *RMA* genera datos de expresión aplicando una normalización no paramétrica.

La gran cantidad de pares de correlación tiene que ser evaluada para que se puedan identificar a los verdaderos positivos. Con este propósito se ideó un método basado en validación cruzada que pudiera identificar correlaciones estables y significativas. Esta estrategia, representada en la **figura 3.11**, seleccionó aleatoriamente 1000 conjuntos de *microarrays* con un tamaño del 25% del conjunto inicial de *microarrays* (12 *microarrays*, de los cuales 6 son réplicas) y calculó la correlación entre todos los *probesets* para cada uno de los 1000 conjuntos de *microarrays*. Solo cuando el coeficiente de correlación (**r**) tomaba un valor igual o superior a 0.7 fue considerado como una validación cruzada positiva. Estos positivos se fueron acumulando para cada pareja de genes y así se obtuvo el coeficiente de validación cruzada (**N**). De esta

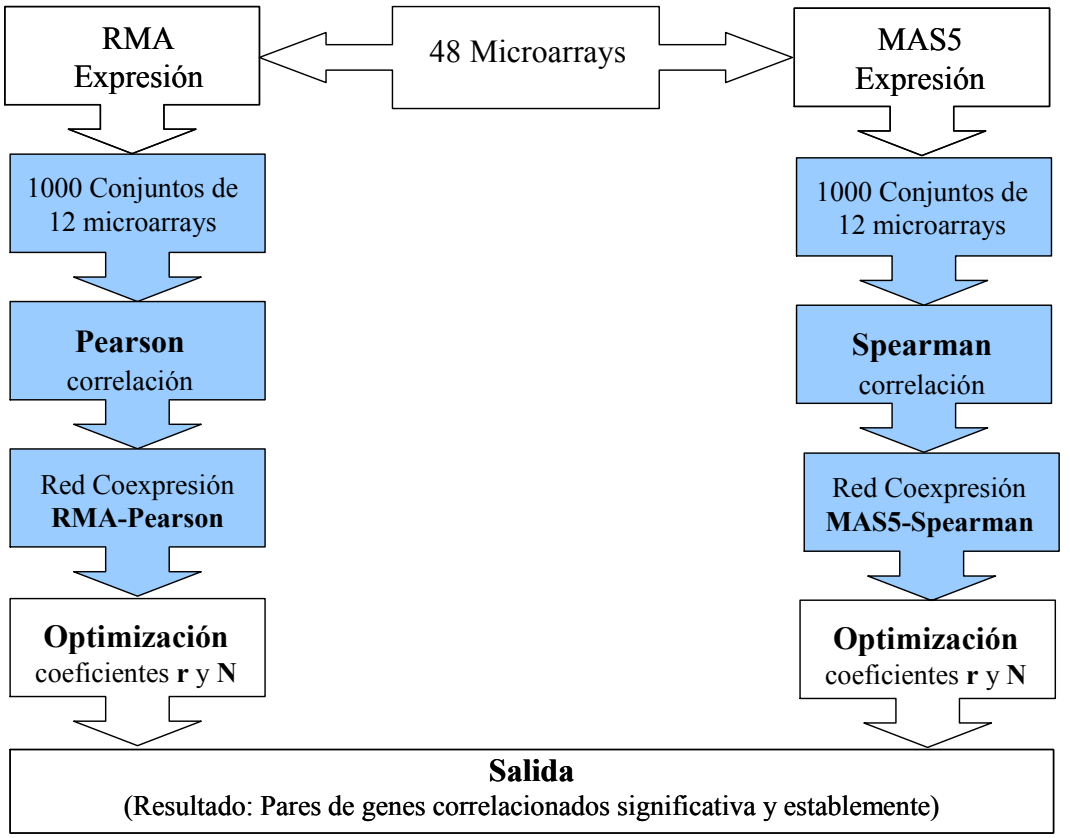


Figura 3.11: Cálculo de correlación con validación cruzada.
Esquema del estudio que se ha realizado, en el que se ha propuesto hacer un cálculo de correlación con validación cruzada.

forma, que una pareja de genes tenga 620 sucesos de correlación cruzada positivos (**N**), significa que ha tenido un valor de correlación (**r**) igual o superior a 0.7 en 620 de los 1000 conjuntos de *microarrays* generados. En la **figura 3.12** tenemos una representación de los resultados del método de validación cruzada propuesto. En ella se representa un sistema de coordenadas cartesianas en cuyo eje de ordenadas (**y**) están los coeficientes de correlación (**r**) obtenidos para una pareja de *probesets* y en el eje de abscisas (**x**) están los valores del coeficiente de validación cruzada (**N**) obtenidos.

Estos gráficos permiten identificar parejas de *probesets* que tienen una correlación significativa con validación cruzada, que están segregadas de las que no tienen una correlación diferenciable del ruido, es decir, que tienen valores bajos de **r** y **N**. Los círculos de la figura representan parejas de *probesets* divididas en 3 grupos (diferenciados por el color) en función de los parámetros **r** y **N**. En color negro se han representado 10000 parejas de *probeset* elegidas al azar para contrastarlas con parejas que en principio deberían de coexpresar. De esta forma las parejas de *probesets* que mapean el mismo gen se han representado en color rojo, estás son, por ejemplo, todas las posibles parejas entre los 4 *probesets* que mapean el gen ALDOB (fructosa

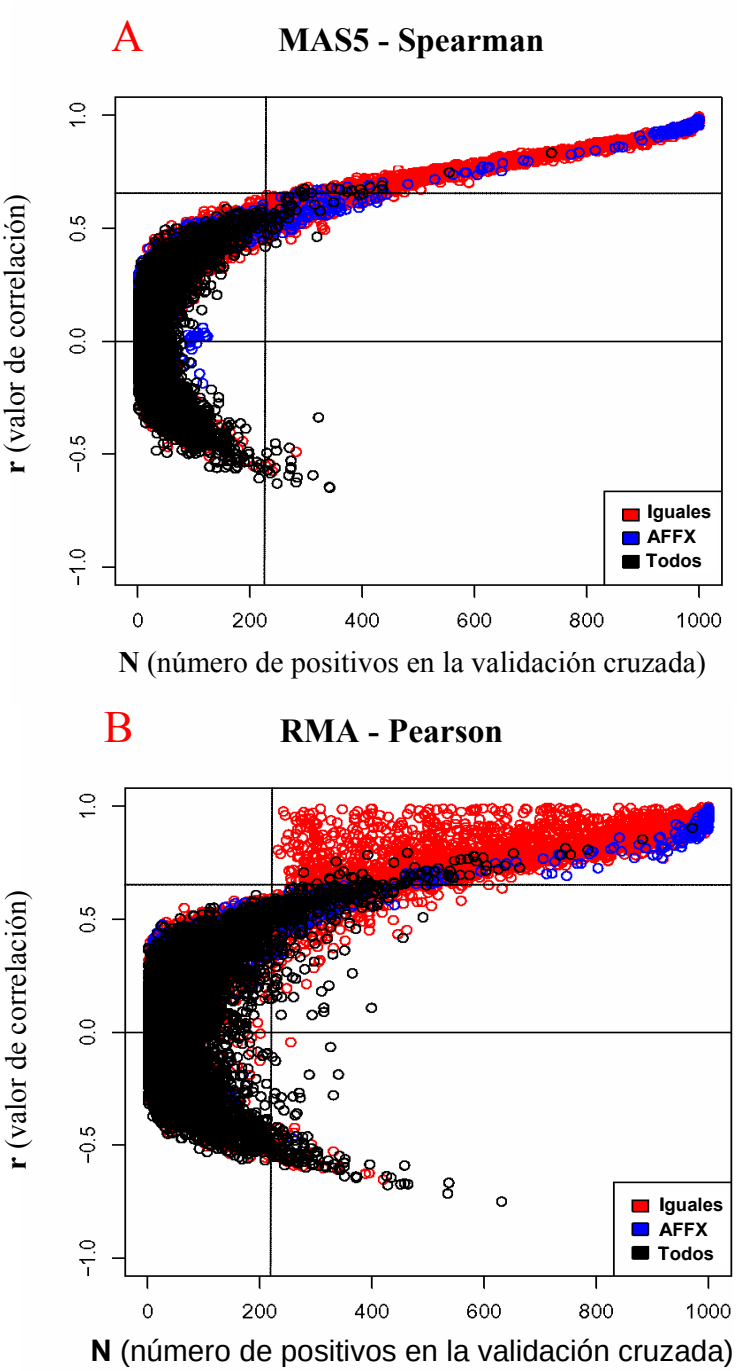


Figura 3.12: Comportamiento de r y N en distintos tipos de parejas de genes.
rN-Plots que representan para cada pareja de *probesets* el valor del coeficiente de validación cruzada (**N**) y el coeficiente de correlación (**r**) obtenido por dos métodos: **A**) *MAS5-Spearman* y **B**) *RMA-Pearson*. Las relaciones de coexpresión entre *probesets* que mapean el mismo gen se representa con puntos rojos. Las parejas de *probesets* que son controles del *microarray* se representan con puntos azules. Con puntos **negros** se representan 10000 parejas de *probesets* coexpresantes que se han obtenido aleatoriamente de la matriz global de datos de coexpresión.

bifosfato aldolasa: *204704_s_at*, *204705_x_at*, *211357_s_at* y *210622_x_at*). Como se ve en la **figura 3.12**, estas parejas de *probesets* se acumulan para valores altos de **r** y **N** (**r>0.65** y **N > 200**) en contraste con los puntos de color negro que se acumulan en valores bajos de **r** y **N**.

Para hacer una evaluación más estricta se representaron en color azul las parejas de *probesets* usados como controles, que son añadidos durante el protocolo de hibridación en los experimentos de *microarrays*. Estos controles, que se nombran con el prefijo AFFX en el chip, son genes *housekeeping* o *mRNAs* no correspondientes a genes humanos que se añaden en el protocolo de hibridación en diferentes concentraciones controladas. Como se ve en la **figura 3.12**, los controles añadidos en concentraciones similares tienen una coexpresión muy fuerte, es por ello que los controles están acumulados a valores muy altos de **r** y **N** y, nos indican los valores de los parámetros en los que hay una coexpresión significativa.

3.2.3 Filtros de genes para el cálculo de correlación

Hasta ahora se ha trabajado con matrices de expresión completas, es decir usando todos los *probesets* presentes en el *microarray*. Pero no todos los *probesets* aportan información. En función de las muestras y las condiciones en las que se han hecho los *microarrays*, hay una gran proporción de genes que no están expresados y sus *probesets* tienen una señal cercana al ruido. Es importante descubrir tanto la presencia como el efecto de estos genes no cambiantes (también llamados "genes planos") (Prieto, *et al.*, 2006) a la hora de aplicar el método de coexpresión, ya que pueden aparecer fácilmente en los resultados como falsos positivos.

Para evitar estos genes se aplicaron en combinación dos filtros de genes que podemos llamar "no informativos": uno basado en la variabilidad de la expresión y otro basado en el nivel de intensidad. Este filtrado dejó solo los *probesets* que cumplían alguna de estas dos condiciones: **1**) Que la diferencia de expresión del *probeset* esté por encima de la mediana de las diferencias de expresión calculadas para todos los genes: $\mathbf{\Delta Exp^{gi}_{máximo\text{-}mínimo} > mediana(\Delta Exp^{gx}_{máximo\text{-}mínimo})}$. **2**) Que la expresión media del *probeset* en todas las muestras esté por encima de la mediana de todas las señales de expresión calculadas para cada gen: $\mathbf{media\ Exp^{gi} > mediana\ (media\ Exp^{g})}$.

El uso de este filtro dio resultados muy diferentes entre los datos obtenidos con *RMA* y *MAS5*. Se filtraron 6893 *probesets* (69.06%) en el primer caso y 3682 *probesets* (83.48%) en el segundo del total de 22283 *probesets*. La diferencia en estos números muestra que estos dos métodos no dan un cálculo equivalente de la señal de expresión y por tanto es posible que puedan aportar resultados complementarios.

3.2.4 Estimación de la precisión y la cobertura estadísticas en los datos de coexpresión

Para poder validar estos filtros así como para estudiar si el ajuste de los parámetros definidos dan una mejora de la señal biológica, se hizo un estudio de precisión y cobertura en función de los parámetros **r** y **N**. La precisión se midió como el "valor positivo predicho" (**VPP**) definido como **VP / (VP+FP)**, donde **VP** es el número de verdaderos positivos y **FP** es el número de falsos positivos (Loong, 2003; Suojanen, 1999). La cobertura se midió como la proporción de verdaderos positivos que hay en un punto de corte determinado en relación con el conjunto inicial de verdaderos positivos (**VP / Todos los positivos**).

Estas medidas del error estadístico necesitan conocer, o al menos estimar, qué parejas de coexpresión son verdaderas. La estimación se hizo bajo el postulado de que los genes que trabajan en la misma ruta o proceso biomolecular es más probable que coexpresen, que los que no están relacionados en algún proceso común. Para anotar los gen *probesets* del *microarray* a rutas biomoleculares se utilizó la base de datos *KEGG*. Esta base de datos es uno de los repositorios más completos de rutas biomoleculares curadas manualmente (Kanehisa, 2002). Por tanto, seleccionando solo el grupo de genes que está anotado a *KEGG*, se consideraron verdaderos positivos a las parejas de genes coexpresantes anotadas al mismo identificativo de ruta molecular *KEGG*. Esta estrategia permite calcular la precisión y la cobertura antes definida y, por tanto, se puede explorar como evolucionan dichos parámetros en función de **r** y **N**. Para poder tener una medida más adecuada de la precisión y de la cobertura, se transformaron los identificativos de *probesets* a genes por medio de la anotación que proporciona *Affymetrix*, es por ello que a partir de ahora se habla de genes en lugar de *probesets*.

Los análisis de precisión y cobertura están representados en la **figura 3.13**. EL **VPP** en función de **r** y **N** está representado en color rojo para los datos filtrados y en azul para los no filtrados, la cobertura está representada en negro. Los gráficos muestran que la tasa de verdaderos positivos **VP** se incrementa para valores altos de **r** y **N**. Este incremento es más significativo para los datos provenientes de *MAS5-Spearman* que toman un **VPP > 80%** con **r > 0.8** o **N > 700**. Esto es debido a que *RMA-Pearson*, para unos parámetros similares, tiene un número muy superior de datos coexpresados.

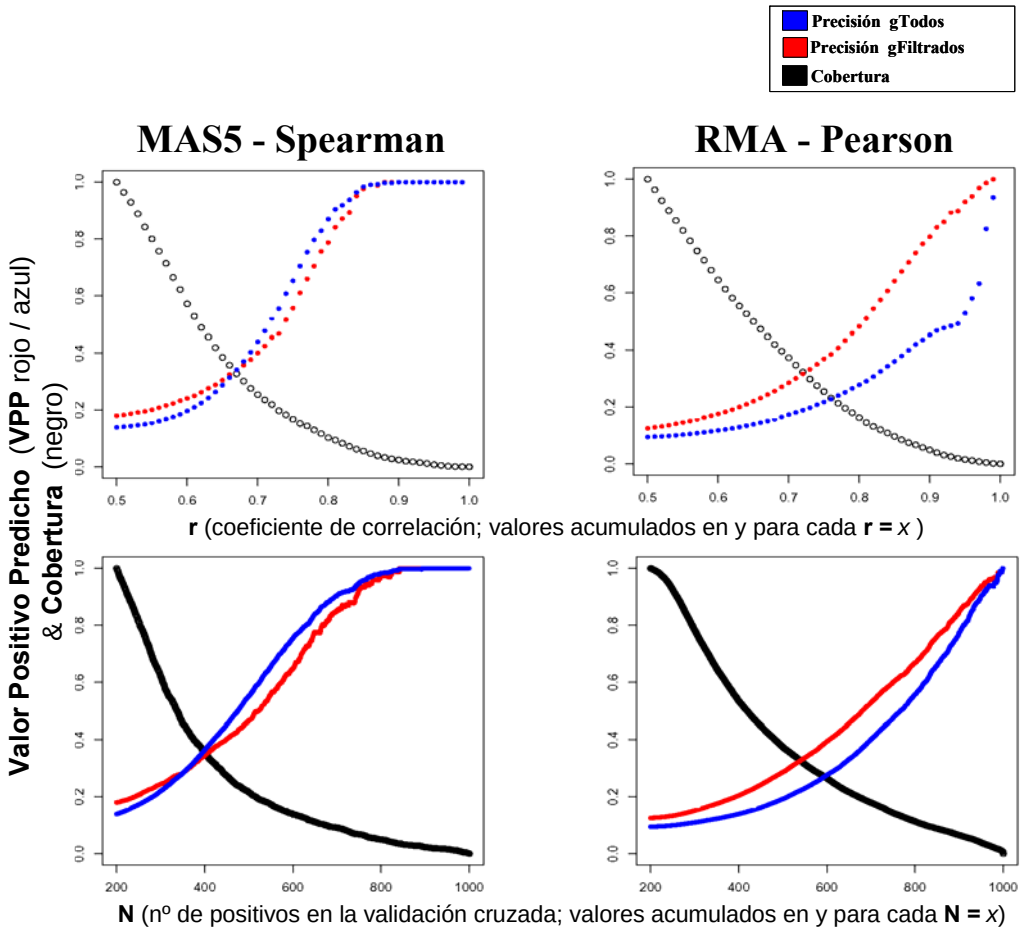


Figura 3.13: Precisión y cobertura de los datos de coexpresión en función de r y N. Representación para ambos métodos (*RMA-Pearson* y *MAS5-Spearman*) de la evolución de la cobertura (en **negro**) y la precisión (medida como **VPP**), para los genes prefiltrados (en rojo) o sin filtrar (en azul), en función del coeficiente de correlación (**r**) y de validación cruzada (**N**).

La cobertura es la proporción de **VP** que queda respecto al total inicial, calculada para cada valor de los parámetros **r** o **N**. Se calcula tomando como número total de parejas positivas a las existentes en los puntos de corte donde empieza la representación (**r = 0.5** y **N = 200**). La cobertura indica cómo va decreciendo el número de positivos encontrados a medida que crecen los parámetros **r** y **N**. Esta disminución es más acusada para *MAS5-Spearman* ya que para **r >= 0.75** la cobertura toma un valor de 16,7% mientras que *RMA-Pearson* para el mismo **r** tiene una cobertura del 25,4%. Lo mismo ocurre para un **N >= 600** en el que *MAS5-Spearman* tiene un 13,9% de los positivos mientras que *RMA-Pearson* tiene un 26,4%. La cantidad total de positivos al principio de la curva es 15657 para *RMA-Pearson* y 2198 para *MAS5-Spearman*. Este número puede parecer pequeño pero se corresponde solo con las parejas positivas. Si tomamos el total de parejas de coexpresión (no incluyendo solo las anotadas a *KEGG*), tenemos 1340472 parejas de genes para *RMA-Pearson* y 180305 para *MAS5-Spearman*. Estos datos indican que *RMA-Pearson* tiene una cobertura mayor.

En conclusión, el estudio muestra que el método con *RMA-Pearson* tiene una cobertura del coexpresoma superior a *MAS5-Spearman*, pero es menos preciso. El problema de la precisión en *RMA-Pearson* es suavizado prefiltrando los datos de entrada tal y como se explica en el punto anterior (3.2.3).

La **figura 3.13** también representa el efecto del filtrado en la precisión, en color rojo están representados los **VPP** de los datos de coexpresión filtrados en función de los parámetros **r** y **N**. Este análisis comprobó que se consigue una mejora considerable si se aplica el filtro propuesto a los datos de coexpresión calculados con *RMA-Pearson*. Sin embargo, para los datos calculados con *MAS5-Spearman*, no hay una mejora significativa: el filtro elimina casi la misma cantidad de verdaderos positivos que de falsos positivos y, por tanto, no aporta ninguna mejora al método, perdiendo además cobertura. Según esto lo más adecuado es aplicar el filtro solo a los datos calculados con *RMA-Pearson*. De esta forma se obtuvieron dos *sets* de datos con un número muy similar de relaciones de coexpresión: por ejemplo, para los parámetros **N >= 200** y **r >= 0.5** el set de *MAS5-Spearman* no filtrado incluye 15623 parejas de coexpresión; que es un número muy similar al de *RMA-Pearson* filtrado que incluye 15657 parejas coexpresadas.

3.2.5 Generación de redes de coexpresión fiables

Hasta ahora se han representado los parámetros **N** y **r** de forma separada para ver cómo afectan al resultado. En ambos casos se ha observado que a medida que éstos son más estrictos se aumenta el **VPP**. Para ver el efecto de estos parámetros de forma combinada se representaron figuras con tres dimensiones en las que el color se corresponde con el **VPP** en función de **N(x)** y **r(y)**. Los resultados se muestran en la **figura 3.14**. Los colores y la intensidad de los mismos reflejan los valores de **VPP** que varían de 0.05 a 1. Este gráfico está representado para los métodos, *MAS5-Spearman* no filtrado y *RMA-Pearson* con prefiltrado. Como se ha indicado antes, en estas condiciones ambos *sets* tienen un número similar de parejas de coexpresión. Partiendo de estas figuras, para la generación de redes de coexpresión fiables se utilizó una estrategia de optimización de cobertura para un determinado nivel de precisión. De este modo, se seleccionaron tres *sets* de datos derivados de cada método a tres **VPP**: 0.60, 0.70 y 0.80. Los valores de correlación **r** y de validación cruzada **N** correspondientes a estos *sets* están reflejados en la **tabla 3.1**. Estos valores muestran

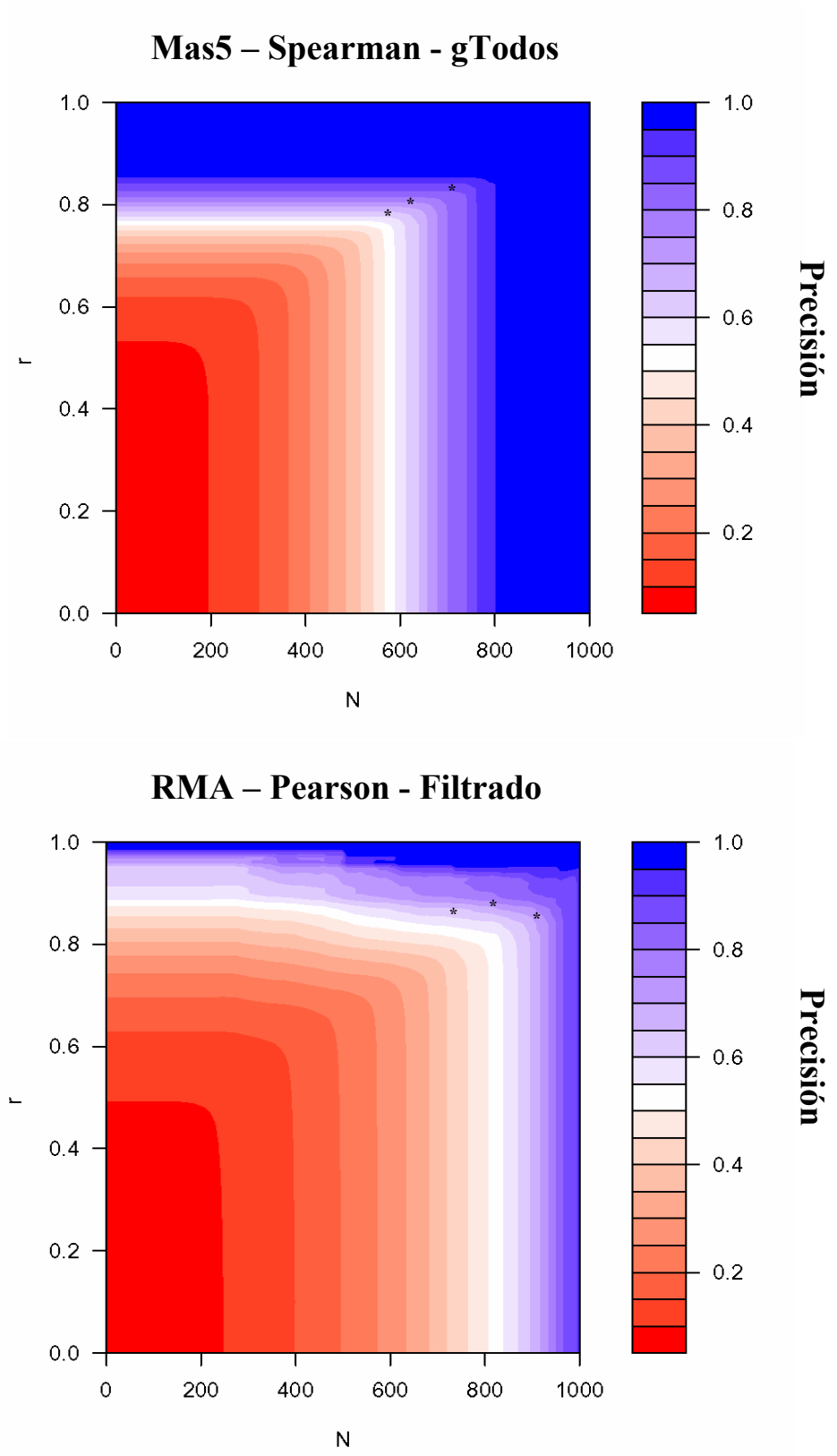


Figura 3.14: Precisión en función de los parámetros r y N.
Gráficos en tres dimensiones que representan la precisión (**VPP**) de cada *set* de coexpresión (*RMA-Pearson* con prefiltrado y *MAS5-Spearman*) en función del coeficiente de correlación (**r**) y el coeficiente de validación cruzada (**N**).

VPP [1]	Coeficientes		Nº de Nodos [2]	Nº de *Links* [2]
	N	r		
MAS5-Spearman (Todos los genes)				
≥ 0.60	605	0,77	3052	**12669**
≥ 0.70	645	0,79	2295	7874
≥ 0.80	695	0,81	1762	4910
RMA-Pearson (Genes filtrados)				
≥ 0.60	765	0,85	1672	**5945**
≥ 0.70	835	0,87	1215	3273
≥ 0.80	925	0,84	983	2423

1 VPP que se corresponde con las redes derivadas para genes anotados a KEGG
2 Valores que se corresponden con las redes completas que incluyen todos los genes

Tabla 3.1: Redes de coexpresión generadas a distintos valores de precisión.
Tabla con los coeficientes de correlación (**r**) y de validación cruzada (**N**) que producen una cobertura mayor para un valor de precisión (**VPP**) determinado. En la tabla se especifica también el número de nodos y de relaciones de coexpresión de las redes resultantes.

	Redes de coexpresión			
PPV [1]	Unión		Intersección	
	Nº Nodos	Nº *Links*	Nº Nodos	Nº *Links*
≥ 0.60	3327	**15841**	731	**2249**
≥ 0.70	2411	9264	542	1447
≥ 0.80	1863	5935	387	1008

Tabla 3.2: Unión e intersección de redes de coexpresión.
Número de nodos y de relaciones de coexpresión que resultan de combinar (por intersección y unión) los datos provenientes de *RMA-Pearson* y *MAS5-Spearman*, a 3 valores de precisión (**VPP**) diferentes 0.6, 0.7, 0.8.

que el método *RMA-Pearson* es menos estricto ya que para conseguir un mismo **VPP** necesita tener valores más altos de **r** y **N**. Los *sets* de datos obtenidos pese a tener una precisión similar son bastante heterogéneos; esto se puede ver en la **tabla 3.2** en la que se hizo la intersección y la unión de los distintos *sets* seleccionados. Por ejemplo, la unión para un **VPP** >= 0.6 da una red de coexpresión que incluye 3327 genes y 15841 relaciones de coexpressión. Esta unión nos permite tener una visión más general de la red de coexpresión no sesgada por la aplicación de un único método de normalización. En el caso de la intersección se pueden ver las relaciones de coexpresión encontradas por ambos métodos. Al mantenerse en ambos métodos reflejan relaciones de coexpresión estables, seguramente esenciales para el funcionamiento celular.

3.2.6 Coherencia funcional de los módulos de las redes generadas

Una vez definido el método de búsqueda de perfiles coexpresados y evaluados los parámetros estadísticos, se hizo un estudio sobre el significado biológico y la coherencia funcional de los *sets* de datos. En primer lugar se investigó la coexpresión de genes *housekeeping* en los *sets* generados. Es de esperar que esta coexpresión se dé a niveles altos de **N** y de **r**, ya que reflejarían pares de coexpresión conservados en casi todos los tejidos seleccionados. Para hacer este estudio se tomaron dos grupos de genes *housekeeping* generados con diferentes métodos descritos en dos publicaciones (Eisenberg y Levanon, 2003; Hsiao, *et al.*, 2001). *Hsiao et al.* identifican 451 genes expresados en 19 tejidos humanos, *Eisemberg et al.* identifican 575 genes humanos que están expresados en diferentes *sets* de *microarrays* públicos. En las **figuras** superiores **3.15 A y B**, se representan las distribuciones de las parejas coexpresadas tomando todos los genes (en negro), y tomando solamente los genes *housekeeping* definidos por Hsiao (en rojo) y Eisemberg (en verde). Como se podía esperar, los genes *housekeeping* coexpresan en todos los tejidos y por tanto su distribución está desplazada a valores altos de **N**. En las gráficas inferiores **3.15 A** y **B**, se han representado las parejas de coexpresión en función de los parámetros **r** y **N**, en negro se han representado 10000 parejas de coexpresión elegidas al azar y en rojo las parejas de coexpresión en las que están implicados genes *housekeeping* definidos por Hsiao. Además de ver que los *housekeeping* coexpresan para valores más altos de **r** y **N**, se puede ver que *MAS5-Spearman* muestra principalmente relaciones entre genes *housekeeping* mientras que en *RMA-Pearson* hay relaciones de coexpresión, con niveles altos de **r** y bajos de **N**, que no se dan para genes *housekeeping*. Con esta observación se puede decir que *RMA-Pearson*, utilizando un método de validación cruzada, permite discernir entre relaciones generales de la célula y relaciones específicas de tejido.

También es interesante ver cómo se distribuyen las parejas de coexpresión (calculadas con *RMA-Pearson*) que están implicadas en determinadas rutas biomoleculares (*pathways*). La **figura 15C** consta de 6 gráficos correspondientes a distintos *pathways*

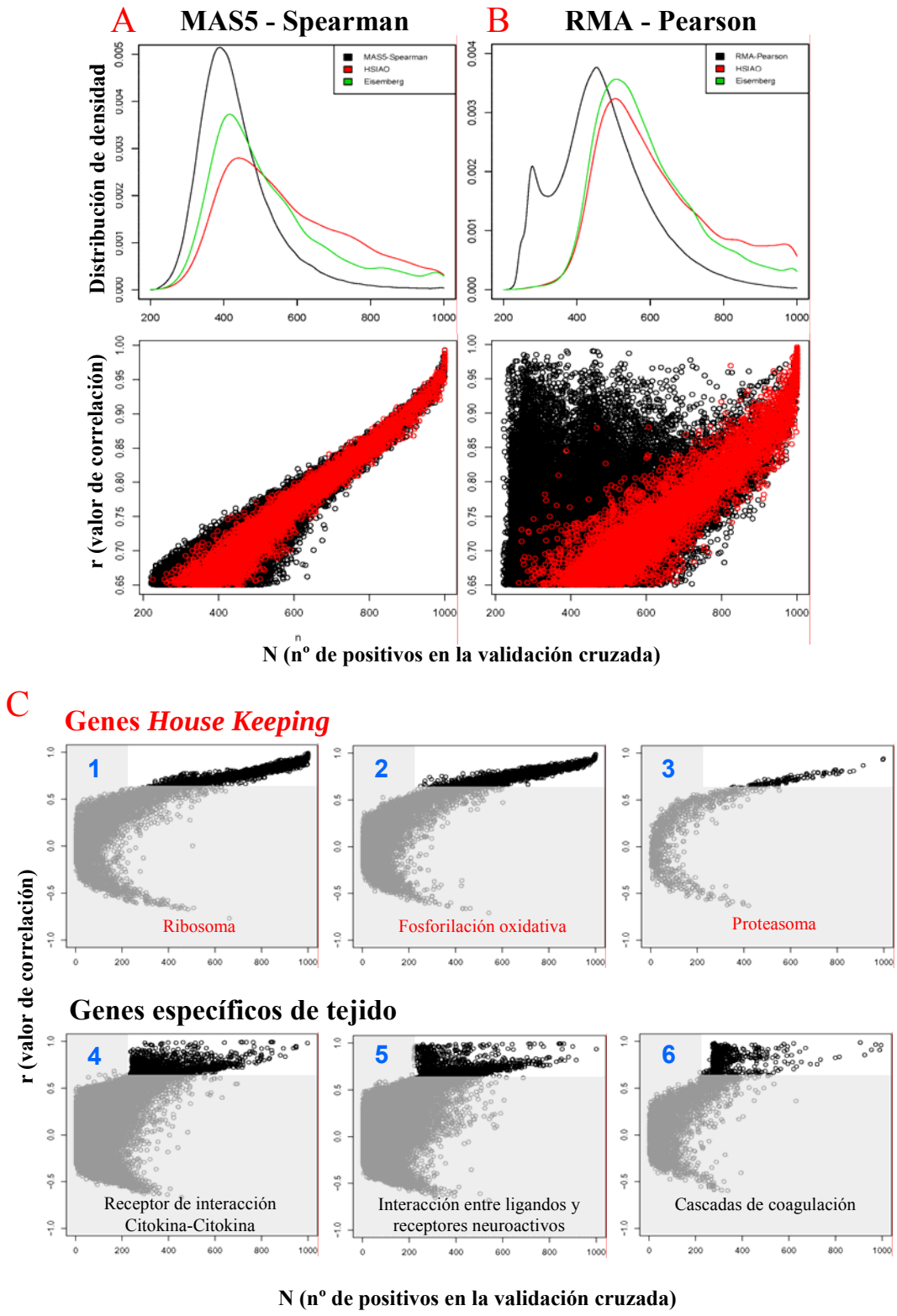


Figura 3.15: Genes housekeeping en los datos de coexpresión.
Gráficos A y **B superiores**: Distribuciones de densidad de los datos de coexpresión para **N>220**, en **negro** se representa para todos los genes del *microarray*, en **verde** para los genes *housekeeping* de *Eisenberg et al.* y en **rojo** para los genes *housekeeping* de *Hsiao et al.* **Gráficos A** y **B inferiores**: *rN-plots* con las parejas de expresión de todos los genes (**negro**) y de genes housekeeping de *Hsiao et al.* (**rojo**), para **N>220** y **r>0.65**. **Gráficos C**: 6 *rN-plots* que representan las parejas de coexpresión en función de **r** y **N**, de genes involucrados en 6 *pathways* diferentes: **1**) ribosoma, **2**) fosforilación oxidativa, **3**) proteasoma, **4**) receptor de interacción citokina-citokina, **5**) ligando-receptor de interacción neuroactiva y **6**) cascadas de coagulación.

definidos en KEGG; en ellos se representaron las relaciones de coexpresión de los genes anotados a esos *pathways* en función de **r** y **N**. Se escogieron por un lado 3 *pathways* que desempeñan funciones constitutivas de la célula, es decir, presentes en todos los tejidos y tipos celulares: ribosoma (KEGG id = hsa03010), fosforilación oxidativa (hsa00190) y proteosoma (hsa03050). Además se eligieron por otro lado 3 *pathways* específicos de tejido: receptores de interacción citokina-citokina (hsa04060), interacción entre ligandos y receptores neuroactivos (hsa04080), y cascadas de coagulación (hsa04610). Las diferencias de especificidad funcional están claramente reflejadas en las distribuciones de coexpresión, de forma que los **gráficos 1**, **2** y **3** muestran relaciones de coexpresión para niveles altos de **N** y **r**, mientras que para los **gráficos 4**, **5** y **6** las relaciones de coexpresión tienen niveles bajos de **N** y altos de **r**. En conclusión este análisis nos ha servido como validación del método de coexpresión propuesto y para ver la existencia, en el método *RMA-Pearson*, de relaciones de coexpresión con valores altos de **r** y bajos de **N**, que tienden a ser entre genes específicos de tejido.

3.2.7 Comparación de redes de coexpresión de genes humanos

Es interesante comparar los resultados de coexpresión, con otros conjuntos de genes coexpresantes humanos propuestos por otros autores anteriormente. Para ello hicimos dos análisis midiendo la cobertura y la fiabilidad de los coexpresomas. Para testar la cobertura usamos la estrategia de *Stuart et al.*, en la que se exploró la cobertura funcional de redes de coexpresión obtenidas para 4 organismos (Stuart, *et al.*, 2003), calculamos el porcentaje de genes de una categoría funcional que está conectado con otros de la misma categoría. Este análisis se hizo basándose en asignaciones funcionales a *KEGG* y se empleó: (**i**) la red *RMA-Pearson* para **r >0.63** y **N >500**, (**ii**) los resultados publicados por *Lee et al.* y (**iii**) los publicados por *Griffith et al*. El resultado de esta comparativa se muestra en la **tabla 4.3**, que incluye las 10 rutas moleculares con mejor porcentaje de genes asignados al mismo *KEGG*. Este análisis comparativo de la cobertura funcional muestra varias conclusiones: (**i**) todos los *sets* de datos estudiados muestran la coexpresión más significativa para 3 máquinas moleculares: **ribosoma**, **proteasoma** y **fosforilación oxidativa**; (**ii**) los genes involucrados, en interacción o anclaje de células, tienen una cobertura muy alta tal y como indica la presencia de *pathways* como **adhesión** focal, interacción de la **matriz**

Este trabajo (2008)				
Ruta/Vía funcional (KEGG ID)	**nº gn [1]**	**gn coexp / gn [2]**	**% gn coexp**	**r media [3]**
Proteasome (3050)	31	28 / 28	**100,0%**	0,69
Ribosome (3010)	120	52 / 55	**94,5%**	0,75
Oxidative phosphorylation (190)	129	88 / 95	**92,6%**	0,73
Focal adhesion (4510)	194	154 / 168	**91,7%**	0,68
Antigen processing and presentation (4612)	86	71 / 78	**91,0%**	0,75
Glycan structures - degradation (1032)	30	20 / 22	**90,9%**	0,65
Neuroactive ligand-receptor interact. (4080)	299	227 / 255	**89,0%**	0,68
Cell cycle (4110)	114	90 / 102	**88,2%**	0,66
Regulation of actin cytoskeleton (4810)	208	141 / 161	**88,2%**	0,66
Cytokine-cytokine receptor interact. (4060)	256	196 / 223	**87,9%**	0,69
Lee et al. (2004)				
Ruta/Vía funcional (KEGG ID)	**nº gn [1]**	**gn coexp / gn [2]**	**% gn coexp**	
Ribosome (3010)	120	43 / 44	**97,7%**	
Proteasome (3050)	31	19 / 22	**86,4%**	
Oxidative phosphorylation (190)	129	31 / 44	**70,5%**	
Cell cycle (4110)	114	33 / 47	**70,2%**	
ECM-receptor interaction (4512)	87	16 / 23	**69,6%**	
Gap junction (4540)	92	9 / 13	**69,2%**	
Pathogenic Escherichia coli infection (5130)	49	11 / 16	**68,8%**	
Pathogenic Escherichia coli infection (5131)	49	11 / 16	**68,8%**	
T cell receptor signaling pathway (4660)	93	15 / 22	**68,2%**	
Metabolism of xenobiotics by cytP450 (980)	70	7 / 11	**63,6%**	
Griffith et al. (2005)				
Ruta/Vía funcional (KEGG ID)	**nº gn [1]**	**gn coexp / gn [2]**	**% gn coexp**	
Ribosome (3010)	120	36 / 38	**94,7%**	
Proteasome (3050)	31	20 / 24	**83,3%**	
Oxidative phosphorylation (190)	129	55 / 67	**82,1%**	
Val, Leu and isoleucine degradation (280)	50	15 / 19	**78,9%**	
ECM-receptor interaction (4512)	87	16 / 22	**72,7%**	
Cell cycle (4110)	114	36 / 51	**70,6%**	
Propanoate metabolism (640)	34	9 / 14	**64,3%**	
Butanoate metabolism (650)	44	9 / 14	**64,3%**	
Hematopoietic cell lineage (4640)	88	18 / 28	**64,3%**	
beta-Alanine metabolism (410)	24	7 / 11	**63,6%**	

1 nº gn = número total de genes incluido en esa ruta KEGG
2 gn coexp / gn = genes que coexpresan en la red con genes anotados a esa ruta KEGG
3 valor medio de la correlación (**r**) para las parejas de genes coexpresantes incluidas en esa ruta KEGG

Tabla 3.3: Comparativa de redes de coexpresión según su cobertura funcional.
Estudio de la cobertura de tres redes de coexpresión (la generada en este estudio, la de *Lee et al.* y la de *Griffith et al.*) medida como el porcentaje de genes de una vía o ruta molecular de *KEGGs* que tienen alguna relación de coexpresión detectada con otro gen anotado a esa ruta molecular.

extracelular y regulación del **citoesqueleto** en la parte alta de la lista; (**iii**) los genes relacionados con el **ciclo celular** son también frecuentes en los 3 *sets* de datos, indicando que las células guardan una gran regulación de genes involucrados en las funciones esenciales para la vida de la célula (mantenimiento, proliferación y supervivencia); (**iv**) otra diferencia importante entre nuestro coexpresoma y los definidos por *Lee* y *Griffith*, es que nuestro trabajo solo incluye muestras que no provienen de tejidos patológicos, la inclusión de estos tejidos puede sesgar los

	Genes [1]	Vínculos [2]	VP [3]	VP+FP [4]	VPP [5]
Este trabajo (2008)	3052	12669	729	1189	**0,613**
Lee et al. (2004)	1751	12187	1275	2265	**0,563**
Griffith et al. (2005)	2922	12686	1265	2588	**0,489**

1 Nº de genes en la red (los valores se refieren a la red de genes completa)
2 Nº de vinculos de coexpresión (los valores se refieren a la red de genes completa)
3 Verdaderos Positivos = parejas de genes que coexpresan y están anotados al mismo KEGG
4 Todos los genes que coexpresan y están anotados a KEGG
5 VPP de las redes, derivados de la anotación delos genes a KEGG

Tabla 3.4: Comparativa de redes de coexpresión en función de su precisión.
Comparativa de la precisión (**VPP**) de las redes generadas en este estudio frente a las redes de *Lee et al.* y de *Griffith et al.*.

resultados y esta puede ser la razón de la aparición de rutas biomoleculares como infección patógena en los datos de *Lee*; (**v**) finalmente nuestros datos también incluyen mucha coexpresión entre genes involucrados en comunicación celular como son interacciones entre citoquinas y ligandos.

Para hacer la comparativa de fiabilidad, se tomaron las 3 redes de coexpresión estudiadas de forma que tuvieran un número similar de relaciones de coexpresión (unas 12000 relaciones de coexpresión) y se calculó el **VPP** para cada *set*. El resultado de esta comparativa se muestra en la **tabla 3.4**, en la que se ve que nuestra red dió un valor de fiabilidad mayor que las otras redes. Nuestros datos obtuvieron un **VPP** de 0.61, mientras que los otros datos obtuvieron un **VPP** de 0.56 para *Lee* y de 0.49 para *Griffith*. En general estos números dan una idea de que la red de coexpresión obtenida en este trabajo es una red muy coherente de genes que tienden a estar involucrados en un mismo *pathway*. Además hay que tener en cuenta el número de *microarrays* usados para cada estudio: mientras que en *Lee* se usaron 3924 *microarrays* provenientes de 60 estudios y para *Griffith* se integraron un total de 2111 *microarrays* provenientes de diversas tecnologías de *arrays*; para nuestros resultados solo se emplearon 48 *microarrays* cuidadosamente seleccionados y tratados con el método de validación cruzada desarrollado.

3.2.8 Descubrimiento de información funcional y transcripcional en redes de coexpresión

Como hablamos en la introducción, en este estudio abordamos la estrategia de ingeniería reversa llevando a cabo tanto la aproximación física como la deductiva. Para demostrar que nuestro método tiene capacidad de generar datos adecuados para

definir relaciones funcionales entre genes e identificar factores de transcripción, se optó por usar un *set* de coexpresión muy fiable. Éste se obtuvo haciendo la intersección de los *sets* de coexpresión calculados por las dos aproximaciones con un **VPP >= 0.60** (ver **tabla 3.2**). De esta intersección se obtuvo la red de coexpresión que se va a analizar funcionalmente, compuesta por 731 genes y 2249 relaciones de coexpresión de la que se eliminaron las subredes que solo incluyen dos nodos quedando 615 genes y 2190 relaciones de coexpresión. En la **figura 3.16** se representa la red de coexpresión obtenida. Los principales grupos conectados están resaltados en diferentes colores y tienen etiquetas con una función general asignable a la mayoría de sus genes. De esta forma se puede ver que la subred más grande se corresponde con genes involucrados en actividad nuclear y metabolismo relacionado con el núcleo (región azul), con una parte (en azul oscuro) que incluye la mayoría de las proteínas ribosómicas y proteínas involucradas en la función ribosomal. El segundo grupo más grande (en verde) incluye muchos genes involucrados en metabolismo mitocondrial y homeostasis reducción-oxidación (*redox*) (con genes de las familias *COX*, *NDUF* y *UQCR*). La tercera región más grande (en rojo) se corresponde con genes involucrados en respuesta inmune, genes del complejo mayor de histocompatibilidad (*MHC*), genes que produce la superficie celular como los *clusters* de diferenciación (*CD*) y genes que codifican moléculas específicas de antígenos. Finalmente, hay otras regiones más pequeñas pero bien definidas que incluyen genes involucrados en homeostasis de iones metálicos (en gris), relacionados con la matriz extracelular y adhesión celular (naranja) o relacionados con la estructura del citoesqueleto (en amarillo).

En el apartado 2.5 de este capítulo, se vio el predominio de genes *housekeeping* a medida que los parámetros **N** y **r** van siendo más astringentes. Ésto también se comprobó identificando una gran cantidad de genes *housekeeping* esenciales en la red representada en la **figura 3.16**. Para ello se usó el conjunto de genes *housekeeping* definido por *Hsiao* y un conjunto de genes ortólogos de humano que se han identificado como esenciales en levadura (obtenidos de la base de datos *SGD* (Hong, *et al.*, 2008)). De esta forma encontramos que los dos principales grupos de la red, relacionados con el metabolismo redox mitocondrial y con funciones nucleares, tienen un porcentaje de un 63% y de un 58% respectivamente de genes identificados como *housekeeping*. Este resultado revela que la red está claramente enriquecida con genes esenciales.

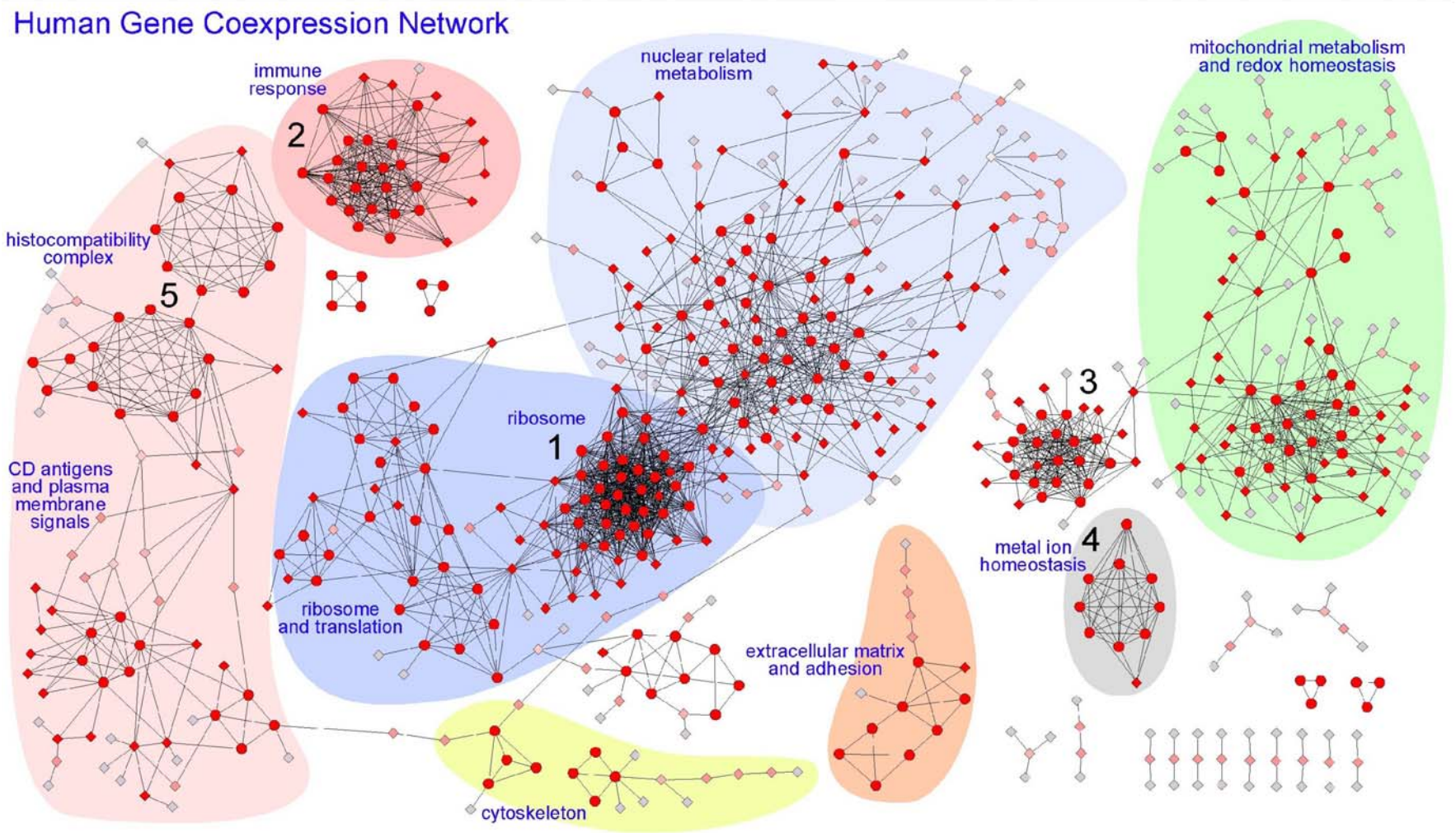


Figura 3.16: Red fiable de coexpresión analizada funcionalmente.
Representación de la red de coexpresión producida con la intersección de conjuntos de datos provenientes de los dos métodos (*RMA-Pearson* y *MAS5-Spearman*) con **VPP ≥ 0.6**. Se eliminaron las parejas de genes que solo estaban conectados entre ellos y quedó una red con 615 genes y 2190 interacciones de coexpresión. Las regiones más significativas se han etiquetado con la función asignable a la mayoría sus genes y se han resaltado con colores de fondo. Los nodos tienen diferentes colores (de rojo a gris) y formas (círculos o rombos). La intensidad del color rojo está en función de la **conectividad** del gen, cuanto más conectado esté el gen mas intenso es el rojo. La forma de los nodos está en función de su **coeficiente de *clustering***, los de bajo coeficiente son rombos y los de alto círculos.

En conclusión, la coherencia funcional observada en las principales regiones de la red y la gran presencia de genes *housekeeping*, nos prueban la validez de las relaciones de coexpresión encontradas, que en muchos casos van a revelar procesos biológicos esenciales u omnipresentes. Esta red, como veremos posteriormente, también nos revela relaciones entre genes no descritas.

Para deducir relaciones funcionales y transcripcionales, hay que reconocer grupos de genes altamente conectados en la red de coexpresión. Estos grupos fueron extraídos aplicando algoritmos de *clusterización* de redes. Trabajamos principalmente con dos algoritmos: *MCODE* (Bader y Hogue, 2003) y *MCL* (Enright, *et al.*, 2002). Los parámetros sobre los que trabaja *MCODE* se muestran en la **figura 4.16** en la que los nodos circulares son los que tienen un **coeficiente de *clusterización*** (**cc**) alto, mientras que los nodos en forma de rombo son los de **coeficiente de *clusterización*** bajo. La intensidad del color rojo indica la **conectividad** (**cn**) de los nodos de forma que cambia de rojo intenso a gris en función de tener una conectividad alta o baja. Con estos parámetros (**cc** y **cn**) *MCODE* divide la red en grupos que tienen genes muy conectados entre si. En la **figura 4.16** se han numerado los 5 grupos más importantes encontrados por *MCODE* en la red de coexpresión: (***clúster* 1**) corresponde a genes ribosomales, tiene 29 nodos y 366 interacciones, siendo la mayoría de genes *RPL* o *RPS*; (***clúster* 2**) corresponde a inmunoglobulinas y respuesta inmune (genes de las familias IGH, IGK e IGL) y tiene 19 nodos y 151 interacciones; (***clúster* 3**) es un grupo heterogéneo muy agrupado sin una aparente correspondencia funcional, tiene 19 nodos y 140 interacciones; (***clúster* 4**) se corresponde con genes relacionados con homeostasis de iones metálicos (muchos *MT1* y *MT2*) y tiene 9 nodos y 36 interacciones; (***clúster* 5**) en él la mayoría de los genes están relacionados con el complejo mayor de histocompatibilidad (*MCH*), tiene 17 nodos (la mayoría *HLA*) con 63 interacciones.

Esta *clusterización* fue verificada aplicando el algoritmo *MCL* que proporcionó resultados similares para los 5 grupos antes comentados, sin embargo *MCL* dividió también la red en grupos de menos genes que tenían mayor coherencia funcional que los generados por *MCODE*. Por ejemplo, *MCL* encontró un grupo con 15 genes de los cuales 7 estaban relacionados con unión a *RNA* y 3 con unión a *DNA* (todos ellos pertenecientes a la región azul), incluyendo 3 miembros de la familia *HNRP* (proteínas relacionadas con el ensamblaje del nucleosoma) y 2 iniciadores de traducción.

Genes desconocidos	*Cluster* de coexpresión	Genes que tiene conectados en el cluster de coexpresión	
LOC440055	Ribosome (1)	EEF1B2	Elongation factor 1-beta
		HNRPA1	Heterogeneous nuclear ribonucleoprotein A1
		MDS1	Myelodysplasia syndrome 1 protein
		RPL11	60S ribosomal protein L11
		RPL17	60S ribosomal protein L17
		RPL21	60S ribosomal protein L21
		RPL22	60S ribosomal protein L22
		RPL23	60S ribosomal protein L23
		RPL30	60S ribosomal protein L30
		RPL31	60S ribosomal protein L31
		RPL32	60S ribosomal protein L32
		RPL35A	60S ribosomal protein L35a
		RPL36A	60S ribosomal protein L36a
		RPL37	60S ribosomal protein L37
		RPL39	60S ribosomal protein L39
		RPL4	60S ribosomal protein L4
		RPL9	60S ribosomal protein L9
		RPS15A	40S ribosomal protein S15a
		RPS17	40S ribosomal protein S17
		RPS20	40S ribosomal protein S20
		RPS23	40S ribosomal protein S23
		RPS3A	40S ribosomal protein S3a
		RPS4X	40S ribosomal protein S4, X isoform
		RPS6	40S ribosomal protein S6
		RPS7	40S ribosomal protein S7
LOC645745	Metal Ion Homeostasis (4)	MT1E	Metallothionein-1E
		MT1F	Metallothionein-1F
		MT1G	Metallothionein-1G
		MT1H	Metallothionein-1H
		MT1L	Metallothionein-1L
		MT1M	Metallothionein-1M
		MT1X	Metallothionein-1X
		MT2A	Metallothionein-2
LOC91316	Inmune Response (2)	IGHA1	Ig alpha-1 chain C region
		IGHG3	Ig gamma-3 chain C region
		IGHM	Ig mu chain C region
		IGKC	Ig kappa chain C region
		IGLC1	Ig lambda chain C regions
		IGLC2	Ig lambda chain C regions
C6orf12	Histocompatibility Complex (5)	HLA-A	HLA class I histocompatibility antigen, A alpha chain
		HLA-B	HLA class I histocompatibility antigen, B alpha chain precursor
		HLA-C	HLA class I histocompatibility antigen, Cw alpha chain precursor
		HLA-F	HLA class I histocompatibility antigen, alpha chain F precursor
		HLA-G	HLA class I histocompatibility antigen, alpha chain G precursor
FAM96B	Mitochondrial Metabolism and Redox H. (9)	NDUFA10	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 10, mitochondrial precursor
		NDUFA2	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 2
		NDUFA3	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 3
		NDUFS3	NADH dehydrogenase [ubiquinone] iron-sulfur protein 3, mitochondrial precursor
		NDUFS7	NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial precursor
		NDUFS8	NADH dehydrogenase [ubiquinone] iron-sulfur protein 8, mitochondrial precursor
		NDUFV1	NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial precursor
		TUFM	Elongation factor Tu, mitochondrial precursor
LOC391020	Quartet under cluster 2	IFITM1	Interferon-induced transmembrane protein 1
		IFITM2	Interferon-induced transmembrane protein 2
		IFITM3	Interferon-induced transmembrane protein 3
LOC388344	Ribosome and Translation (7)	GLTSCR2	Glioma tumor suppressor candidate region gene 2 protein
		RPL13	60S ribosomal protein L13
		RPL18	60S ribosomal protein L18
		RPL18A	60S ribosomal protein L18a
		RPS15	40S ribosomal protein S15
		RPS2	40S ribosomal protein S2

Tabla 3.5: Anotación funcional de genes desconocidos.
Ejemplos de genes que tienen una anotación con función desconocida en los ficheros de *Affymetrix* y que sin embargo, su función puede deducirse por los genes con los que están relacionados en la red de coexpresión.

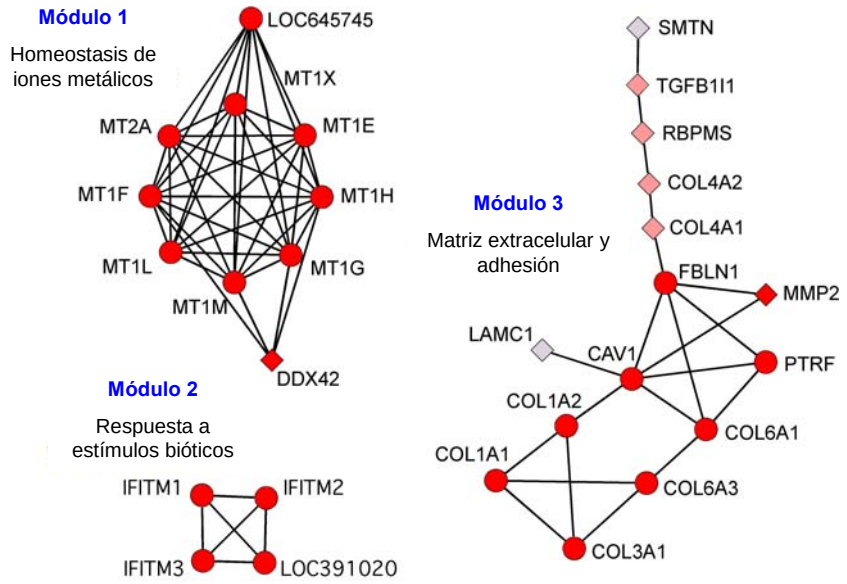
La identificación de grupos o *clusters* de genes muy conectados dentro de la red de coexpresión, puede ayudar a comprender mejor la función de genes humanos y las relaciones entre ellos. Este es el objetivo de la aproximación deductiva explicada en la introducción, que se ha desarrollado estableciendo relaciones asociativas entre los genes y posibilitando la inferencia de anotación funcional para los genes desconocidos incluidos en los grupos. De este modo se puede deducir la función de un gen desconocido, si éste queda relacionado con un grupo de genes de función conocida muy interconectado. Unos ejemplos de esto se muestran en la **tabla 3.5**.

Como se ha explicado anteriormente, la aproximación física de la ingeniería reversa trata de identificar moléculas que regulan la síntesis de m*RNA* y con ello la expresión de genes concretos. Por ello, exploramos la similitud de elementos regulatorios entre genes coexpresados. Se analizaron 2 grupos de genes muy conectados presentes en la red de coexpresión. El primer grupo tiene 10 genes de los cuales 8 forman una

estructura octogonal con los vértices totalmente relacionados. Los 8 genes se corresponden con metalotioninas: *MT1E*, *MT1F*, *MT1G*, *MT1H*, *MT1L*, *MT1M*, *MT1X*, *MT2A*. Los otros dos genes relacionados no están muy anotados: *DDX42* (es un miembro de la familia *DEAD box* que no tiene una función clara) y *LOC645745* (que ha sido recientemente identificado como una proteína putativa *MT1*). La coexpresión de estos dos genes con el cluster de metalotieninas permite inferir que su función debería estar relacionada con homeostasis de iones metálicos. Este módulo puede verse en la **figura 3.17**. Para ver si estos genes tienen algún factor de trascripción común que actúe sobre sus promotores y regiones reguladoras, se usaron dos herramientas bioinformáticas que buscan factores de transcripción comunes a grupos de genes: *PAP* (Chang, *et al.*, 2007) y *Factory* [33]. *PAP* encontró que los genes coexpresados en este grupo son regulados por un factor de transcripción común *MTF1* (pvalue = 0.001). Este resultado es muy probable ya que *MTF1* es un factor de transcripción que induce la expresión de metalotioninas y de otros genes involucrados en homeostasis de iones metálicos (como zinc y cobre). La asociación de *MTF1* con este módulo es muy coherente y muestra que esta red de coexpresión está relacionada con entidades de corregulación de transcripción inherentes.

El segundo módulo que se muestra en la **figura 3.17** tiene 4 genes de los cuales 3 se corresponden con proteínas transmembrana interferón-inducible (*IFITM1*, *IFITM2*, *IFITM3*) y el otro es un gen desconocido nombrado como *LOC391020*, que ha sido recientemente anotado por inferencia como un gen similar a *IFITM3*. Estos genes se analizaron con *PAP* y *Factory* en busca de factores de transcripción comunes y ambos métodos coincidieron en que el factor de transcripción *CRE-BP1* (También llamado *ATF2*, factor de transcripción activador 2) está relacionado con estos genes. ATF2 es una proteína que se une al promotor *CRE* y forma un homodímero o un heterodímero con *JUN*. La suposición de que los genes *IFITM* son regulados por *ATF2* es muy sensata, ya que se ha observado que para la activación transcripcional de genes relacionados con interferón, se necesita el ensamblaje de un complejo de activación formado por *ATF2* y *JUN* (Falvo, *et al.*, 2000; Panne, *et al.*, 2004).

Finalmente, el tercer módulo de la **figura 3.17** incluye 15 genes de los cuales 6 son colágenos (*COL1A1*, *COL1A2*, *COL3A1*, *COL4A1*, *COL4A2*, *COL6A1*), es decir proteínas fibrilares que están en la mayoría de los tejidos conectivos y que conforman su matriz extracelular. Otras proteínas de este módulo, también relacionadas con



	Buscando en	Resultado	p-valor	TransFac ID	Nombre del Factor de Transcripción
Módulo 1 10 genes	*PAP*	MTF-1	**0.001**	T02354	**MTF1** metal-regulatory transcription factor 1
	Factory	–	**–**		
Módulo 2 4 genes	*PAP*	CRE-BP1	**0.0172**	T00167	**ATF2** activating transcription factor 2
	Factory	CRE-BP1	**0.0033**		
Módulo 3 15 genes	*PAP*	Sp1	**0.13**	T00759	**SP1** Sp1 transcription factor
	Factory	Sp1	**0.017**		

Búsqueda de factores de transcripción en módulos muy conectados.
(**A**) Ampliación de 3 módulos muy conectados extraídos de la red de la **figura 4.16**. **Módulo 1**) homeostasis de iones metálicos, **Módulo 2**) respuesta a estímulos bióticos, **Módulo 3**) matriz extracelular y adhesión. (**B**) Tabla que muestra los resultados de buscar factores de transcripción que comparten la mayoría de los genes incluidos en cada módulo. Esta búsqueda se hizo mediante las herramientas bioinformáticas *PAP* y *Factory*.

adhesión celular y matriz extracelular son: Fibulin 1 (*FBLN1*), que es una glicoproteina segregada para ser incorporada a la matriz extracelular fibrilar; Laminina gamma 1 (*LAMC1*) que es otra glicoproteína de la matriz extracelular de tipo no colágeno, constitutiva de la mayoría de las membranas; y por último la metaloproteinasa de matriz 2 (*MMP2*) que participa en la ruptura de la matriz extracelular, y codifica una enzima que degrada el colágeno tipo IV. El análisis de estos 15 genes muestra una asociación significativa con el factor de transcripción *SP1* (tanto por *PAP* como por *Factory*). Esta observación ha sido validada por datos experimentales recientes, que han descubierto que el factor de transcripción *SP1* está involucrado en la regulación de promotores de colágeno (Kypriotou, *et al.*, 2007; Magee, *et al.*, 2005; Poree, *et al.*, 2008).

Resultados similares a los descritos pueden obtenerse de la mayoría de los módulos de la red, indicando que la red de coexpresión agrupa genes que suelen estar ligados a factores de transcripción que regulan su actividad transcripcional. Esta observación da una clara coherencia funcional y biológica a la red de coexpresión generada.

3.3 CONCLUSIONES

Los datos de **perfiles de expresión génica** están siendo ampliamente utilizados en investigaciones bioinformáticas gracias al gran desarrollo de la tecnología de *microarrays* que miden expresión a nivel genómico global. Sin embargo, debido al ruido de estas técnicas es necesario el desarrollo de métodos robustos que generen **redes de coexpresión** fiables. Estos métodos deben considerar todos los pasos que se tienen que realizar para obtener perfiles de expresión correlacionados (selección de muestras, cálculo de señal, búsqueda de correlaciones binarias), ya que los resultados son muy sensibles a un cambio en cualquiera de estos pasos. De hecho, la variedad de métodos de cálculo de señal de *microarrays* de expresión, hace que los resultados se deban tratar de forma diferente dependiendo del método o algoritmo empleado. También, hemos comprobado que la selección de muestras es importante y no debe incluir estados alterados que provoquen ruido o sesgo en los resultados, ya que el uso de muestras alteradas dificulta la detección adecuada de perfiles correlacionados propios de estados normales.

Respecto a la comparación de perfiles de expresión, hemos comprobado que los resultados obtenidos al aplicar distintos **métodos de búsqueda de coexpresión** se pueden combinar para aumentar la cobertura y precisión de los datos. Así en nuestro estudio hemos puesto a punto los métodos *RMA-Pearson* y *MAS5-Spearman*, que generan resultados algo diferentes no solapantes, con una buena fiabilidad, siendo por ello susceptibles de combinarse. Además hemos propuesto un filtro que se puede aplicar sobre los datos crudos de expresión para el método *RMA-Pearson*, con objeto de mejorar la astringencia y fiabilidad de los resultados obtenidos al buscar perfiles correlacionados con dicho método. Este filtro no es necesario aplicarlo en datos calculados con el método *MAS5-Spearman* que es más estricto en la búsqueda de correlaciones. Por último, para el cálculo final de las correlaciones más fiables hemos utilizado **validación cruzada**, que permite obtener datos de coexpresión robustos eliminando las relaciones que no se distinguen significativamente de las encontradas al azar.

Como métrica para **estimar la fiabilidad** de las redes de coexpresión nos hemos basado en el axioma de que "genes que trabajan juntos tendrán una mayor tendencia a estar coregulados o coexpresados". Así hemos comprobado que el mapeo y asignación

de los genes a rutas biomoleculares y vías metabólicas definidas permite hacer una buena estimación del número de verdaderos positivos en las redes de coexpresión.

Finalmente se ha demostrado como la obtención de redes de coexpresión fiables, permite el descubrimiento de **información biológica funcional y transcripcional**. Estas redes son útiles para reanotar funcionalmente genes desconocidos basándose en el entorno de la red en el que están y facilitan la búsqueda de factores de transcripción comunes a grupos de genes muy conectados entre sí dentro de la red.

Capítulo 4

Búsqueda de alteraciones y desregulación génica en perfiles de expresión correlacionados

4.1 INTRODUCCIÓN

La expresión genómica es un proceso fuertemente regulado que es crucial para el funcionamiento adecuado de una célula. En datos de *microarrays*, la regulación común se refleja por correlaciones fuertes entre niveles de expresión (Eisen, *et al.*, 1998) y es habitual que genes con funciones biológicas relacionadas tengan un perfil de expresión similar en diversas condiciones (Segal, *et al.*, 2003; Stuart, *et al.*, 2003). Por otro lado, es frecuente que los mecanismos moleculares de una enfermedad

provoquen anormalidades en la regulación de genes produciendo fuertes alteraciones en los niveles de expresión (Rhodes, *et al.*, 2002). Los cambios en los perfiles de expresión pueden ayudar a identificar genes desregulados relacionados con enfermedad (Golub, *et al.*, 1999) y así facilitar mejoras en el diagnóstico y en el pronóstico, basadas en el descubrimiento de dichos genes alterados (West, *et al.*, 2001).

La alteración de la regulación genómica suele provocar perfiles de expresión con gran variabilidad. Sin embargo, las estrategias de análisis habituales (Tusher, *et al.*, 2001) examinan generalmente genes diferencialmente expresados (sobreexpresados o reprimidos), pero no exploran la variabilidad que aparece en un estado alterado cuando se compara con un estado control.

En el transcurso de este capítulo se explicará el desarrollo de un algoritmo nuevo que permite analizar y comparar los perfiles de expresión de genes en muestras normales control contra perfiles en muestras alteradas, encontrando de esta forma los genes que sufren una fuerte desregulación o alteración en su variabilidad. El algoritmo está especialmente adaptado y desarrollado para trabajar con *microarrays* de oligonucleotidos de alta densidad, en particular con los *GeneChips* (manufacturados por *Affymetrix*), y proporciona una lista de grupos de genes desregulados con un valor de significación estadística para cada grupo.

4.2 MÉTODOS Y RESULTADOS

4.2.1 Varianza residual relativa como medida de pérdida de coexpresión y de alteración

El objetivo de esta parte del trabajo es encontrar grupos de genes que tienen un patrón de expresión similar en las muestras control y que sufren diversas alteraciones en los perfiles de expresión en muestras alteradas, que normalmente corresponden con estados patológicos.

Cada conjunto de microarrays, que implica a **I** genes en **J** muestras, puede representarse matemáticamente como una matriz **M** bidimensional **|I| x |J|** con elementos $\mathbf{e_{ij}}$, donde $\mathbf{e_{ij}}$ es la intensidad de expresión del gen **i** en la muestra **j**. De esta forma, cada gen tiene un perfil de expresión a lo largo de las muestras que define su estado. Para un subconjunto de genes $\mathbf{G \subset I}$ y un subconjunto de muestras $\mathbf{S \subset J}$, la submatriz **G x S** representa las intensidades de los genes **G** en las muestras **S**. Debido a que el objetivo es comparar solo dos estados entre las muestras, el estado control contra el estado alterado o enfermo, las muestras están siempre divididas en dos subconjuntos: $\mathbf{S_a}$ (muestras alteradas) y $\mathbf{S_c}$ (muestras control). Por tanto:

$$|\mathbf{J}| = |\mathbf{S_c}| + |\mathbf{S_a}|$$

Por cada valor de expresión $\mathbf{e_{ij}}$, una vez normalizado y corregido el *background*, se puede considerar un modelo aditivo que divide conceptualmente el valor de expresión en un conjunto de cuatro elementos diferentes: efecto del gen (α_i), efecto de la muestra (β_j), un valor constate común a todos los datos (γ) y el error (ε_{ij}). Por tanto:

$$\mathbf{e_{ij}} = \alpha_i + \beta_j + \gamma + \varepsilon_{ij}$$

Una asunción que habitualmente se hace en la mayoría de los algoritmos que trabajan con valores de expresión es que los errores son independientes y que sus varianzas son iguales a lo largo de los genes y las muestras. Bajo el modelo aditivo definido anteriormente, los estimadores de máxima similitud se aplicarán a la media global de todos los valores de expresión en todas las muestras ($\bar{e}_{\mathbf{..}}$) para los efectos comunes (γ); a la diferencia entre las medias de los valores de expresión de las muestras en el gen **i** ($\bar{e}_{i\mathbf{.}}$) y la media global ($\bar{e}_{\mathbf{..}}$) para los efectos de los genes (α_i); y a la diferencia entre las

medias de los valores de expresión de los genes en la muestra **j** ($\bar{e}_{\bullet j}$) y la media global ($\bar{e}_{\bullet\bullet}$) para los efectos de las muestras (β_j). Por tanto el estimador de máxima similitud para el error $\varepsilon_{ij} = e_{ij} - \alpha_i - \beta_j - \gamma$ será:

$$e_{ij} - (\bar{e}_{i\bullet} - \bar{e}_{\bullet\bullet}) - (\bar{e}_{\bullet j} - \bar{e}_{\bullet\bullet}) - \bar{e}_{\bullet\bullet} = e_{ij} - \bar{e}_{i\bullet} - \bar{e}_{\bullet j} + \bar{e}_{\bullet\bullet}$$

Usando estos estimadores el objetivo es comparar todos los valores de expresión e_{ij} de la matriz **M** para encontrar variabilidad y correlación entre genes en las distintas muestras. Para encontrar esa variabilidad en los valores de expresión, se tiene que calcular un parámetro estadístico de variabilidad. Un parámetro obvio es el coeficiente de correlación estándar, pero varios autores han encontrado que la **varianza residual** (**VR**) permite una búsqueda más eficiente de grupos de genes candidatos (Cheng y Church, 2000; Kostka y Spang, 2004). Como estos autores, nosotros elegimos la varianza residual **VR(G,S)** como parámetro para medir la variabilidad entre un subgrupo de genes **G** en un subgrupo de muestras **S**. Este parámetro se define como:

$$VR(G,S) = \frac{1}{(|G|-1)(|S|-1)} \sum_{G,S} (e_{ij} - \bar{e}_{i\bullet} - \bar{e}_{\bullet j} + \bar{e}_{\bullet\bullet})^2$$

Del modelo lineal, el sumatorio formado sigue una distribución chi-cuadrado de *Pearson* con (**|G|-1**) y (**|S|-1**) grados de libertad (Draghici, 2003; Montgomery, 2008). De esta forma, la **varianza residual (VR)** es el *ratio* entre una distribución chi-cuadrado de *Pearson* y sus grados de libertad. El siguiente paso en el algoritmo es implementar la posibilidad de comparar las varianzas residuales de dos subgrupos independientes y diferentes de muestras alteradas y control ($\mathbf{S_a}$ y $\mathbf{S_c}$), en un subset de genes *G* dado. Para ambos subgrupos calculamos $\mathbf{VR_c} = \mathbf{VR(G,S_c)}$ y $\mathbf{VR_a} = \mathbf{VR(G,S_a)}$, y la comparación se hizo con la **varianza residual relativa** (**VRR**) que es el *ratio* entre la varianza residual de los controles ($\mathbf{VR_c}$) y la varianza residual de las muestras alteradas ($\mathbf{VR_a}$):

$$VRR(G,S) = \frac{(|G|-1)(|S_a|-1)}{(|G|-1)(|S_c|-1)} \frac{\sum_{G,S_c} (e_{ij_c} - \bar{e}_{i\bullet_c} - \bar{e}_{\bullet j_c} + \bar{e}_{\bullet\bullet_c})^2}{\sum_{G,S_a} (e_{ij_a} - \bar{e}_{i\bullet_a} - \bar{e}_{\bullet j_a} + \bar{e}_{\bullet\bullet_a})^2}$$

4.2.2 Significación estadística de los grupos de genes alterados

En la varianza residual relativa la división de dos distribuciones chi-cuadrado de *Pearson* sigue una distribución de *Fisher-Snedecor* ($\mathbf{F_{n1,n2}}$) con **(|G|-1)*(|S$_c$|-1)** y **(|G|-**

1)*(|S_a|-1) grados de libertad (Montgomery, 2008). El ajuste a una distribución de probabilidad permite construir un test estadístico de hipótesis para cada subgrupo de genes **G**, donde la hipótesis nula $\mathbf{H_0}$ será la no existencia de diferencia en el parámetro de variabilidad (**VR**) de expresión entre los controles ($\mathbf{VR_c}$) y las muestras alteradas ($\mathbf{VR_a}$) para un subgrupo de genes **G**. En términos estadísticos, la no diferencia se corresponde a la situación donde la varianza residual de las muestras alteradas $\mathbf{VR_a}$, es menor o igual que la varianza residual de las muestras control $\mathbf{VR_c}$. Por otro lado, la hipótesis alternativa $\mathbf{H_1}$ será la existencia de una diferencia significativa entre las muestras control y las alteradas. Como estimador de varianza usamos la **varianza residual** (**VR**):

$\mathbf{H_0}$: $\sigma_a^2 \leq \sigma_c^2$: el subconjunto de genes **G** no presenta una diferencia de variabilidad en las muestras alteradas frente a las control.

$\mathbf{H_1}$: $\sigma_a^2 > \sigma_c^2$: el subconjunto de genes **G** presenta una diferencia de variabilidad en las muestras alteradas frente a las control.

Esto es un test de una cola (*one-tailed test*) donde la región de rechazo es **{F ≤ k}** para un valor crítico **k**. De este modo, se puede calcular un valor de probabilidad correspondiente al valor del estadístico **F**, y dicho valor se toma como indicador de la diferencia de variabilidad entre las muestras control y alteradas de un subconjunto de genes **G: F-score = P{F≤Fobs}**. Este valor se calcula para cada subconjunto de genes **G** y permite evaluar progresivamente la significación estadística de los grupos, de forma que será mayor cuanto más se alejen de la hipótesis nula.

4.2.3 Diseño e implementación del algoritmo *AlteredExpression*

Siguiendo la **varianza residual relativa** (**VRR**) definida previamente, escribimos un algoritmo en R [25] llamado *AlteredExpression*. Un esquema detallado del funcionamiento del algoritmo esta representado en la **figura 5.1** en la que se especifican todos los pasos que realiza en un diagrama de flujo. El algoritmo fue diseñado para ejecutarse recibiendo como entrada un conjunto de *microarrays* de expresión con **I** genes y **J** muestras subdivididas en dos clases: muestras de control $\mathbf{S_c}$ y muestras alteradas $\mathbf{S_a}$. Como salida, en cada ejecución del algoritmo se da el grupo de genes con mejor varianza residual relativa. Este grupo se va generando progresivamente de forma que partimos de un grupo de genes $\mathbf{G_1}$ elegido al azar, al

que se van a ir añadiendo o quitando genes progresivamente, para formar un grupo $\mathbf{G_n}$ que mejore la varianza residual relativa del anterior. Este proceso continua hasta que la varianza residual relativa de los grupos generados no cambia significativamente, es decir se estabiliza alcanzándose un óptimo. El algoritmo tiene definidos unos parámetros de entrada que se pueden modificar en función del tipo de muestras analizadas. Estos parámetros son los siguientes:

- *initialSize*: indica el número de genes que se van a elegir de forma aleatoria para formar el grupo inicial de genes $\mathbf{G_1}$. Tras explorar varios tamaños, vimos que con un conjunto inicial de 20 genes el algoritmo se comportaba bien en la mayoría de los casos. Las ejecuciones hechas con valores de *initialSize* diferentes probaron que este parámetro no afecta a los resultados.
- *maxiter*: indica el número máximo de iteraciones que se permiten en el algoritmo para encontrar un grupo óptimo. Este parámetro es sólo para evitar la ejecución del algoritmo sin interrupción y está inicialmente fijado en 500, que es un número de ejecuciones más que razonable para obtener el óptimo.
- *vSize*: modula el efecto del tamaño de grupo. Varía entre 0 y 1, siendo mayor el efecto del tamaño cuando es cercano a 0 y menor cuando está próximo a 1. El *vSize* esta fijado a 0.5 por defecto.
- *pctSubset*: es un parámetro, que varía entre 0 y 1, introducido para fijar el porcentaje de genes que entran o salen del grupo $\mathbf{G_n}$ para optimizar la varianza residual relativa en cada iteración del bucle externo (*loop* 1). En el diagrama de flujo *ioGenes* son todos los genes que han mejorado individualmente la varianza residual relativa en el bucle 2 (*loop* 2). Que el parámetro *pctSubset* tenga un valor 0.1 por defecto, significa que el 10% de los genes de *ioGenes* son cambiados en $\mathbf{G_n}$ en cada iteración del bucle 1 (**figura 4.1**).

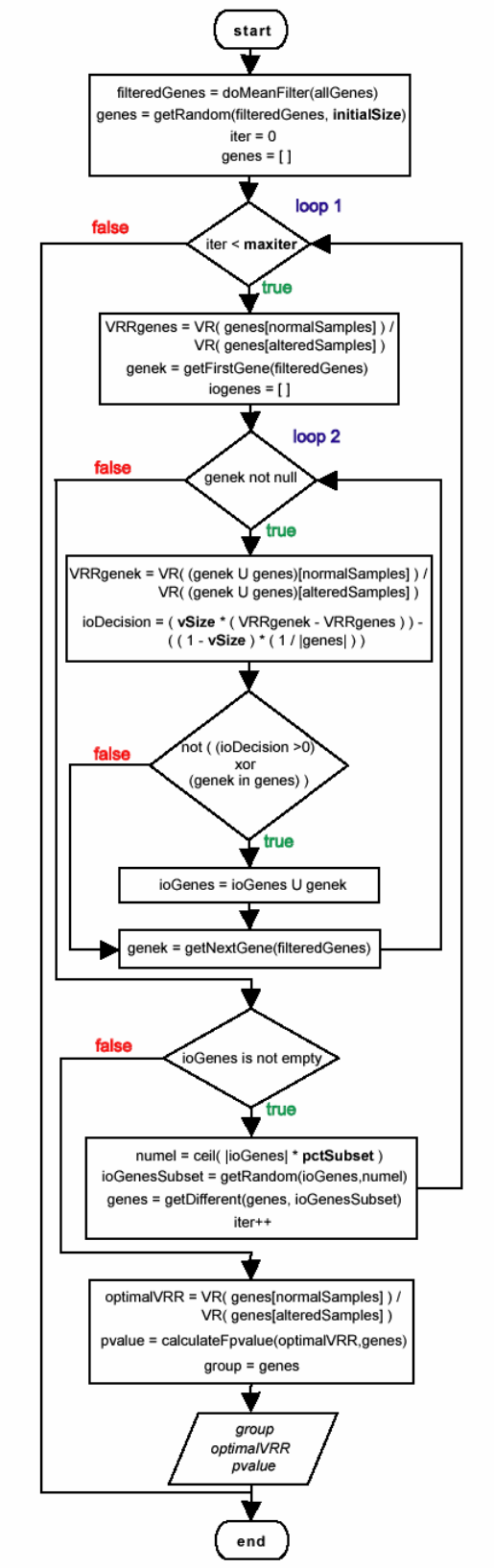
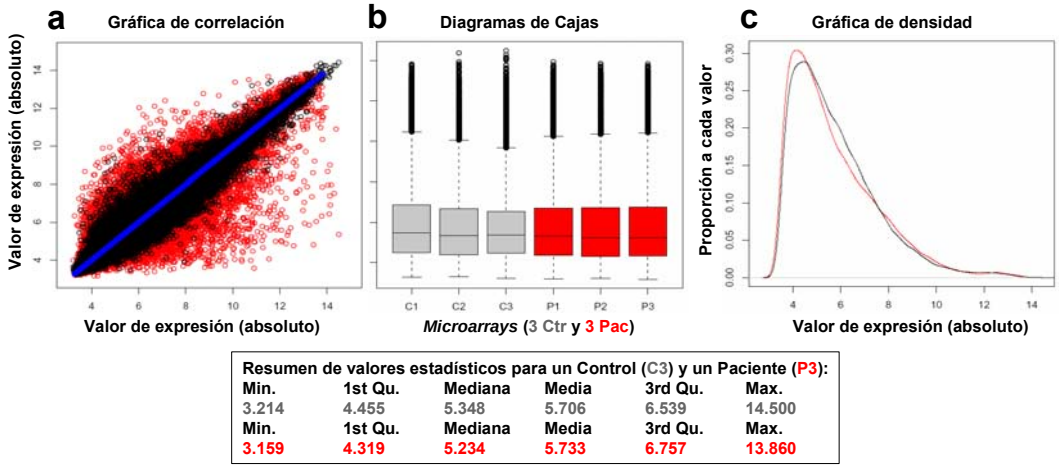


Figura 4.1: Diagrama de flujo de *AlteredExpression*:
Muestra los pasos que realiza el algoritmo para seleccionar un grupo de genes a partir de una matriz de datos inicial. Como entrada se pueden pasar los parámetros: *initialSize*, *maxiter*, *vSize*, *pctSubset*. Como salida devuelve un grupo estable de genes con la *VRR* del grupo y su *F-score*.



Resumen de valores estadísticos para un Control (C3) y un Paciente (P3):

Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
3.214	4.455	5.348	5.706	6.539	14.500
Min.	**1st Qu.**	**Mediana**	**Media**	**3rd Qu.**	**Max.**
3.159	4.319	5.234	5.733	6.757	13.860

Figura 4.2: Cálculo de señal de los datos de entrada.
Representación de algunos de los *microarrays* normalizados con los que se ha trabajado: datos de expresión de muestras de control (**C**) y de pacientes con *APL* (**P**). **(a)** Gráfica de correlación de los valores de expresión de 3 controles contra 3 pacientes (puntos **rojos**), 3 controles contra 3 controles diferentes (puntos **negros**), un control contra si mismo (puntos **azules**). **(b)** Diagramas de cajas (*boxplots*) que muestran la distribución de los datos de 3 controles (**C1**,**C2**,**C3**) y 3 pacientes (**P1**,**P2**,**P3**). **(c)** Gráficos de densidad (*density plots*) de las distribuciones acumuladas de 3 controles (línea **negra**) y 3 pacientes (línea **roja**). Debajo se muestra un resumen de los parámetros estadísticos calculados para la muestra control **C3** y la alterada **P3**.

4.2.4 Puesta a punto del algoritmo con muestras de cáncer

Con el algoritmo diseñado e implementado, quedaba evaluar y validar si se comporta correctamente al aplicarlo a muestras de *microarrays* de expresión. Se seleccionaron 16 muestras (*Affymetrix GeneChip HGU133a*) que provenían de biopsias de médula ósea de 16 personas diferentes: 6 sanos y 10 pacientes con leucemia promielocítica aguda (APL). Se usó el algoritmo *RMA* para corregir el *background*, hacer la normalización y calcular los valores de expresión de las muestras. La complejidad del conjunto de *microarrays* se puede ver en la **figura 4.2a** que representa la correlación de los valores de expresión para cada gen (cada *probeset* en el microarray) en un subconjunto de 6 muestras: tres pacientes y tres controles. Se hicieron 9 comparaciones tomando las posibles parejas de 3 pacientes con 3 controles (puntos rojos), 9 comparaciones tomando las posibles parejas de 3 controles con otros 3 controles (puntos negros) y una comparación de un control contra si mismo (línea azul central). Los puntos negros muestran el grado de variabilidad entre individuos que es menor que en el caso de los puntos en rojo que muestran las diferencias entre los datos control y los pacientes. Las **figuras 4.2b y c** representan diagramas de cajas (*boxplots*) y gráficos de densidad (*density plots*) de las distribuciones de expresión de las 3 muestras control y 3 de pacientes que se compararon en la **figura 4.2a**. Los números

corresponden a una transformación logarítmica en base 2. Se puede observar en las representaciones de densidad, que los valores de expresión no siguen una distribución normal. Otra observación son las pequeñas diferencias entre las distribuciones de diferentes muestras, incluso entre muestras control y de pacientes. Este hecho indica que se ha hecho una buena normalización que hace posible usar estos datos en la búsqueda de genes desregulados.

Además de la importancia de la normalización, una observación aprendida de las pruebas realizadas para mejorar la eficiencia del algoritmo, fue que la búsqueda de genes alterados mejoraba cuando se hacía un filtro de diferencia de expresión en los datos de entrada; esto es, se hacía una preselección de genes con un determinado umbral de diferencia de expresión entre el estado control y el alterado. Pese a que el algoritmo puede trabajar con datos de *microarrays* sin hacer un filtro previo, para obtener unos resultados más correctos y significativos es mejor evitar que la expresión media de los genes sea igual entre las muestras control y las alteradas. Para conseguir esto, los datos de entrada se pueden filtrar previamente con un algoritmo que busque genes expresados diferencialmente en los dos estados. Uno de los algoritmos mas sólidos y útiles para esto es *SAM* (*Significant Analysis of Microarrays*) (Tusher, *et al.*, 2001) que usa permutaciones de las muestras para estimar el porcentaje de genes elegidos al azar, y calcular un *false discovery rate* (*FDR*) que corresponde al cálculo de un error de tipo I es decir, una estimación del número de falsos positivos (Benjamini y Hochberg, 1995).

Una forma más simple de quedarnos con genes que tienen una diferencia de expresión mínima entre las muestras control y las alteradas, es eliminar los genes que no tienen un cambio mínimo en la media de expresión entre los dos estados. Ésto es un filtro de medias (**ΔFM**), que está incluido al principio del algoritmo, y el umbral de diferencia de expresión media se puede fijar antes de su ejecución.

La **figura 4.3** muestra como se mejora la eficiencia del algoritmo cuando se aplican filtros de diferencia de expresión. En ella se representan los genes que cambian en el primer grupo obtenido por el algoritmo al aplicar los filtros explicados previamente. Los datos crudos se corresponden a utilizar como parámetro de entrada los valores de expresión de todos los genes del *microarrray* sin ningún filtro (22.283 *probesets*) (**figura 4.3a y d**). Los datos filtrados están calculados con el filtro de medias **ΔFM** (**figura 4.3g**) o usando *SAM* con o sin el filtro de medias **ΔFM** (**figura 4.3b, c, e, f, h e**

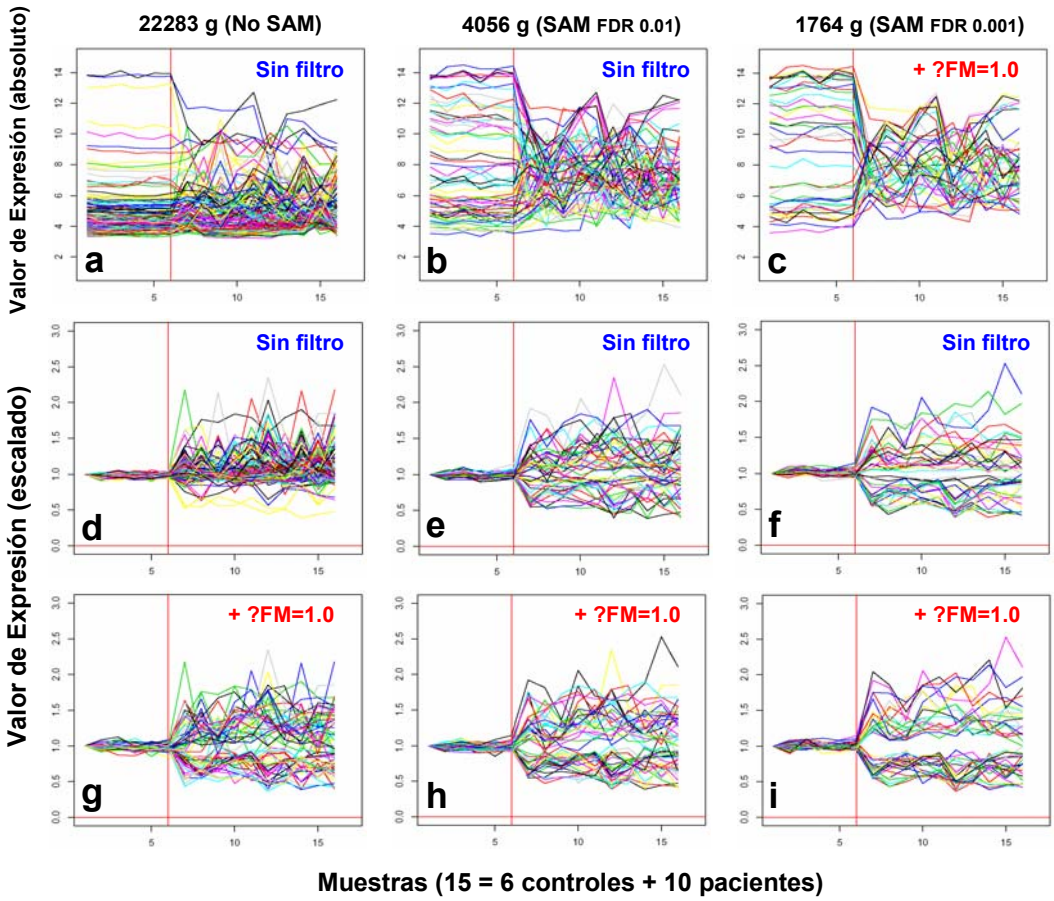


Figura 4.3: omparativa de la salida del algoritmo usando filtros:
Gráficos que muestran los valores de expresión de genes en 16 muestras diferentes: 6 controles y 10 pacientes (*APL*) separados por una línea roja. Los genes representados son los del primer grupo generado por el algoritmo pasando como entrada los datos de expresión: de todos los probesets (22283) del microarray (**a**,**d**,**g**); de 4056 probesets obtenidos después de aplicar *SAM* con un **FDR=0.01** (**b**,**e**,**h**); o de 1764 genes obtenidos aplicando *SAM* con un **FDR=0.001** (**c**,**f**,**i**). Sobre estos datos de entrada se ha aplicado también un filtro de medias de 1.0 en los diagramas que así se indica. De cada gen se representan (con líneas en color) 16 valores de expresión (uno por cada muestra) en escala $\log_2$ (**a**,**b**,**c**), estos valores se representan también en escala normalizada al valor de expresión de la primera muestra (**d**,**e**,**f**,**g**,**h**,**i**).

i). Como se ve en los gráficos escalados de la **figura 4.3**, usando un filtro de medias igual a 1.0 (que se corresponde a un 8-9% del rango de expresión) se elimina una gran cantidad de genes que no tienen cambio entre las muestras control y las alteradas (llamados genes "planos") (**figura 4.3g, h e i**). El uso del filtro de medias combinado con *SAM* da al algoritmo el mejor rendimiento en la búsqueda de genes alterados.

Ésta es una estrategia muy rigurosa que ayuda al método a encontrar genes que tienen una diferencia de expresión y una alteración clara, de forma que se evitan al máximo los falsos positivos. La ventaja de usar el filtro de medias en combinación con *SAM* se muestra en la **figura 4.3**, en la que se puede ver que usar solamente *SAM*, con un valor muy estricto de **FDR = 0.001**, no es suficiente para filtrar todos los genes "planos" (**figura 4.3e y f**). Estos genes se eliminan completamente aplicando un filtro de

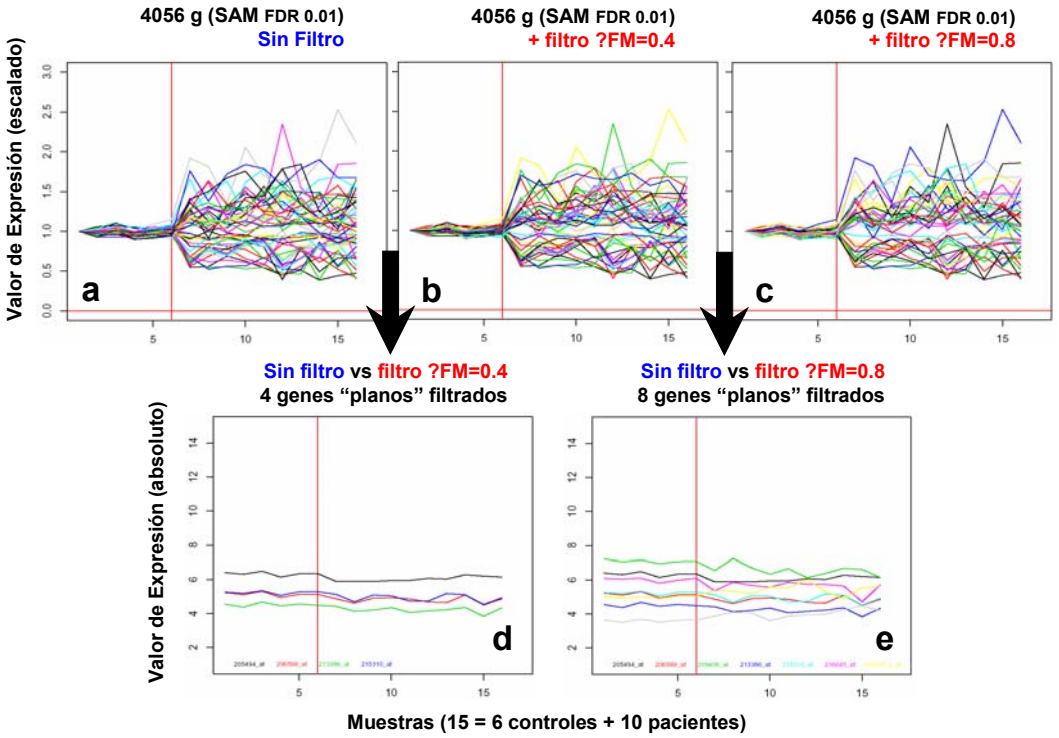


Figura 4.4: Filtrado de genes "planos" en *SAM*:
Gráficos, similares a los de la **figura 3**, que representan valores de expresión de genes. Los genes son los encontrados en el grupo 1 aplicando diferentes filtros en el conjunto inicial de datos de expresión: (**a**) 4056 genes (*SAM* a **FDR** 0.01) sin **ΔMF**; (**b**) con **ΔMF** = 0.4; (**C**) con **ΔMF** = 0.8. Los gráficos (**d**) y (**e**), muestran los genes "planos" que se han filtrado de los gráficos (**a**) y (**b**) respectivamente, cuando se aplica el filtro de medias (**ΔMF**) indicado.

medias de 1.0 (**figura 4.3h e i**). Para una mejor demostración de que el filtro de medias implementado en nuestro algoritmo, mejora el uso exclusivo de *SAM*. La **figura 4.4** muestra que aún usando *SAM* con un **FDR = 0.01** quedan genes "planos" que no cumplen el criterio deseado de tener una mínima diferencia de medias de expresión. Estos genes se quedan al aplicar *SAM* porque tienen en general una desviación estándar muy baja, y por tanto se pueden tomar como significativos en test estadísticos que usan medias y desviaciones estándar como es el caso de *SAM* (que es un *t-test* modificado). Por contra, el filtro de medias es más eficiente filtrando estos genes y permite obtener unos resultados más consistentes en la ejecución del algoritmo *AlteredExpresion*.

Como orientación para usar el filtro de medias, hemos hecho una correlación numérica entre el número relativo de genes que se filtra a un determinado valor del filtro de medias y a un determinado valor de *FDR* en *SAM*. La correlación obtenida indica que para los datos usados en el capítulo, un filtro de medias de 0.8-1.0 es equivalente a un *FDR* de 0.01 (99% de especificidad estadística).

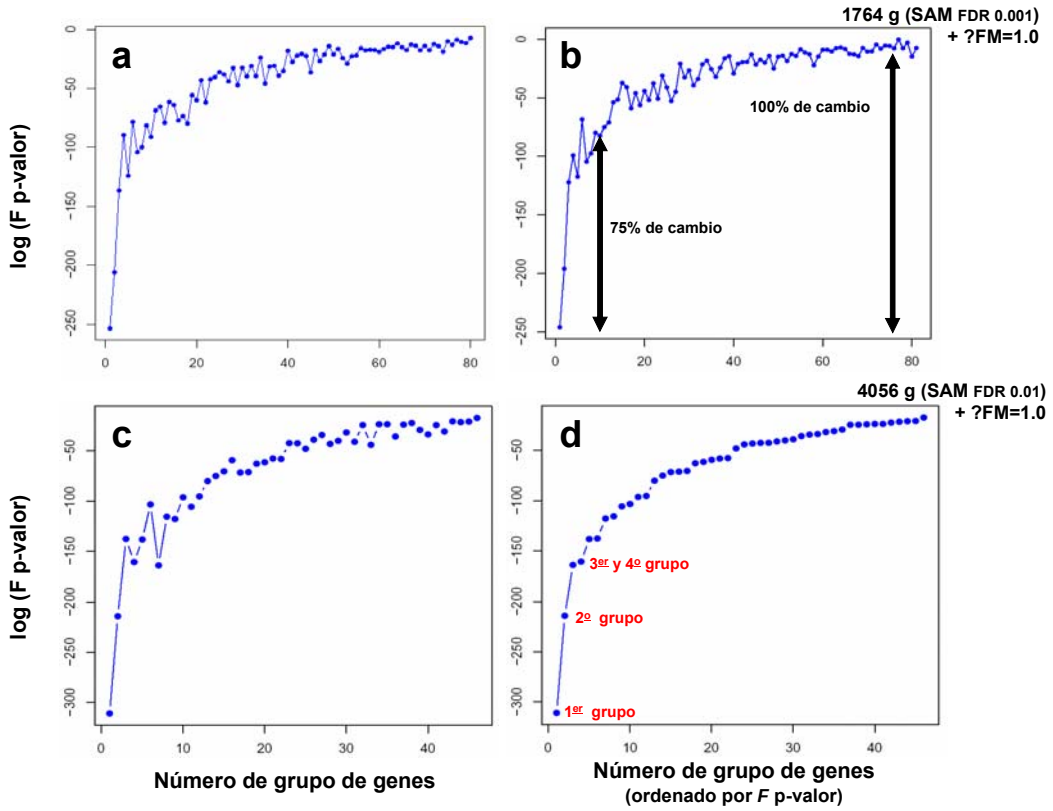


Figura 4.5: Representación de los *F-scores* generados:
Gráfico que representa los cambios de los *F-scores* (en escala $\log_{10}$) para cada grupo de genes generado. Los grupos se obtuvieron aplicando el algoritmo a un conjunto de 1764 genes (*SAM* **FDR=0.001** y **ΔMF=1.0**) para (**a**) y (**b**) y a un conjunto de 4056 genes (*SAM* FDR=0.01 y **ΔMF=1.0**) para (**c**) y (**d**). En el gráfico (**b**) se han ajustado los *F-scores* con el método de *Bonferroni*. En el gráfico (**d**) los grupos están ordenados por *F-score*.

4.2.5 Evolución del F-score en AlteredExpression

Como se describe en el algoritmo, *AlteredExpression* explora la matriz de datos de expresión en busca de grupos de genes que tienen una variación mínima en las muestras control pero que muestran una variabilidad clara en las muestras de casos patológicos. De este modo, el algoritmo produce una serie de grupos consecutivos que tienen un valor mínimo de *F-score* para el primer grupo generado y se va incrementando para los grupos sucesivos. El comportamiento de los *F-scores* de los grupos encontrados después de ejecutar el algoritmo se muestra en la **figura 4.5**, que representa los *F-scores* para cada grupo en escala $\log_{10}$ (**figura 4.5a**) y los valores de probabilidad ajustados mediante el método de *Bonferroni* (Bonferroni, 1937) para cada grupo (**figura 4.5b**). Estos gráficos indican que los *F-scores* presentan una tendencia asimptótica y que el cambio más acusado se produce en los grupos iniciales. Un 75% del cambio total del *F-score* se da entre el primer grupo y el décimo, además el cambio entre los primeros tres grupos es extraordinario. Los grupos de la **figura 4.5a y b** se obtuvieron partiendo del conjunto de 1764 genes (obtenidos con *SAM* a un

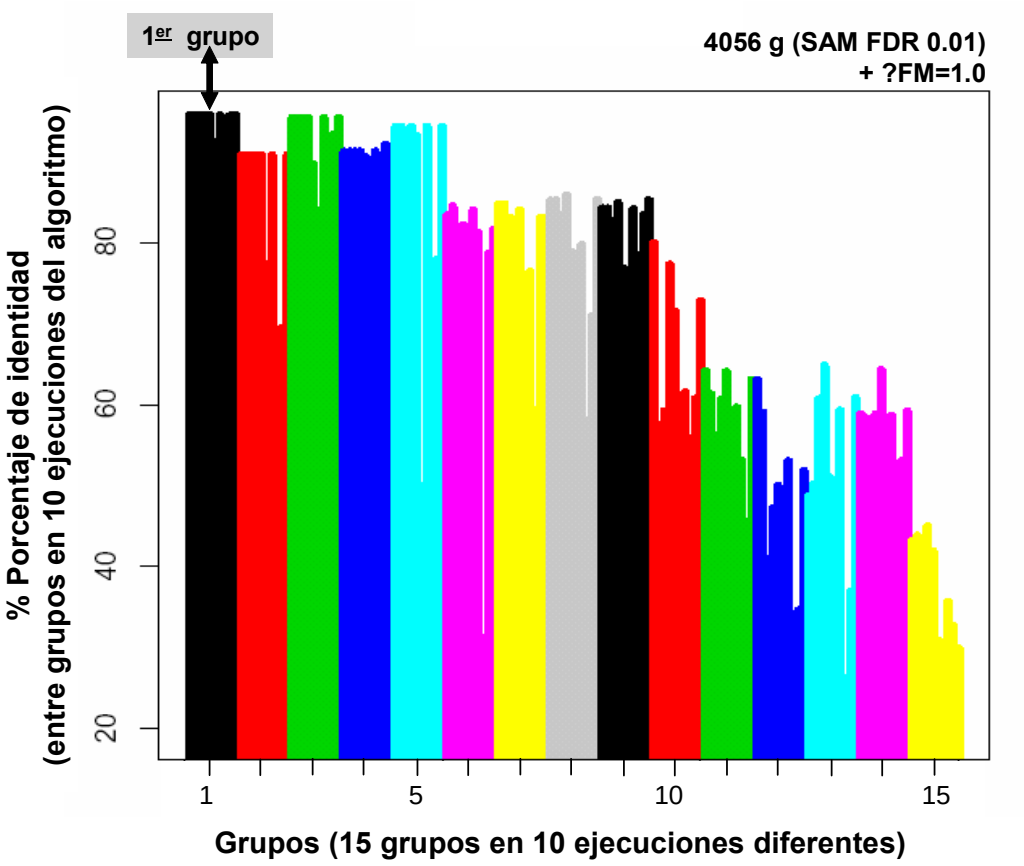


Figura 4.6: Estabilidad del algoritmo:
Comparación de los 15 grupos obtenidos en 10 ejecuciones diferentes del algoritmo. De cada grupo de genes se saca el máximo porcentaje de identidad con los grupos obtenidos en las otras 9 ejecuciones. Los 15 grupos están representados por 10 barras coloreadas que muestran los porcentajes de similitud de ese grupo en las 10 ejecuciones del algoritmo.

FDR=0.001 y **ΔFM=1.0**), se observa un resultado parecido usando el conjunto de 4056 genes (obtenidos con *SAM* **FDR=0.01** y **ΔFM=1.0**) en el que el comportamiento de los *F-scores* es similar y de nuevo los grupos iniciales tienen los cambios más significativos. La tendencia general a incrementar los *F-scores* no siempre sigue el orden en el que el algoritmo genera los grupos (**figuras 4.5a, b y c**). Para evitar estas fluctuaciones los grupos se ordenan por *F-score* en orden ascendente (**figura 4.5d**), de forma que se muestra mejor el continuo crecimiento del parámetro estadístico y permite una selección mas adecuada de los grupos de genes más significativos.

4.2.6 Estabilidad del algoritmo

Un punto crítico en los algoritmos de selección de grupos de genes es la reproducibilidad que tienen, o en otras palabras, cual es la consistencia, la coherencia y la estabilidad de los grupos más significativos seleccionados. Para comprobar esto se usó un conjunto de 4056 genes (obtenido con *SAM* **FDR=0.01** y **ΔFM=1.0**) y se ejecutó el algoritmo 10 veces, obteniendo 10 resultados diferentes de 15 grupos de

genes. Estos resultados fueron comparados entre sí para ver los genes que se habían mantenido en las distintas ejecuciones del algoritmo. Los resultados obtenidos se representan en la **figura 4.6** como el porcentaje de identidad de cada grupo con los generados en las otras 10 ejecuciones. El gráfico indica que del primer al quinto grupo, la reproducibilidad de los resultados es muy alta ya que los grupos tienen una identidad > 90%. La consistencia es alta, alrededor de un 80%, para los grupos 6-9, y empieza a ser baja después del grupo 10. Este resultado indica también que hay una correlación entre la coherencia de los primeros grupos y los *F-scores* bajos que tienen (**figura 4.5**).

La fuerte estabilidad y los bajos *F-scores* de los grupos iniciales seleccionados por el algoritmo, son dos buenos indicadores de que los genes incluidos en esos grupos mantienen algún tipo de relación biológica que hace que estén juntos en un *cluster*. Esta observación es acorde con a la hipótesis básica sobre el estado biológico alterado, que presuponía la existencia de grupos de genes con expresión alterada en muestras provenientes de células o tejidos enfermos. Como se ha demostrado, el algoritmo propuesto es capaz de encontrar estos grupos de genes con expresión alterada con respecto a unas muestras control. La siguiente cuestión a abordar es si esos grupos de genes de expresión alterada tienen genes con función biológica similar o relacionada.

4.2.7 Significación biológica de los grupos generados

A la hora de comprobar la consistencia biológica de los grupos generados se anotaron los genes a categorías funcionales definidas en *Gene Ontology* (GO) (Gene Ontology Consortium, 2005). El resultado de esta anotación en los grupos generados por el algoritmo con los datos prefiltrados por *SAM* (conjunto de 4056 genes) se muestra en la **figura 4.7**. Los datos indican que la anotación funcional de los grupos obtenidos es bastante diferente. Por ejemplo, el primer grupo incluye genes relacionados con funciones que no se encuentran en ninguno de los otros grupos (grupos 2-5) como son: desarrollo de los vasos sanguíneos (*GO:0001568*), diferenciación de células endoteliales (GO:0045446), diferenciación mioblástica (*GO:0045445*), etc. Este resultado muestra una consistencia y coherencia funcional de los grupos obtenidos. Dicha consistencia indica que el algoritmo descubre grupos estables con funciones definidas, explorando eficazmente la variabilidad observada entre el conjunto de muestras alteradas y las control.

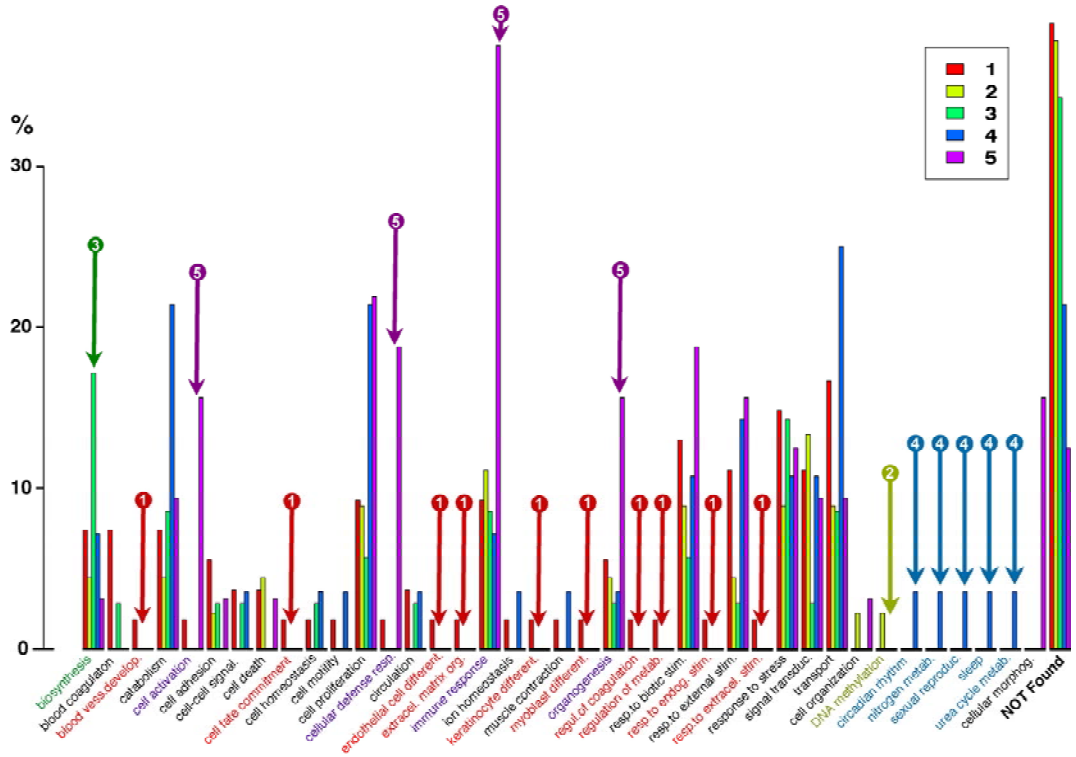


Figura 4.7: Estudio funcional de los resultados:
Asignación a términos *GO* del primer al quinto grupo de genes generados por el algoritmo. La anotación está hecha a categorías *GO* de proceso biológico a un nivel ≥ 3. La escala representa el porcentaje de genes de cada grupo que están anotados a la categoría funcional, teniendo en cuenta que un gen puede estar anotado a varias categorías.

4.3 CONCLUSIONES

En esta parte del trabajo hemos comprobado en datos transcriptómicos que es común que exista **desregulación y alteraciones en los niveles de expresión de genes**, debidas a anormalidades en los procesos de regulación de transcripción. Estas desregulaciones es esperable que se sucedan de modo drástico en estados biológicos anómalos, como los que se dan en muchas enfermedades y patologías, y por ello es interesante buscar grupos de genes que tienen una **pérdida de correlación** en sus niveles de expresión y que sufren cambios o variabilidad extrema respecto a sus controles normales.

Usando como parámetro de variabilidad la *varianza residual relativa* (**VRR**) hemos desarrollado e implementado un **nuevo algoritmo** que agrupa genes que han sufrido una fuerte desregulación o alteración en su variabilidad. También hemos comprobado que para optimizar los resultados de este algoritmo en la búsqueda de perfiles alterados a partir de datos de expresión genómica, es conveniente aplicar previamente filtros (como un filtro de medias o un filtro de expresión diferencial) que evitan la inclusión de genes con perfiles de expresión "planos" en los resultados. Este tipo de genes son frecuentes entre la gran proporción de genes que se expresan poco o que casi no cambian dentro de los datos de *microarrays*.

El nuevo algoritmo denominado *AlteredExpression* consigue generar **grupos bien definidos de genes alterados** que han sufrido una desregulación en su expresión. La significación o categorización de los grupos obtenidos se ha logrado desarrollando y aplicando el estadístico *F-score* que calcula los grupos más estables. Finalmente, se ha comprobado que los grupos generados tienen **significación biológica** ya que los genes incluidos en cada grupo tienen funciones relacionadas, mostrando que cuando se produce una alteración ésta afecta a un conjunto de genes implicados en una alguna función o proceso biológico común.

CONCLUSIONES GENERALES

Respecto a **redes de interacción de proteínas**:

- La heterogeneidad existente entre las bases de datos y repositorios de **interacción de proteínas**, hace necesaria la organización y la recopilación de esta información en una **plataforma unificada ágil** de acceso *web* por internet. Esto se ha logrado en la plataforma bioinformática construida llamada *APID*.
- Los datos de interacción de proteínas tienen un elevado número de falsos positivos, que es importante filtrar mediante el uso de **estratégias de validación**. En *APID* se han implementado estratégias basadas en el número de experimentos que validan la interacción y en la existencia de datos estructurales dominio-dominio entre las proteínas interactuantes.
- La comparación y combinación de varios estudios que analizan **interacciones estructurales dominio-dominio** permite generar conjuntos de datos fiables, que integrados ayudan a conseguir mayor cobertura sobre los interactomas.
- Las aplicaciones de **visualización de datos de interacción** entre proteínas que hemos desarrollado, *APIN* y *APID2NET*, facilitan el análisis y la exploración dinámica de las redes biomoleculares complejas.

Respecto a **redes de coexpresión de genes**:

- La búsqueda de **genes coexpresados** es una cuestión comunmente planteada que necesita de técnicas y métodos robustos que den buena fiabilidad.
- La aplicación de varios **algoritmos combinados** de cálculo de señal y de correlación para buscar perfiles de expresión correlacionados, consigue obtener mejores resultados que aproximaciones simples. Además, el empleo de técnicas de validación cruzada en el cálculo de coexpresión, aporta un claro incremento en la robustez y fiabilidad de los resultados.
- La obtención de datos de coexpresión génica fiables permite un mejor descubrimiento de la **información biológica** funcional y transcripcional subyacente.
- Los resultados del desarrollo y aplicación del nuevo algoritmo *AlteredExpresión* demuestran que se pueden encontrar grupos de genes con expresión altamente variable presentes en estados alterados respecto a estados controles normales. Estos **genes alterados** pueden ser marcadores de estados biológicos patológicos o enfermos.
- Se ha demostrado que los grupos de genes que se encuentran con grados de alteración semejante elevada, están involucrados en **funciones o procesos biológicos comunes**, y de esta forma pueden marcar los estados patológicos alterados..

APÉNDICE 1: PUBLICACIONES

- **Prieto C**, De las Rivas J. (2009). Comparison and assessment of structural domain-domain interactions to improve protein-protein interaction data. *Proteomics*. En preparación.

- **Prieto C**, Risueño A, Fontanillo C, De las Rivas J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *Plos One 3(12)*: e3911.

- Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, **Prieto C**, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology 25(8)*: 894-8.

- Hernández-Toro J, **Prieto C**, De Las Rivas J (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics 23(18)*: 2495-7.

- **Prieto C** and De Las Rivas J (2006). APID: Agile Protein Interaction DataAnzalizer. *Nucleic Acids Res. 34 (Web Server issue)*: W298-302.

- **Prieto C**., Rivas-Lopez M. J., Sánchez-Santos J. M., Lopez-Fidalgo J. and De Las Rivas J. (2006). Algorithm to find gene expression profiles of de-regulation and identify families of disease-altered genes. *Bioinformatics 22(9)*: 1103-10.

- **Prieto C**, Risueño A, Fontanillo C, De las Rivas J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *Plos One 3(12)*: e3911.

OPEN ACCESS Freely available online

PLoS one

Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles

Carlos Prieto, Alberto Risueño, Celia Fontanillo, Javier De Las Rivas*

Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CIC-IBMCC, CSIC/USAL), Salamanca, Spain

Abstract

Background: Analysis of gene expression data using genome-wide microarrays is a technique often used in genomic studies to find coexpression patterns and locate groups of co-transcribed genes. However, most studies done at global "omic" scale are not focused on human samples and when they correspond to human very often include heterogeneous datasets, mixing normal with disease-altered samples. Moreover, the technical noise present in genome-wide expression microarrays is another well reported problem that many times is not addressed with robust statistical methods, and the estimation of errors in the data is not provided.

Methodology/Principal Findings: Human genome-wide expression data from a controlled set of normal-healthy tissues is used to build a confident human gene coexpression network avoiding both pathological and technical noise. To achieve this we describe a new method that combines several statistical and computational strategies: robust normalization and expression signal calculation; correlation coefficients obtained by parametric and non-parametric methods; random cross-validations; and estimation of the statistical accuracy and coverage of the data. All these methods provide a series of coexpression datasets where the level of error is measured and can be tuned. To define the errors, the rates of true positives are calculated by assignment to biological pathways. The results provide a confident human gene coexpression network that includes 3327 gene-nodes and 15841 coexpression-links and a comparative analysis shows good improvement over previously published datasets. Further functional analysis of a subset core network, validated by two independent methods, shows coherent biological modules that share common transcription factors. The network reveals a map of coexpression clusters organized in well defined functional constellations. Two major regions in this network correspond to genes involved in nuclear and mitochondrial metabolism and investigations on their functional assignment indicate that more than 60% are house-keeping and essential genes. The network displays new non-described gene associations and it allows the placement in a functional context of some unknown non-assigned genes based on their interactions with known gene families.

Conclusions/Significance: The identification of stable and reliable human gene to gene coexpression networks is essential to unravel the interactions and functional correlations between human genes at an omic scale. This work contributes to this aim, and we are making available for the scientific community the validated human gene coexpression networks obtained, to allow further analyses on the network or on some specific gene associations. The data are available free online at http://bioinfow.dep.usal.es/coexpression/.

Citation: Prieto C, Risueño A, Fontanillo C, De Las Rivas J (2008) Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles. PLoS ONE 3(12): e3911. doi:10.1371/journal.pone.0003911

Editor: Nicholas James Provart, University of Toronto, Canada

Received July 3, 2008; **Accepted** November 5, 2008; **Published** December 15, 2008

Funding: Funding and grant support was provided by the Ministery of Health, Spanish Government (ISCiii-FIS, MSyC; Project reference PI061153) and by the Ministery of Education, Castilla-Leon Local Government (JCyL; Project reference CSI03A06). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jrivas@usal.es

Introduction

Exploration and analysis of gene expression data using genome-wide microarrays is a technique often used in genomic studies to find coexpression patterns and locate groups of co-transcribed genes. This kind of studies has been used in model organisms, like yeast [1], to discover gene functions, to define biological processes and to find related transcription factors and their products. The main features of expression patterns that give a wide utility in bioinformatic studies are: the functional information associated [2], the high conservation of gene coexpression groups along evolution [3] and the high correlation of these groups with biomolecular pathways or reactions [4]. All these features leverage genome-wide expression profiling, and convert this topic in a hot research area.

Despite the described interest, coexpression studies done at global "omic" scale are not focused in many cases on human samples [5], and, when they correspond to human, very often they include heterogeneous datasets, mixing "normal" samples with "disease altered" samples from patients suffering from some kind of pathological state. This is the case, for example, in several human gene expression large studies [2,6]. The inclusion of many disease datasets (mainly from cancer) in such meta-analyses may introduce strong bias and produce a lot of biological noise in the results. In fact, it is well known that cancer cells have altered genomes. Therefore, these kind of studies cannot be used to clarify

how a normal-healthy human cellular system works, and they cannot be used to draw a reliable map of the human gene coexpression landscape.

The technical noise in the genome-wide expression microarray studies is another well reported problem that can not be ignored when gene coexpression studies at "omic" scale are undertaken. Considering all these problems and knowing the interest of having a reliable normal human gene coexpression network, we have undertaken this task selecting human genome-wide expression microarrays from a controlled set of different normal tissues to build a confident human transcriptomic network using several statistical and computational methods. These methods (which include robust data normalization and signal calculation, combined parametric and non-parametric correlation and random cross-validation) help to avoid both biological and technical noise and provide a human gene coexpression network that shows good accuracy and coverage. Moreover, the network reveals well defined biological functions and pathways that map to specific coexpression clusters.

Results and Discussion

Genome-wide expression profiles from a broad set of human samples

An expression matrix was calculated for a dataset of human genome-wide microarrays hybridized with mRNA samples coming from different human tissues, glands and organs from healthy normal individuals. As indicated in **Materials and Methods** the dataset included two biological replicates of samples from 24 parts of the body: *adrenal gland, appendix, blood, bone marrow, brain, kidney, liver, lung, lymph node, muscle heart, ovary, pancreas, pituitary gland, prostate gland, salivary gland, skin, spinal cord, testis, thymus gland, thyroid gland, tongue, tonsil gland, trachea* and *uterus*. **Figure 1** presents the heatmaps and clustering of the 48 samples analyzed by two different methods following the strategy and steps described in **Methods**: (**1st**) "MAS5-Spearman" method, that applies MAS5 algorithm for signal calculation and Spearman correlation coefficient (**r**) for distance calculation (based on the sample expression profiles and displayed in the heatmap as **1 r**); (**2nd**) "RMA-Pearson" method, that applies RMA algorithm for signal calculation and Pearson correlation coefficient (**r**) for distance calculation (also based on the sample expression profiles and displayed as **1 r**). We use "Spearman with MAS5" and "Pearson with RMA" because it has been shown that the inclusion of at least one non-parametric step based on ranks in the analyses of microarray data offers statistically more robust and more accurate estimation of expression values [7] and expression correlations [8]. The two methods proposed provide such non-parametric transformation (i.e. change to ranks), because Spearman is a rank correlation coefficient and RMA includes a quantile normalization.

The heatmaps (**Figures 1A and 1B**) show a clear and coherent clustering of each pair of biological replicates. A color bar with scales for each heatmap is included in the figure, indicating that **dark-red** corresponds to minimum distance (i.e. maximum correlation) and **dark-blue** to maximum distance (i.e. minimum correlation). White color corresponds to medium values and the distributions inside the color bars show that the two methods are similar but not identical: MAS5-Spearman provides larger distances between samples (more blue values in the heatmap) than RMA-Pearson (more red values in the heatmap).

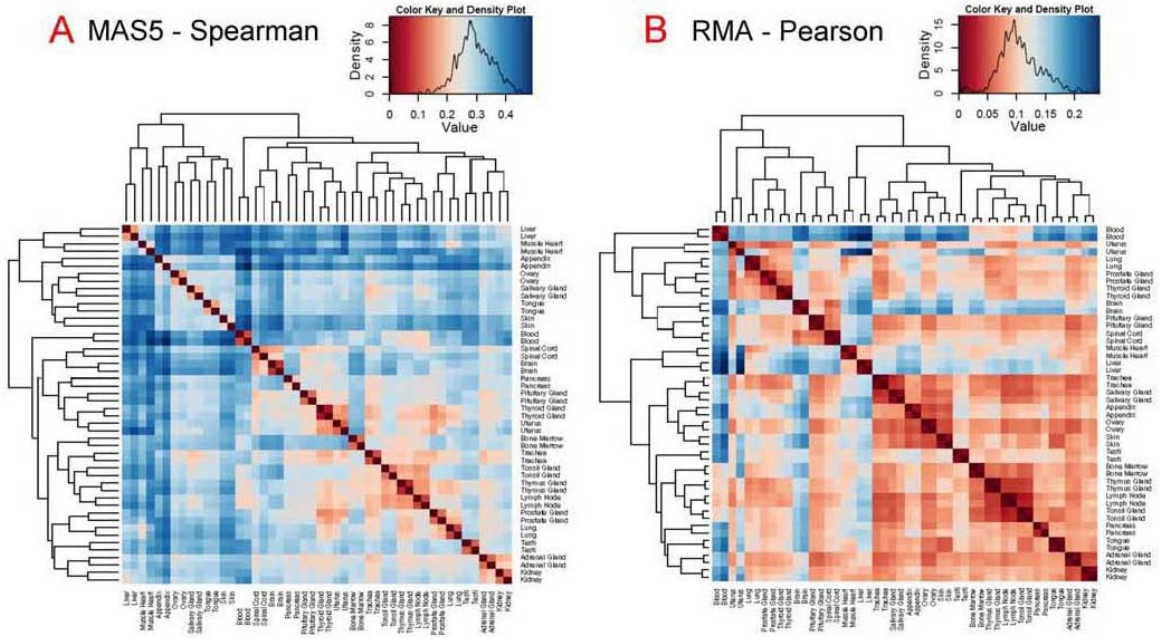


Figure 1. Clustering of human tissue expression profiles. Heatmaps and clustering of the 48 human genome-wide expression microarray samples from 24 different tissues and organs analyzed by two different methods: **(A)** MAS5-Spearman: MAS5 for signal calculation and Spearman for distance calculation based on the sample expression profiles; and **(B)** RMA-Pearson: RMA for signal calculation and Pearson for distance calculation based on the sample expression profiles. A color bar with scales for each heatmap is included, indicating that **dark-red** corresponds to minimum distance and **dark-blue** to maximum distance. The color distributions observed in the heatmaps are also included inside the bars.
doi:10.1371/journal.pone.0003911.g001

The similarity and proximity of the replicates is closer in the case of the second method, but in both cases there is not confusion or separation of any pair of replicates. By contrast to this clear clustering, the ordering and clustering of the different tissues, glands and organs is not fixed in the heatmaps, changing quite a lot from **1A** to **1B**. This observation was confirmed by bootstrap analysis done with *pvclust* [9] which allows the assessment of the uncertainty in hierarchical clusters (see **Methods**). The results of *pvclust* showed that the biological "replicate pairs" gave in all cases stable groups with optimum probability values (AU and BP = 100%). However, within the tissues and organs only two stable groups were found with both methods: the group that includes *lymph node*, *thymus gland* and *tonsil gland* (that gave a AU value of 0.98); and the group that includes *kidney* and *adrenal gland* (with AU value 0.97). These groups have clear biological meaning since they correspond to physiologically and functionally related organs (i.e. *lymph node*, *thymus* and *tonsil* are related to the lymphatic and immune systems). Thus the functional relationship between samples is captured by the gene expression profiles. However, all the other tree branches produced low AU values, therefore the overall sample clustering observed in the heatmaps indicates a lack of well defined and stable groups. In conclusion, these results show neat separation of most of the sample expression profiles, which is an adequate condition for the exploration of a global broad human gene expression landscape.

In order to consider if these observations are reliable enough, we explored the data changing some conditions following another two different strategies (data not shown). **First** strategy, the same analyses with 48 microarrays were done again twice: one not using the total number of genes (i.e. 22 283 gene probesets) but only the 25% of the genes that showed the largest variance; and another using only the 25% of the genes that showed the highest signal. In both cases, the heatmap and trees obtained were very similar to the ones presented in **Figure 1**, and the bootstrap gave similar results. **Second** strategy, we included in the data set two new groups of microarrays corresponding to samples from specific organs: 16 microarrays from different parts of the brain and 10 microarrays from different hematologic cell types. In this case (data not shown) the analyses provided larger trees, where two main clusters were segregated from other branches: one corresponding to brain related samples (i.e. nervous system) including the two whole-brain samples; and another cluster corresponding to the hematologic related samples including the two whole-blood samples. These results indicate again that any functional relation between samples is well captured by the global gene expression profiles, and provide validity to the genome-wide expression profiles of human normal tissues obtained, allowing us to proceed to the next step of the study.

From sample expression profiles to gene expression signatures

The main data presented so far correspond to the analysis of the genome-wide expression profiles of samples from different human normal tissues, organs or glands. These genome-wide "sample profiles" are numerical vectors including the expression values of each one of the gene probesets present in the microarray (i.e. each one of the detectable genes of the human genome). As shown above, the "sample profiles" can resemble the physiological relationships expected between the samples (tissues, glands and organs). However, in order to achieve a mapping of the human gene coexpression landscape, we needed to move from the analysis of the "sample expression profiles" based on the genes, to the analysis of each "gene expression signature" based on the sample set.

It is difficult to achieve a proper gene coexpression study due to several obstacles that have to be taken in consideration: **(i)** the technical noise present in the microarrays at genomic scale [10], despite the fact that the *Affymetrix* high density oligonucleotide genechips have been reported quite reliable and reproducible [11,12]; **(ii)** the small number of samples used to define each gene expression signature (specially in comparison to the large number of genes); **(iii)** the strong heterogeneity of the data sets frequently studied, that include in many cases samples from pathological or altered states [2,13] which are not adequate samples to find "normal" gene expression behavior.

The approach and strategies taken in this study to solve or minimize these problems were the following: **(a)** careful selection of expression samples from different parts of the human body (tissues, whole glands and whole organs) from normal healthy individuals; **(b)** calculation of expression signals and correlations using two different independent methods: MAS5-Spearman, RMA-Pearson; **(c)** use of a robust random cross-validation strategy to find the most stable correlation pairs and distinguish the consistent biological-signal from the noise-signal; **(d)** statistical estimation of the accuracy and the coverage for each coexpression dataset obtained. All the details and description of these strategies are presented in **Materials and Methods**. The results associated with them have been partially described above and are explained in the following paragraphs.

Gene pairs coexpression analyzed with cross-validated correlations

The complete expression data matrix analyzed had, as indicated, 48 samples (24 duplicates) and 22,283 gene probesets (which correspond to 13,068 distinct known human genes according to *Affymetrix* annotation). Therefore the global pair-wise gene coexpression matrix including all possible pairs had 248,254,903 data points and was calculated twice, once for each independent method used (MAS5-Spearman and RMA-Pearson). These huge data matrices have many pairs that are false coexpression pairs and to detect those positive gene pairs that had stable and significant correlation we use cross-validation. The results corresponding to the gene pairs correlation obtained with the cross-validation method (described in **Methods**) are presented in **Figure 2**, that shows what we called "**rN-plots**". The rN-plots are graphics representing: **r** at *y* axis, that is, for each gene probeset pair, the "correlation coefficient" of their expression signatures along the complete dataset of 48 samples, calculated as Spearman or Pearson distance (for MAS5 or RMA data, respectively) (with values from 0 to 1 for positive correlations and from 0 to −1 for negative correlations); **N** at *x* axis, that is the "cross-validation coefficient" defined as the number of times that a given gene pair has a significant correlation (i.e. $r \geq |0.70|$) out of the 1000 times random selection (as explained in **Methods**). This graphical analysis presents the positive and negative correlations well segregated and it allows to identify those gene pairs that have a significant "cross-validated correlation", discriminated from those false gene-pairs that have low **r** or low **N** values. Such false gene-pairs do not correlate in a stable and consistent way, being undistinguishable from noise.

To demonstrate how the **rN-plots** represent stable and consistent correlations, we selected in the case of the **red** circles or dots only the gene probeset pairs that correspond to probesets assigned to "the same gene". For example, pairs between the 4 probesets that correspond to gene ALDOB, *fructose bisphosphate aldolase B* (204704_s_at, 204705_x_at, 211357_s_at, and 217238_s_at in microarray HGU133A); or pairs between the 3 probesets that correspond to gene CDK10, *cell division protein kinase 10* (203468_at, 203469_s_at and 210622_x_at in HGU133A). When correlation is found between these kind of "common gene probesets" they are drawn as **red** circles in **Figure 2**. The analysis indicates that the red circles accumulate at high **r** correlations and

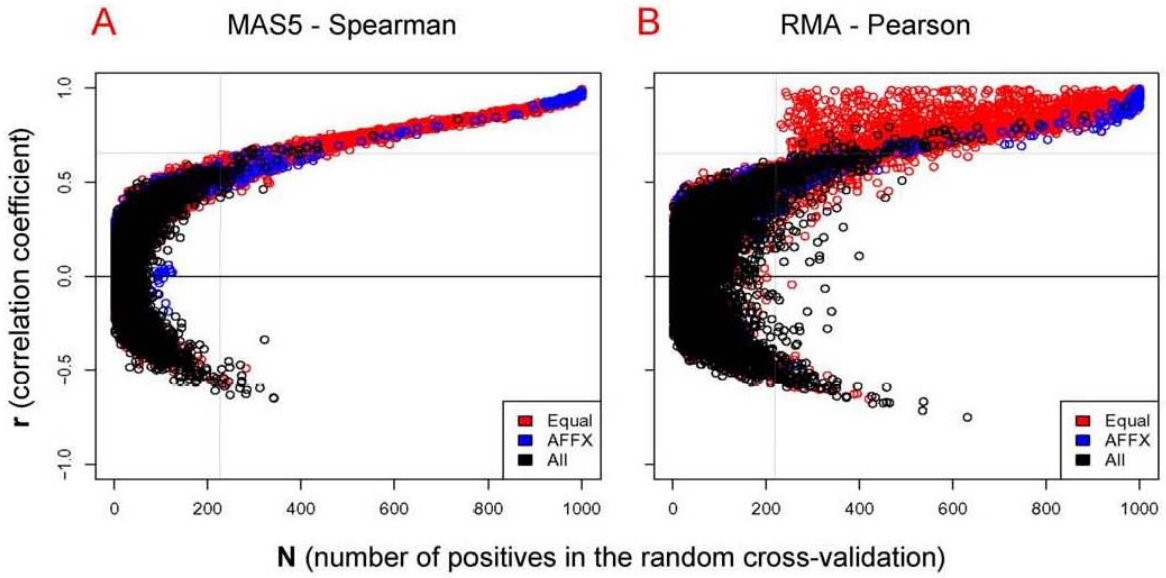


Figure 2. Plot of r and N coefficients calculated for each gene coexpression pair. rN-plots that represent the correlation coefficient (from 0 to 1) versus the cross-validation coefficient (from 0 to 1000) of each gene pair by two different methods: **(A)** MAS5-Spearman and **(B)** RMA-Pearson. The cross-validation is considered positive for a given gene pair when it gives **r**>|0.7| in each sampling. As indicated in **Methods** 1000 samplings are run for each gene-probeset pair. The gene probeset pairs that correspond to the same gene are drawn as **red** circles. The probeset pairs of *Affymetrix* controls are drawn as **blue** circles. A random selection of 10,000 coexpressed gene probeset pairs are drawn as **black** circles. Two dotted lines are drawn to indicate an approximate threshold that can be considered the border of noisy data. These lines are drawn just to show the minimal **r** and **N** values bellow which the coexpressed gene pairs are mainly noise; therefore the coexpression signal appears mostly at **r**>0.65 and **N**>220.
doi:10.1371/journal.pone.0003911.g002

high **N** values. This is the result that should be expected considering that these groups of probesets are measuring the same gene; and, despite the fact that this is not always true, it is a good way to evaluate the meaning of the **rN-plot**. A more stringent evaluation was to find out the correlation between probesets that correspond to "control RNAs" that are added in each microarray assay in the hybridization process. Such controls, named with prefix AFFX in the chip, are spike controls (i.e. series of mRNAs added during hybridization protocol that correspond to different concentrations of non-human genes like AFFX-BIO) and human house-keeping controls (like AFFX-HUMGAPDH). These controls should have strong correlation since they have been added to the microarrays in known concentrations. We draw such correlations in the **rN-plots** as **blue** circles (**Fig. 2**); and it could be seen that the distribution of these true positive gene correlated pairs was very much accumulated at high **N** values and high **r** correlations. This observation again shows that the **rN-plots** are very useful and valuable to separate noisy false correlations from stable true correlations.

The differences observed between **Fig. 2A and 2B** are due to the differences in the methods and to the characteristics of the cross-validation (described in **Methods**). Some **red** circles with high-**r** and low-**N** appear only in the RMA-Pearson method because the correlations derived from this method give in some instances high correlation values to gene pairs that are correlated just in only one tissue (shown in **Fig. 2B**). The cross-validation values of these gene pairs are low because they only appear when such one tissue samples are selected. The probability to select one sample pair out of 24 is: 1 $(23/24)^6 = 0.225$; and this is why the red circles with high-**r** and low-**N** only appear for values **N**>225. By contrast, the MAS5-Spearman method does not find any **red** circle in the high-**r** and low-**N** region, because Spearman is a "rank correlation coefficient" which does not produce high correlation values for gene pairs that correlated in only one tissue (just once out of 6). The **r** value obtained with the Spearman method is proportional to the number of tissues or samples that co-express and so it is quite proportional to **N**.

Data filtering to clear genes with low information content

The calculations and analysis presented in **Figure 2**, were done without using any previous filter of gene probesets. No filtering means using the complete gene expression matrices with all the human gene probesets present in the microarrays. It is known that in most samples and conditions genome-wide microarrays include a large proportion of the genes that are not expressed and therefore they give signal close to the background or noise. This situation is not very likely to occur all along the complete sample set of 24 different tissues and organs studied here. However, out of the 22,283 gene probesets some may have no significant change, and therefore, it is important to find out the possible presence and effect of these "non-changing genes" (that we also called "flat-genes") [14]. The most adequate filter to be used in most of the expression analyses is a variance-filtering between samples (i. e. between-array variability), because this approach filters out elements of low information content within the sample set and covers the complete signal range (from low to high expression), therefore, it does not bias the data by signal intensity or signal ratios [14,15]. However some genes with high signal may be significant despite showing relative low variance, and for these reasons it is better to apply combined filters that explore the variance, but also consider the intensity of the probes [15].

As described in **Methods** we use a combined filter based on between-sample variability and gene minimal signal, that is designed to get rid of genes with low information content. The use of this filter with the 48 microarrays sample set gave different results for the data expression matrix obtained with RMA method and the expression matrix obtained with MAS5 method. In the first case the filter leaves out 6,893 gene probesets (leaving 69.06%) and in the second 3,682 (leaving 83.48%) from 22,283 total gene probesets. The difference in these numbers shows that these two methods do not provide an equal calculation of expression signal and variance and therefore, as explained bellow, both methods can be considered complementary.

Analysis of accuracy and coverage along gene coexpression data

Using the filtered data sets we follow a more thorough analysis of the coexpression distributions with respect to the parameters **r** and **N**. In the rN-plots (**Fig. 2**) two dotted lines were drawn to indicate an approximate threshold for coexpressed gene pairs that could be considered the border of noisy data. These lines are tentatively drawn just to show the minimal **r** and **N** values bellow which the coexpression pairs are mainly noise; therefore, the coexpression signal appears mostly at **r**>0.65 and **N**>220. However, this estimation is not robust enough and a proper calculation of the statistical "accuracy" and "coverage" along all the gene coexpression data matrices was done. The details about the calculation of these parameters are described in **Materials and Methods**. KEGG pathway database was used to estimate the true positives. After these calculations, for all data presented (i.e. all next **Figures**) the nodes correspond to genes and not any more to "gene probesets" from the microarrays. This change was done taking the correspondence of the probesets to the specific genes according to the *Affymetrix* annotation files for HG-U133A from 31.May.2007 (that can be found in URL: http://www.affymetrix.com/support/technical/byproduct.affx?product=hgu133). In this conversion all probesets of the microarray were used. Previously, we calculated the coexpression values for each gene pair considering each probeset independently. When multiple probesets map to one gene, we merged the multiple probesets to the corresponding gene and we only take the gene coexpression pairs with maximum values of correlation (**r**) and cross-validation (**N**) in which its probesets participate.

In **Figure 3** the positive predictive value (PPV) was computed for each coexpression data set obtained at a given correlation factor **r** (**Fig. 3** top graphs) or at a given the number of cross-validations **N** (**Fig. 3** bottom graphs). The change or evolution of the accumulated PPV is drawn as a curve (solid **red** and **blue** circles) for both methods (**Fig. 3A**: MAS5-Spearman; **B**: RMA-Pearson). The graphs show that the rate of true positives increases with higher expression correlation and with higher number of cross-validation. The increase is more significant for the MAS5-Spearman method that achieves PPV about 80% for **r**≥0.8 and

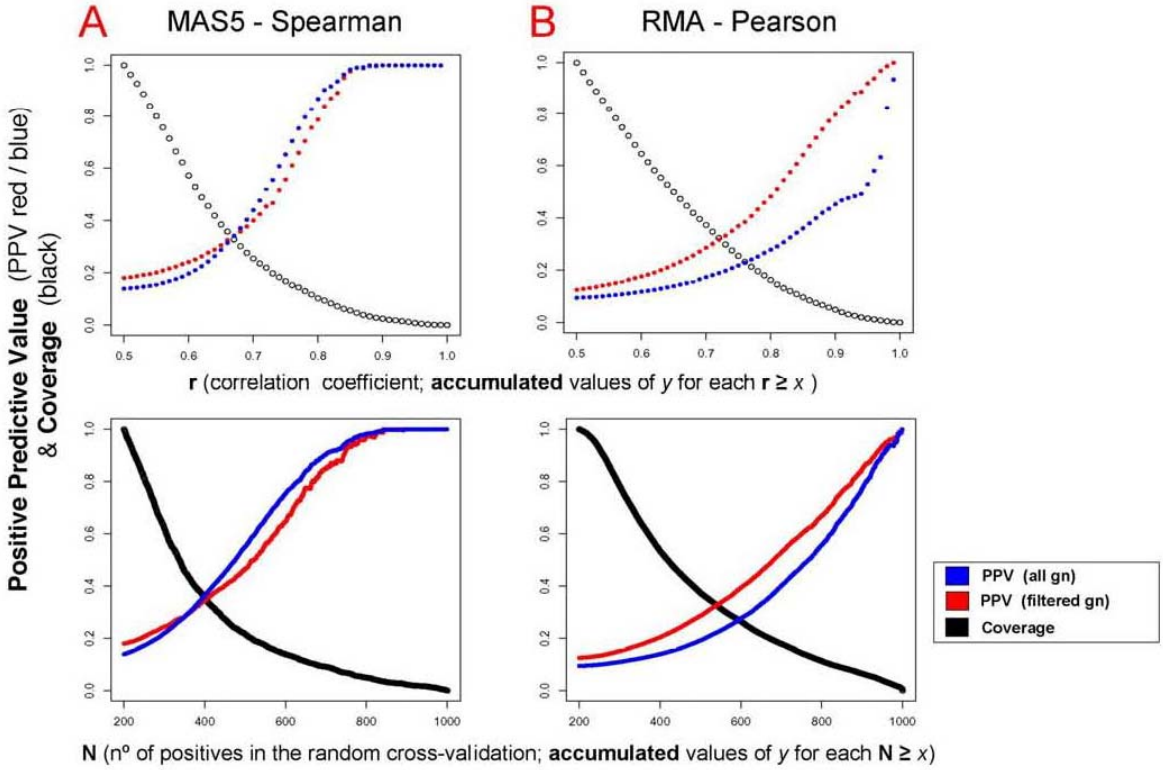


Figure 3. Accuracy and coverage of the coexpression data. Accuracy measured as Positive Predictive Value PPV (for all genes in **blue** and filtered genes in **red**) and coverage as True Positive Rate TPR (in **black**) computed for each coexpression dataset obtained at a given correlation coefficient **r** (top figures) or at a given number of cross-validations **N** (bottom figures) for both methods: **(A)** MAS5-Spearman and **(B)** RMA-Pearson. The accuracy and coverage (in *y* axis) correspond to accumulated values for each **r**≥*x* or for each **N**≥*x*.
doi:10.1371/journal.pone.0003911.g003

for **N**≥700. However, RMA-Pearson provides higher coverage since the amount of positive gene coexpression pairs annotated to common KEGGs for **r** and **N** values is quite different in both methods (larger for RMA-Pearson). The results for the coverage calculated for each method are shown by the curves in black in **Figure 3** (**black** circles), presenting the amount of gene pairs annotated to common KEGGs that remain at each **r**≥*x* or **N**≥*x*. This is calculated considering as "total amount of positive pairs" (value 1.0 at the beginning of the curve, 100%): the number of gene coexpressing pairs annotated to common KEGGs at **r**≥0.5 and **N**≥200. This coverage parameter indicates, as it should be expected, that the number of gene coexpressing pairs decreases when the conditions (**r** and **N**) are more stringent. The decrease is steeper for the MAS5-Spearman method since for **r**≥0.75 it retains about 16.7% of the positive data points, but RMA-Pearson retains 25.4%. Equally for **N**≥600 the MAS5-Spearman method retains 13.9% of the positive data points and RMA-Pearson retains 26.4%. The total amount of positive pairs, which corresponds to value 100% at the beginning of the curve, was: 15,657 for RMA-Pearson and only 2,198 for MAS5-Spearman. These numbers seem small but they only correspond to the "positive pairs", and so, if we take the total number of gene probeset coexpression pairs of the study (i.e. not including only the genes annotated to KEGGs but the complete coexpression data sets) the figures are much larger: 1,340,472 for RMA-Pearson and 180,305 for MAS5-Spearman. These results also indicate that the coverage is larger with the RMA-Pearson method.

In conclusion, the study shows that the RMA-Pearson method has better coverage of the coexpression landscape and the MAS5-Spearman is more accurate to find coexpression pairs. These results support the use of both methods in order to find a confident human coexpression network, since they do not find exactly the same expression signal and both provide important and complementary data allowing a progressive improvement of the significance and confidence of the coexpression set. Moreover, a better knowledge of the strength of each method is a discovery that complements previous comparative studies about RMA [7] and MAS5 [8].

Effects of gene filtering

The original coexpression data used in **Figure 2** are obtained without any gene filtering, however for the analyses in **Figure 3** it was convenient to study the effect of gene filtering upon the accuracy and coverage of the methods. The evolution of the coverage did not show any significant change (data not shown). The evolution of the accuracy was studied by plotting the relative changes of the positive predictive values (PPV) of the coexpressing data with **r** (**Fig. 3** top graphs) and **N** (**Fig. 3** bottom graphs) for each method. In these graphs the **blue** circles correspond to non-filtered data and **red** circles to filtered data. This analysis indicates that for the case of RMA-Pearson method (**Fig. 3B**) a significant improvement was obtained with the gene filtered versus non-filtered. However, in the case of MAS5-Spearman there was not any relative improvement, as it can be seem in **Fig. 3A** both for **r** and **N**. This means that **r** and **N** are already very stringent in MAS5-Spearman dataset and the filter takes out approximately the same amount of estimated true positives and false positives within the data, and so it does not improve the coexpression accuracy (i.e. PPV). This observation, together with the fact that filtered data with the MAS5-Spearman method gives low coverage (as indicated above the total amount of positive pairs was only 2,198), brings us to the resolution of not using the filter for MAS5 dataset. By doing this, the MAS5-Spearman non-filtered dataset at **r** = 0.5 and **N** = 200 included 15,623 positive coexpression pairs; and this number was very similar to the 15,657 pairs found for RMA-Pearson filtered.

Integration of correlation, cross-validation and PPV for datasets obtained with two balanced methods

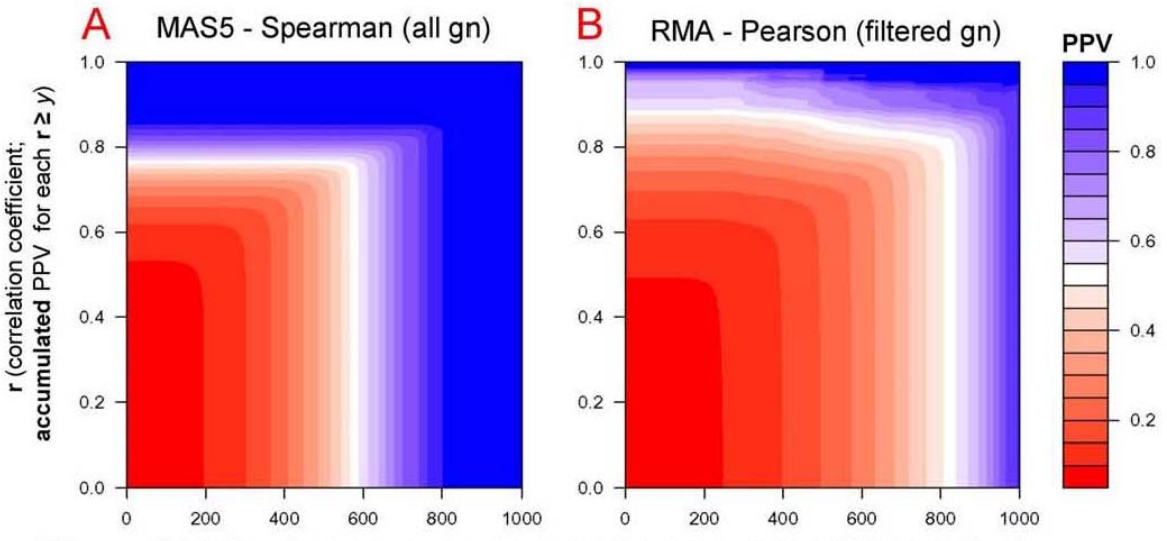
Following the observations and arguments described above we proceed to integrate in "three-dimensions color plots" the data corresponding to the values of correlation (**r**), cross-validation (**N**) and PPV obtained with each method. The results are shown in **Figure 4**. The graphic considers all the calculated subsets of coexpression gene pairs and represents, for each one, the numerical relationship between the accumulated values of the estimated accuracy (**PPV**) corresponding to the correlation coefficients (**r** in *y* axis) and to the cross-validation coefficients (**N** in *x* axis). PPV ranges from 0.05 to 1.0 as indicated in the color scale of **Fig. 4**: **red** low and **blue** high. The graph are calculated for the data corresponding to two methods: MAS5-Spearman without gene filtering (all gn) (**Fig. 4A**) and RMA-Pearson with gene filtering (filtered gn) (**Fig. 4B**). As indicated above, in these conditions both methods include a similar number of coexpression pairs and so they are "balanced" with respect to the coverage.

The three-dimensions color plots allow to assess in a graphic way the level of confidence for a given coexpression data subset. We use them to select three data subsets derived from each method at three specific PPV values: ≥0.60, ≥0.70 and ≥0.80. The values of the correlation and cross-validation coefficients that correspond to these data subsets are indicated in the table enclosed as **Fig. 4C**. The figures show that the second method (RMA-Pearson) is more stringent, since the same given PPVs correspond to higher values of **N** and **r**. The size of the gene coexpression networks that correspond to the three selected accuracy values are also presented in **Fig. 4C**, including for each network the number of nodes (i.e. number of genes) and the number of links (i.e. number of coexpression pairwise relations). The selection and combination of these subsets at well defined and precise accuracy allows the identification of stable and confident human coexpression networks. This was done in the table enclosed as **Fig. 4D**, where the results of the union and the intersection of the datasets provided by the two methods at each PPV are presented. The union with accuracy ≥0.60 provides a full confident and cross-validated human gene coexpression network that includes 3327 genes and 15841 coexpression links. As indicated bellow, we have analyzed in detail a core transcriptomic network that corresponds to the intersection of both methods with accuracy ≥0.60 and includes 731 gene nodes and 2249 coexpression links.

Biological significance of the coexpression datasets: house-keeping gene pairs and tissue-specific gene pairs

Once significant human gene coexpression datasets have been found and evaluated using statistical parameters, we started exploring the biological meaning and functional consistency of these datasets.

In a first approach, we investigate the location of house-keeping gene pairs in the coexpression datasets, taking two different published compendiums of human house-keeping genes [16,17]. *Hsiao et al.* identified 451 genes that are expressed in all 19 different human tissue types. *Eisenberg et al.* identified 575 human genes that show constitutive expression in all conditions tested in several publicly available databases. Mapping these genes in the general distribution of coexpression data shows that the ratio of house-keeping genes increases at high **N** and **r** coefficient values (**Fig. 5A,B**). The top panels in **Fig. 5A and B** present the density distributions of coexpression data for **N**>220 corresponding to all gene pairs (in **black**), to *Eisenberg's* house-keeping gene pairs (in **green**) or to *Hsiao's* house-keeping gene pairs (in **red**). Bottom panels in **Fig. 5A and B** show the same information including now



C

PPV [1]	Coefficients		N° of Nodes [2]	N° of Links [2]
	N	r		
MAS5-Spearman (all gn)				
≥ 0.60	605	0.77	3052	**12669**
≥ 0.70	645	0.79	2295	7874
≥ 0.80	695	0.81	1762	4910
RMA-Pearson (filtered gn)				
≥ 0.60	765	0.85	1672	**5945**
≥ 0.70	835	0.87	1215	3273
≥ 0.80	925	0.84	983	2423

1 Accuracy as PPVs that correspond to the networks derived for KEGG annotated genes
2 Values that corresponds to the full networks including all genes

D

PPV [1]	Coexpression networks			
	Union		Intersection	
	N° Nodes	N° Links	N° Nodes	N° Links
≥ 0.60	3327	**15841**	731	**2249**
≥ 0.70	2411	9264	542	1447
≥ 0.80	1863	5935	387	1008

Figure 4. Coexpression networks obtained at different levels of accuracy. Color plots **(A** and **B)** that represent the Positive Predictive Value (**PPV**) calculated for each set of gene coexpression data for different values of correlation coefficient (**r**) and cross-validation coefficient (**N**). The PPV corresponds to accumulated values for **N**≥*x* and **r**≥*y*. Calculations are done for data derived from two methods: **(A)** MAS5-Spearman without gene filtering (all gn) and **(B)** RMA-Pearson with gene filtering (filtered gn). Table **(C)** shows the specific values of correlation and cross-validation for three coexpression datasets derived from each method at 3 specific PPVs: ≥0.60, ≥0.70 and ≥0.80. This table also shows the number of nodes and links included in each coexpression dataset. Table **(D)** shows the number of gene-nodes and interaction-links that are included in the combined coexpression networks at 3 specific PPVs.
doi:10.1371/journal.pone.0003911.g004

all data points of coexpression pairs with **N**>220 and **r**>0.65 for either all gene pairs (in **black**) or only *Hsiao's* house-keeping gene pairs (in **red**). Panels **A** correspond to coexpression data obtained with method MAS5-Spearman and **B** to RMA-Pearson. The results reveal that house-keeping genes have a clear tendency to coexpress in many different tissues. This can be expected from the mere definition of house-keeping; however, since the result is obtained by mapping external datasets [16,17] on our human gene coexpression data, it provides functional validity to our coexpression study. The analysis also reveals a clear difference between the data obtained with different methods. Meanwhile MAS5-Spearman method finds mainly house-keeping gene coexpression, the RMA-Pearson method finds many gene pairs that are not in the major house-keeping region, but rather they show high levels of **r** correlation with lower levels of **N** cross-validation (**N**>220 and **N**<600).

We further investigate this observation by selecting subsets of the coexpression data for genes included in specific KEGG pathways. Examples of this subsetting are presented in **Fig. 5C**, that includes 6 panels with the coexpression data obtained with the RMA-Pearson method for the human genes included in 6 different pathways: **(1)** ribosome (KEGG ID = hsa03010), **(2)** oxidative phosphorylation (hsa00190), **(3)** proteasome (hsa03050), **(4)** cytokine-cytokine receptor interaction (hsa04060), **(5)** neuroactive ligand-receptor interaction (hsa04080), and **(6)** complement and coagulation cascades (hsa04610). First three pathways can be considered as general constitutive, present in all tissues and cellular types. The other three pathways are tissue-specific, only present in some cell types, like: nervous system cells in the case of the neuroactive ligand-receptor interaction pathway or blood cells in the case of the complement and coagulation cascades pathway. These differences in functional specificity are reflected in the coexpression distributions: only the three panels on the right (**Fig. 5C 4,5,6**) present data points with high **r** values but relatively lower **N** values (220<**N**<600). In conclusion, this analysis reveals that such coexpression pairs correspond to genes expressed in specific cells or specific tissue types, and so they are tissue-specific genes.

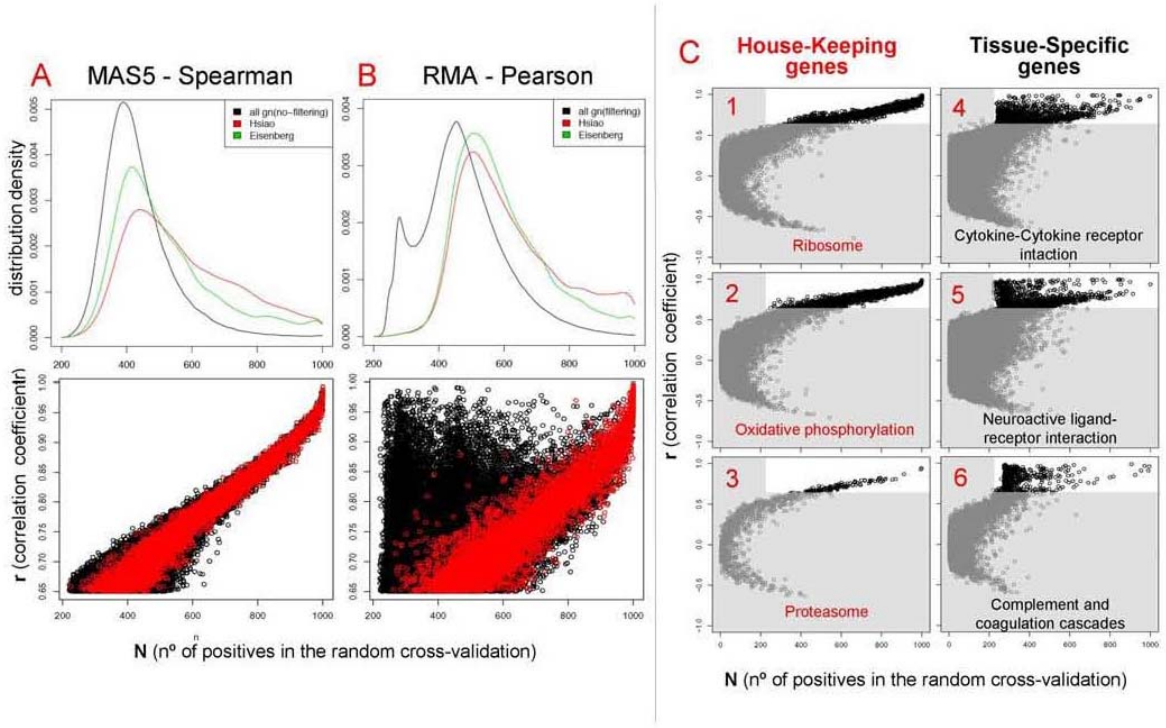


Figure 5. Coexpression of house-keeping and tissue-specific genes. Top panels A and B: Density distributions of coexpression data for N>220 corresponding to all gene pairs (in **black**), to Eisenberg's house-keeping gene pairs (in **green**) or to Hsiao's house-keeping gene pairs (in **red**). **Bottom panels A and B:** rN-plots with all data points of coexpression pairs with N>220 and r>0.65 for either all gene pairs (in **black**) or only Hsiao's house-keeping gene pairs (in **red**). In these panels **(A)** correspond to data from MAS5-Spearman method and **(B)** from RMA-Pearson method. **Panels (C)** 6 rN-plots that present the coexpression data obtained with the RMA-Pearson method corresponding to the human genes included in 6 different pathways: **(1)** ribosome (KEGG ID = hsa03010), **(2)** oxidative phosphorylation (hsa00190), **(3)** proteasome (hsa03050), **(4)** cytokine-cytokine receptor interaction (hsa04060), **(5)** neuroactive ligand-receptor interaction (hsa04080), and **(6)** complement and coagulation cascades (hsa04610). doi:10.1371/journal.pone.0003911.g005

Comparison of human coexpression datasets: molecular machines and pathways consistently co-regulated

In a second approach, we investigate the functional assignment of the gene coexpression data following the strategy taken by *Stuart et al.* [5], who explored functional coverage on a coexpression network obtained for four organisms looking at the percentage of genes that are connected to at least one other gene in the same "functional category". We proceed to the same percentage calculation using the KEGG pathways as "functional categories". The analysis was done for the coexpression dataset derived from RMA-Pearson method with **r**>0.63 and **N**>500. The same functional analysis was also done using two other external human coexpression datasets previously published by *Lee et al.* [2] and *Griffith et al.* [6].

The results are presented in **Table 1**, that includes the ten-top pathways found with best percentage of genes coexpressing within the gene groups assigned to KEGG pathways for 3 different human coexpression datasets (this work, *Lee et al.* and *Griffith et al.*). This comparative analysis of functional coverage shows some interesting results: **(i)** All coexpression datasets find the most significant coexpression for 3 key molecular machines: ribosome, proteasome and oxidative phosphorylation. **(ii)** Genes involved in cell scaffolding and cell to cell interaction or anchoring are also found to coexpress quite often, as indicated by the presence of pathways like focal adhesion, extracellular matrix (ECM) interaction and cytoskeleton regulation. **(iii)** Genes involved in cell cycle pathway are also common to the three datasets, indicating that cells keep a tight regulation of the genes involved in essential living functions (maintenance, proliferation, survival). **(iv)** An important difference between our coexpression dataset and *Lee et al.* or *Griffith et al.* datasets is that this work only includes samples coming from normal non-pathological tissues, but the others include quite heterogeneous samples mixing normal and disease altered samples (for example, *Lee et al.* includes many human cancer samples). The inclusion of pathological samples can bias the results and this may be the reason of the appearance of "pathogenic infection pathways" in *Lee et al.* data. **(v)** Finally, the data obtained in this work also includes many coexpressing pairs involved in cell-cell communication like cytokine-receptor and ligand-receptor interactions.

As a general conclusion of this analysis, we can say that KEGG pathways is revealed as a good database to investigate the biological functions of human genes, because it includes groups of genes that really work together in well defined biomolecular processes.

The comparative calculation of the coverage for the three human coexpression datasets included in **Table 1** indicates that the data obtained in this work present a higher level of functional coherence than previously published datasets [2,6]. This comparison was also done taking coexpression networks of similar sizes (including in each case around 12,000 best coexpression relations) and calculating the statistical accuracy for all of them. The result presented in **Table 2** shows that the accuracy estimated as PPV

Table 1.

This work (2008)				
Pathway Name (KEGG ID number)	**n° gn [1]**	**gn coexp/gn [2]**	**% gn coexp**	**mean r [3]**
Proteasome (3050)	31	28/28	**100.0%**	0.69
Ribosome (3010)	120	52/55	**94.5%**	0.75
Oxidative phosphorylation (190)	129	88/95	**92.6%**	0.73
Focal adhesion (4510)	194	154/168	**91.7%**	0.68
Antigen processing and presentation (4612)	86	71/78	**91.0%**	0.75
Glycan structures - degradation (1032)	30	20/22	**90.9%**	0.65
Neuroactive ligand-receptor interact. (4080)	299	227/255	**89.0%**	0.68
Cell cycle (4110)	114	90/102	**88.2%**	0.66
Regulation of actin cytoskeleton (4810)	208	141/161	**88.2%**	0.66
Cytokine-cytokine receptor interact. (4060)	256	196/223	**87.9%**	0.69
Lee et al. (2004)				
Pathway Name (KEGG ID number)	**n° gn [1]**	**gn coexp/gn [2]**	**% gn coexp**	
Ribosome (3010)	120	43/44	**97.7%**	
Proteasome (3050)	31	19/22	**86.4%**	
Oxidative phosphorylation (190)	129	31/44	**70.5%**	
Cell cycle (4110)	114	33/47	**70.2%**	
ECM-receptor interaction (4512)	87	16/23	**69.6%**	
Gap junction (4540)	92	9/13	**69.2%**	
Pathogenic Escherichia coli infection (5130)	49	11/16	**68.8%**	
Pathogenic Escherichia coli infection (5131)	49	11/16	**68.8%**	
T cell receptor signaling pathway (4660)	93	15/22	**68.2%**	
Metabolism of xenobiotics by cytP450 (980)	70	7/11	**63.6%**	
Griffith et al. (2005)				
Pathway Name (KEGG ID number)	**n° gn [1]**	**gn coexp/gn [2]**	**% gn coexp**	
Ribosome (3010)	120	36/38	**94.7%**	
Proteasome (3050)	31	20/24	**83.3%**	
Oxidative phosphorylation (190)	129	55/67	**82.1%**	
Val, Leu and isoleucine degradation (280)	50	15/19	**78.9%**	
ECM-receptor interaction (4512)	87	16/22	**72.7%**	
Cell cycle (4110)	114	36/51	**70.6%**	
Propanoate metabolism (640)	34	9/14	**64.3%**	
Butanoate metabolism (650)	44	9/14	**64.3%**	
Hematopoietic cell lineage (4640)	88	18/28	**64.3%**	
beta-Alanine metabolism (410)	24	7/11	**63.6%**	

[1]n° gn = whole number of genes included in this KEGG pathway.
[2]gn coexp/gn = genes that coexpress within the genes included for this pathway in the network.
[3]mean value of the correlation factor (r) for the coexpressing gene pairs included in this pathway.
doi:10.1371/journal.pone.0003911.t001

was 0.61 for our dataset obtained with MAS5-Spearman, 0.56 for *Lee et al.* and 0.49 for *Griffith et al.* As a whole these numbers indicate that the human coexpression network derived from this work includes very consistent co-regulation of genes many times involved in common pathways.

A high confidence human coexpression network reveals a map of ubiquitous biological functions

As far as we know, none of the previously published human coexpression networks [2,5,6] has a comprehensive calculation of the estimated statistical error in the datasets at different levels of coverage. However, following the analysis and data presented in **Figure 4** we can select coexpression datasets at specific thresholds of PPV accuracy. In order to gain in reliability, we can also combine the data obtained with 2 methods: MAS5-Spearman and RMA-Pearson. This was done taking the datasets of both methods with PPV≥0.60 (3052 and 1672 genes) to produce an intersect coexpression network that includes 731 genes and 2249 coexpression interactions (see **Fig. 4D**). We also restrict the network including only coexpressing groups including at least three genes. In this way, a high confidence core subset of 615 gene nodes and 2190 coexpression links was obtained.

Table 2.

	Nodes [1]	Links [2]	TP [3]	All [4]	PPV [5]
This work (2008)	3052	12669	729	1189	**0.613**
Lee et al. (2004)	1751	12187	1275	2265	**0.563**
Griffith et al. (2005)	2922	12686	1265	2588	**0.489**

[1]N° of genes as nodes in the network (the values correspond to the full networks including all genes).
[2]N° of coexpression links (the values correspond to the full networks including all links).
[3]True Positives = gene-pairs that coexpress and are annotated to the same KEGG.
[4]All the genes that coexpress and are annotated to KEGG.
[5]Accuracy as PPVs that correspond to the networks derived for KEGG annotated genes.
doi:10.1371/journal.pone.0003911.t002

Figure 6 presents a graphical view of this coexpression network where the nodes correspond to genes and the edges to coexpression. The network was produced introducing the coexpression dataset of 615 genes and 2190 pairwise interactions in *Cytoscape* (a bioinformatics software platform for visualizing molecular interaction networks, [18]. In the graphical view the most significant regions of this human gene coexpression network have been marked with background colors to enhance them as constellations within the coexpression landscape. Labels have been placed to each colored region to describe the main biological processes that are common to most of the genes in each region. The map shows that the larger subnetwork corresponds to genes involved in nuclear activity and nuclear-driven metabolism (region in **blue**), with a side part (in dark **blue**) that includes most of the ribosomal proteins and proteins involved in ribosomal function. The second major constellation (region in **green**) includes many genes involved in mitochondrial metabolism and redox homeostasis (like genes of the COX family, the NDUF family and the UQCR family). The third main region (in **red**) corresponds to genes involved in the immune response, genes of the major histocompatibility complex (MHC), genes that produce the cell surface clusters of differentiation (CD) and genes that encode antigen-specific molecules. Finally some smaller regions include: genes involved in metal ion homeostasis (in **grey**); genes related to the extracellular matrix and cell adhesion (in **orange**); genes related to the cytoskeleton (in **yellow**).

As a whole the network is quite stringent but it is functionally very coherent. Moreover, coming from the intersection of two methods it will be expected to include mainly essential human genes. To prove if this network is enriched in house-keeping and essential genes we identified the nodes of the network that are included in the *Hsiao* human house-keeping gene set [16] and we also identified the nodes that correspond to genes that are orthologous to known essential yeast genes (taken from SGD database). In this way, we found that the two major constellations of the network, including mainly genes involved in nuclear related and mitochondrial related metabolism, show respectively 63% and 58% of genes assigned to be house-keeping. This result reveals that the coexpression network is enriched in essential genes.

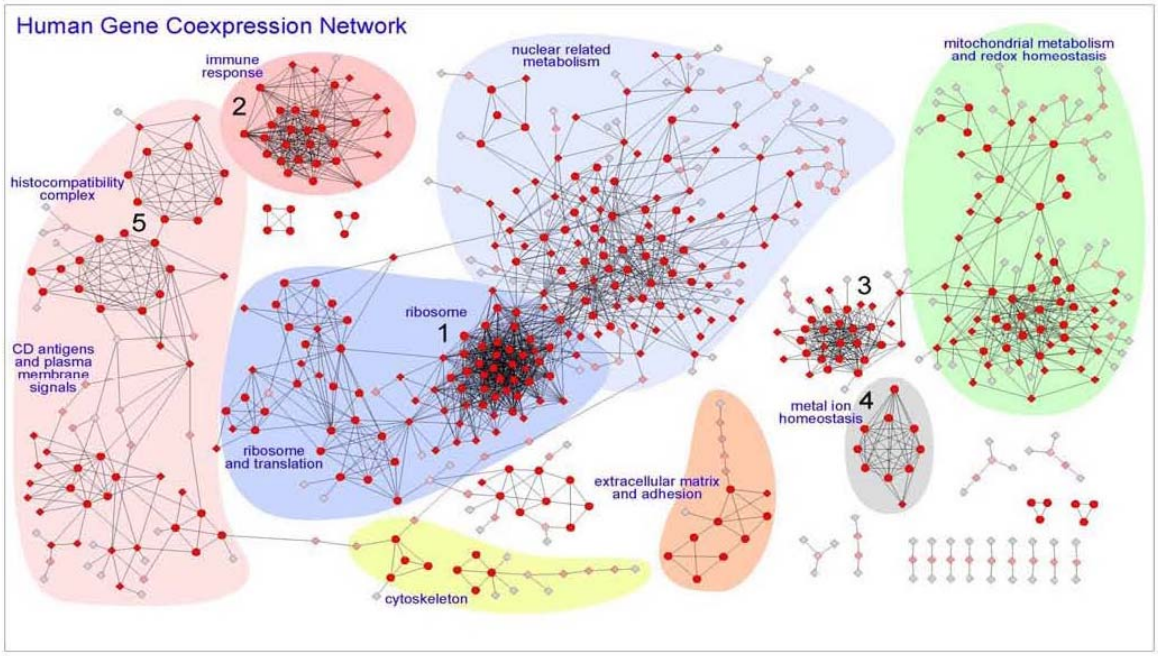


Figure 6. Human Gene Coexpression Network. Graphical view of the human gene coexpression network where the nodes correspond to genes and the edges to coexpression links. The network was produced as the intersection of two datasets (MAS5-Spearman and RMA-Pearson datasets with PPV≥0.60) to provide a confident coexpression network that includes 615 genes and 2190 pairwise coexpression interactions. The network includes only groups of coexpressing genes with at least three nodes. The most significant regions have been marked with background colors and labels describe main functions assigned. For each node the color (from **red** to **grey**) and shape (circles or diamonds) were obtained with MCODE algorithm. The circular nodes are the ones found with high cluster coefficient and the diamond nodes are the ones with lower cluster coefficient. The intensity of the **red** color in the nodes also indicates the degree of clustering, changing till pale **grey** for the most peripheral nodes that only have one link.
doi:10.1371/journal.pone.0003911.g006

In conclusion, the functional consistency observed in the constellations and regions defined by the coexpression network and the enrichment on house-keeping genes place the genes in a new integrative relational context that has strong biological coherence and, in many cases, can reveal essential or ubiquitous biological processes. The network also unravels new non-described human gene associations.

All the details about this coexpression network are provided in a supplementary file for *Cytoscape* (Supporting Information File S1: **S1_HumanCoexpNtw_615g_cys.zip**; that can be downloaded and used as a *.cys* file to be explored interactively using *Cytoscape*). This file also includes information about each node with GO and KEGG functional annotations.

Analysis of the network with clustering algorithms

The network described above was analyzed using a graph theoretic clustering algorithm called MCODE [19] as indicated in **Materials and Methods**. The result of this analysis is presented in **Figure 6**, where the circular nodes are the ones with high "cluster coefficient" and the diamond nodes are the ones with lower "cluster coefficient". The intensity of the **red** color of each node indicates the degree of clustering; changing up to pale **grey** for the most peripheral nodes (that only have one link). MCODE found 5 major gene coexpressing clusters marked with numbers in **Figure 6**: **(cluster 1)** corresponds to ribosomal genes, it includes 29 nodes and 366 links and many of the genes are RPL or RPS; **(cluster 2)** corresponds to immunoglobulins and immune response related genes (many belong to families IGH, IGK and IGL) and it includes 19 nodes and 151 interactions; **(cluster 3)** includes 19 nodes and 140 interactions and corresponds to an heterogeneous group of genes strongly clustered with no apparent common functional theme; **(cluster 4)** includes 9 nodes and 36 interactions and corresponds to genes related to metal ion homeostasis (several MT1 and MT2); and **(cluster 5)** corresponds to genes related to the major histocompatibility complex (MHC), it includes 17 nodes split in two clusters with 63 interactions, where most of the genes are HLA. There are other less dense clusters also found by MCODE that have lower score and significance for this algorithm.

We also applied another cluster algorithm for graphs called MCL [20] (see **Methods**). The analysis with MCL provided similar results to MCODE for the large clusters mentioned, although it splits the network in more clusters being the smaller ones more coherent in functional terms that the ones found by MCODE. For example, MCL algorithm finds another cluster form by 15 genes, with 7 assigned to RNA binding gene products, 3 to DNA binding gene products (all included in region **blue** in **Figure 6**), other 3 genes members of the gene family HNRP (heterogeneous nuclear ribonucleoproteins: HNRPA2B1, HNRPR, HNRPU) and 2 genes translation initiation factors (EIF3M, EIF4G2).

These results show that the gene clusters obtained with the graph algorithms from the coexpression network can help to understand the function of many human genes and the active relations between them. As expected, we find that stable and consistent coexpression clusters of genes are involved in specific functions, at cellular or systemic level. A complete analysis of all clusters is not possible in just one article but, as indicated above, the coexpression datasets of this study are open to new studies.

Functional coherence of the coexpressing modules: finding coregulation and new biological assignments

To show some specific examples about the functional coherence of the gene coexpressing modules and the adequate correlation of the genes with common regulatory elements (i.e. transcription factors, TFs, and corresponding promoters) we analyzed three specific clusters or modules found in the core coexpression network.

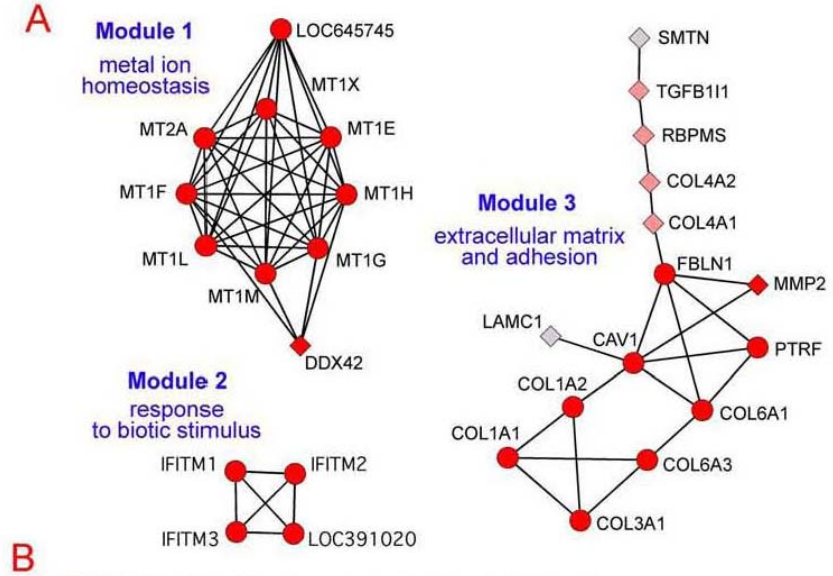
The first module includes 10 genes: 8 forming a full cross-related octogonal structure plus 2 nodes linked to them. The 8 genes are all metallothioneins: MT1E, MT1F, MT1G, MT1H, MT1L, MT1M, MT1X, MT2A. The other 2 genes are not well annotated: DDX42 (that encodes a member of the DEAD box protein family with unclear function) and LOC645745 (that has been recently and provisionally identified as a putative MT1, metallothionein 1 pseudogene 2). The coexpression of these two genes with a well defined and stable cluster of metallothioneins allows to infer that they will be genes also involved in metal ion homeostasis. This module can be seen in **Figure 7**.

A further analysis was done to find if these coexpressing genes have any common transcription factor (**TF**) that can act on the promoters and regulation regions of these genes. Two bioinformatic tools were used to find out **TFs** associated in a significant way to the coexpressing genes: PAP [21] and FactorY (see **Methods**). Using PAP we found that the 10 coexpressing genes of module 1 are regulated in common by the transcription factor MTF1 (found with p-value = 0.001). This result could be expected since MTF1 is a metal-regulatory transcription factor that induces expression of metallothioneins and other genes involved in metal homeostasis (such as zinc and copper). In any case, the association of MTF1 to module 1 provides strong coherence to the data, showing that this coexpression network is correlated with an underlying transcription regulatory entity.

The second module shown in **Figure 7** includes 4 genes: 3 correspond to interferon-induced transmembrane proteins (IFITM1, IFITM2, IFITM3) and the fourth is an unknown gene LOC391020 recently annotated by inference as similar to interferon-induced transmembrane protein 3. The coexpression of these four genes in a full related cluster gives support to the indication that all produce IFITM proteins. The analysis of transcription factors done with PAP and FactorY (**Figure 7B**) indicated that these 4 genes can be significantly correlated with the transcription factor CRE-BP1 (also called ATF2, activating transcription factor 2), that is a protein which binds to the cAMP-responsive element promoter (CRE, an octameric palindrome) and forms a homodimer or heterodimer with JUN. The deduction that IFITM genes can be coregulated by ATF2 makes biological sense because it has been observed that transcriptional activation of interferon related genes requires assembly of an enhanceosome containing the transcription factors ATF2 and JUN [22,23].

Finally, the third module shown in **Figure 7** includes 15 genes: 6 encode for collagen proteins (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL6A1) that are fibrillar proteins found in most connective tissues, related to the extracellular matrix. Other proteins within this module are also related to cell adhesion and extracellular matrix, like: Fibulin 1 (FBLN1), a secreted glycoprotein that becomes incorporated into the fibrillar extracellular matrix; Laminin gamma 1 (LAMC1), another extracellular matrix glycoprotein which is part of the major noncollagenous constituent of basement membranes; and matrix metalloproteinase 2 (MMP2), that belongs to a family of proteins involved in the breakdown of extracellular matrix in normal physiological processes and in altered disease processes. In fact MMP2 gene encodes an enzyme which degrades type IV collagen. All these data indicate functional consistency and proximity for the genes included in this coexpression module. The analysis, using PAP

A



B

Coexp Modules	Search in	TF found	p-value	TransFac_db	TF Gene Name
Module 1	*PAP*	MTF-1	**0.001**	T02354	**MTF1** metal-regulatory transcription factor 1
10 genes	*Factory*	–	–		
Module 2	*PAP*	CRE-BP1	**0.0172**	T00167	**ATF2** activating transcription factor 2
4 genes	*Factory*	CRE-BP1	**0.0033**		
Module 3	*PAP*	Sp1	**0.13**	T00759	**SP1** Sp1 transcription factor
15 genes	*Factory*	Sp1	**0.017**		

Figure 7. Coexpressed gene modules regulated by specific transcription factors. (A) Graphical enlarged view of three coexpressing modules selected from the network presented in Figure 6, indicating the name of each gene corresponding to each node and the functional labels: **(Module 1)** metal ion homeostasis; **(Module 2)** response to biotic stimulus; **(Module 3)** extracellular matrix and adhesion. **(B)** Table showing the results of the search for common transcription factors (**TFs**) most significantly associated to the genes included in each of the three modules described above. The search was done using the bioinformatic tools PAP and FactorY.
doi:10.1371/journal.pone.0003911.g007

and FactorY, of the regulatory promoters of this 15 genes shows a significant association with SP1 transcription factor, and recent experimental data have reported that in fact SP1 transcription factor is involved in the regulation of the collagen promoters [24–26].

The results presented for three coexpression modules can be extended to most of the clusters present in the network, and they indicate that the coexpression network can be correlated with an underlying regulatory network driven by specific transcription factors. This observation provides biological and functional coherence to the human gene pairwise coexpression network presented in this paper deduced from the analysis of normal-healthy human samples (whole tissues, glands or organs).

Finally, it is clear that a complete pairwise coexpression network of human genes will be only obtained using a comprehensive and systematic set of samples including all different human cell types. This achievement is at present quite far and difficult, since there are more than two hundred different cell types in the human body and that each cell type can be at different development or differentiation stages. Meanwhile, however, we think that the present study reports a reliable gene-gene coexpression network that includes very valuable information about many human genes, placing them in an integrated transcriptomic context. These coexpression networks selected at specific levels of confidence include a lot of information to better understand the complexity of the human expressing genome.

Materials and Methods

Sample selection: dataset of genome-wide expression microarrays from human normal whole tissues/glands/organs

The data used in this work corresponds to a set of human genome-wide expression microarrays hybridized with mRNA samples coming from different human tissues, glands or organs from healthy normal individuals. The complete list of tissues, glands and organs is: *adrenal gland, appendix, blood, bone marrow, brain, kidney, liver, lung, lymph node, muscle heart, ovary, pancreas, pituitary gland, prostate gland, salivary gland, skin, spinal cord, testis, thymus gland, thyroid gland, tongue, tonsil gland, trachea* and *uterus*. These 24 samples where selected from a larger set of 68 human samples (GEO GSE1133; Su et al. 2004) that also included some cell specific sources, like: lung bronchial epithelial cells HBEC, blood B-cells CD19 and T-cells CD4. The samples selection done was driven under the criteria of including mRNA samples from whole organs, glands or tissues covering the main parts of the human body and avoiding samples of very specific cell types within a tissue. This selection was validated performing global expression analyses of the samples, using a series of algorithms described bellow. The total mRNA from these 24 different samples came form a mix of 3 different individuals, that were: two men and one woman or one man and two women for the samples non sex-associated; three men for *testis* and *prostate* samples and three women for *ovary* and *uterus* samples. Moreover two biological replicates were used in each case, producing

a total set of 48 microarrays. The microarrays used were high density oligonucleotide microarrays HGU133A GeneChips from *Affymetrix*, that include 22,283 probesets (corresponding to 13,068 human genes according to *Affymetrix* annotation).

Genome-wide sample expression profiles and gene expression signatures

The global expression matrix including the genome-wide expression profiles of each sample and the expression signature of each gene-probeset was calculated and evaluated using a set of algorithms and methods in four consecutive steps: (**1st**) use of two different background correction, normalization and signal calculation methods: MAS5 [8,27] and RMA [28]; (**2nd**) use of two distance measuring methods based in the global gene expression profile of each sample: first, distance based on Spearman correlation coefficient applied to MAS5 data; second, distance based on Pearson correlation coefficient applied to RMA data (both methods provided robust non-parametric distance distributions); (**3rd**) analysis by hierarchical clustering with complete linkage of the samples using the tool *hclust* from **R** (http://www.r-project.org/), taking as distance (**1**−**r**), where **r** is the correlation coefficient between sample expression profiles [29]; (**4th**) analysis by bootstrapping of the sample hierarchical trees to assay the stability of the associations, using the tool *pvclust* from **R**. The *pvclust* algorithm allows to assess the uncertainty in hierarchical cluster analysis via multiscale bootstrap resampling. This assessment is provided by two parameters: the *approximately unbiased p-value* (AU) and the *bootstrap probability value* (BP). The maximum and optimum values of AU and BP are 1 (or 100 in %).

Gene pairs coexpression and cross-validation

As indicated above the global gene to gene (i.e. pair-wise) coexpression matrix was calculated using two different and independent methods: MAS5-Spearman and RMA-Pearson. Furtherly, cross-validation was used to discriminate stable and significant correlations. The cross-validation strategy applied was a 1000 times random selection of a 25% subset sampling (that are 12 samples, corresponding to 6 duplicates out of 24 duplicated samples) and calculation of the **r** correlation coefficient for each gene-probeset pair in such 1000 samplings. Only when the **r** correlation coefficient for a given time was higher than |0.70|, such was considered a positive event (positive cross-validation) and counted for the corresponding gene-probeset pair. In this way, for example, a given gene pair with **N** = 620 means that it gave 620 positive times out of the 1000 samplings. Therefore **N** can be considered a cross-validation coefficient or cross-validation factor (**N** = 620 is equivalent to 620/1000 = 0.62).

Gene filtering method

In order to get rid of genes with low information content a combined filter based on between-sample variability and gene minimal signal was used. The filter leaves out only those gene probesets that fulfilled both of the two following conditions: **1st**.- Genes which have an expression difference or variability between samples ($\Delta Exp^{gi}_{highest\text{-}lowest}$) lower than the median of all the expression differences calculated for each gene ($\Delta Exp^{gi}_{highest\text{-}lowest} < \text{median } \Delta Exp_{highest\text{-}lowest}$); **2nd**.- Genes which have a mean expression signal between samples ($meanExp_{samples}$) lower than the median of all the expression signals calculated for each gene.

Statistical estimation of accuracy and coverage of the coexpression datasets

The ***accuracy*** measured as "**Positive Predictive Value**" (**PPV**) in statistical terms is defined as the ratio TP/(TP+FP), where TP is the number of true positives and FP is the number of false positives [30,31]. This parameter is related to "error type I", and it is the inverse to the ratio of "false positives" (i.e. FP/(TP+FP), percentage of false positives within all the positives). The ***coverage*** (sometimes also named recall) can be measured as the proportion of true positives that remain in a given subset selected, with respect to an initial reference set of positives. We consider that both the accuracy and coverage are critical statistical parameters to evaluate the error and validity of a method. They are directly related to ***specificity*** = TN/(TN+FP), −where (TN+FP) are all the "false"−, and ***sensitivity*** = TP/(TP+FN) −where (TP+FN) are all the "true"− [30], though these can only be applied when the real true and real false data of a test are known; while the accuracy defined as "positive predictive value" and the defined coverage can be applied when it is only possible to know or estimate the "positive data".

Therefore, in this study if the true data are not known (i.e. if we do not know *a priori* which are true gene coexpressing pairs) a proper calculation of the sensitivity and specificity is not possible. This is the most common situation in many biological and biomolecular studies where many of the true occurring relations between molecules are not yet known. Therefore, we need to design a way to at least estimate the percentage or ratio of "true positives" of the method, and so estimate the accuracy and coverage. These parameters will provide a good indication of how valuable is the method that we have applied to find human coexpressing gene pairs. The estimation was done considering the idea that genes that work together in the same biological pathway are much more likely to coexpress than genes that are not involved in a common biological reaction or pathway. This biomolecular axioma in our case was tested annotating all the genes of the microarrays to the KEGG pathway database (www.genome.jp/kegg/), that is one of the most complete and expert curated repository of human genes involved in biological reactions or pathways [32]. Therefore, selecting only the subset of the genes annotated to KEGGs, a gene coexpression pair was considered a "true positive" when both genes of the pair were included in a common KEGG human pathway. This strategy allows to calculate the statistical parameters ***accuracy*** and ***coverage*** defined above, and therefore to explore how the values of the **r** and **N** coefficients change such parameters.

Analytic algorithms to find groups and modules in the coexpression networks

The gene to gene coexpression networks obtained were analyzed using a graph theoretic clustering algorithm called MCODE (Molecular Complex Detection) [19] that allows to detect densely connected regions in large interaction networks which may represent molecular associations. This algorithm follows a vertex weighting by local neighbourhood density and outward traversal from locally dense seed nodes to isolate the dense regions. Furthermore, the networks were also analyzed using another cluster algorithm for graphs called MCL (Markov Cluster algorithm, http://micans.org/mcl/) [20] that finds cluster structure in graphs by a mathematical bootstrapping procedure. MCL has been shown very robust to find relevant modules in protein interaction networks [33].

Mapping transcription factors associated to gene coexpressing modules

Two bioinformatic tools were used to find out transcription factors that can be associated in a significant way to groups or modules of coexpressing genes: Promoter Analysis Pipeline (PAP) and Transcription Factor Enrichment Analysis (FactorY).

PAP is based in a systematic, statistical model of mammalian transcriptional regulatory sequence analysis and it is suitable for the identification of the potential transcriptional regulators of co-expressed genes and the identification of the potential regulatory targets of transcription factors. A typical PAP analysis includes input of a co-expressed gene cluster, identification of several high scoring transcription factors and visualization of the predicted transcription factor binding sites [21]. The bioinformatic tool is at: http://bioinformatics.wustl.edu/webTools/portalModule/PromoterSearch.do.

FactorY is another bioinformatic tool that explores the 1000 bp upstream sequence signature of co-expressed genes to find homology with transcription factor binding sites (TFBs) based on JASPAR and TRANSFAC databases. The tool calculates the significant enrichment in known given TFBs for a group of genes and it was used at the web site: http://www.garban.org/factory/.

Supporting Information

File S1 Human Gene Coexpression Network. Network that corresponds to the core with the most confident human gene pairwise coexpression data and includes 615 gene-nodes and 2190 coexpression-links. This network is provided in Cytoscape format (.cys file compressed as .zip) with full annotations about the genes. The file to be run in Cytoscape should have .cys extension: S1_HumanCoexpNtw_615g.cys

Found at: doi:10.1371/journal.pone.0003911.s001 (0.30 MB ZIP)

Acknowledgments

We thank the support provided by the *Instituto de Salud Carlos Tercero*, Ministery of Health, Spanish Government (ISCiii-FIS, MSyC) and by the *Consejería de Educación*, Castilla-Leon Local Government (JCyL).

Author Contributions

Conceived and designed the experiments: CP JDLR. Performed the experiments: CP AR CF. Analyzed the data: CP AR CF. Contributed reagents/materials/analysis tools: CP AR CF. Wrote the paper: JDLR.

References

1. van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5: 280 284.
2. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085 1094.
3. Tirosh I, Weinberger A, Carmi M, Barkai N (2006) A genetic signature of interspecies variations in gene expression. Nat Genet 38: 830 834.
4. Magwene PM, Kim J (2004) Estimating genomic coexpression networks using first-order conditional independence. Genome Biol 5: R100.
5. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249 255.
6. Griffith OL, Pleasance ED, Fulton DL, Oveisi M, Ester M, et al. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. Genomics 86: 476 488.
7. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185 193.
8. Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics 23: i282 288.
9. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540 1542.
10. Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A (2007) Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. Bioinformatics 23: 2088 2095.
11. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res 33: 5914 5923.
12. Dallas PB, Gottardo NG, Firth MJ, Beesley AH, Hoffmann K, et al. (2005) Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR how well do they correlate? BMC Genomics 6: 59.
13. Choi JK, Yu U, Yoo OJ, Kim S (2005) Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 21: 4348 4355.
14. Prieto C, Rivas MJ, Sanchez JM, Lopez-Fidalgo J, De Las Rivas J (2006) Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. Bioinformatics 22: 1103 1110.
15. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, et al. (2007) Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic Acids Res 35: e102.
16. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, et al. (2001) A compendium of gene expression in normal human tissues. Physiol Genomics 7: 97 104.
17. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. Trends Genet 19: 362 365.
18. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498 2504.
19. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.
20. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575 1584.
21. Chang LW, Fontaine BR, Stormo GD, Nagarajan R (2007) PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis. Nucleic Acids Res 35: W238 244.
22. Falvo JV, Parekh BS, Lin CH, Fraenkel E, Maniatis T (2000) Assembly of a functional beta interferon enhanceosome is dependent on ATF-2-c-jun heterodimer orientation. Mol Cell Biol 20: 4814 4825.
23. Panne D, Maniatis T, Harrison SC (2004) Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. Embo J 23: 4384 4393.
24. Kypriotou M, Beauchef G, Chadjichristos C, Widom R, Renard E, et al. (2007) Human collagen Krox up-regulates type I collagen expression in normal and scleroderma fibroblasts through interaction with Sp1 and Sp3 transcription factors. J Biol Chem 282: 32000 32014.
25. Magee C, Nurminskaya M, Faverman L, Galera P, Linsenmayer TF (2005) SP3/SP1 transcription activity regulates specific expression of collagen type X in hypertrophic chondrocytes. J Biol Chem 280: 25331 25338.
26. Poree B, Kypriotou M, Chadjichristos C, Beauchef G, Renard E, et al. (2008) Interleukin-6 (IL-6) and/or Soluble IL-6 Receptor Down-regulation of Human Type II Collagen Gene Expression in Articular Chondrocytes Requires a Decrease of Sp1{middle dot}Sp3 Ratio and of the Binding Activity of Both Factors to the COL2A1 Promoter. J Biol Chem 283: 4850 4865.
27. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, et al. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 18: 1593 1599.
28. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
29. Murtagh F (1985) Multidimensional Clustering Algorithms. COMPSTAT Lectures. Wuerzburg: Physica-Verlag.
30. Loong TW (2003) Understanding sensitivity and specificity with the right side of the brain. Bmj 327: 716 719.
31. Suojanen JN (1999) False false positive rates. N Engl J Med 341: 131.
32. Aoki-Kinoshita KF, Kanehisa M (2007) Gene annotation and pathway mapping in KEGG. Methods Mol Biol 396: 71 91.
33. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

- Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, **Prieto C**, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology 25(8)*: 894-8.

PERSPECTIVE

nature biotechnology

The minimum information required for reporting a molecular interaction experiment (MIMIx)

Sandra Orchard[1], Lukasz Salwinski[2], Samuel Kerrien[1], Luisa Montecchi-Palazzi[1], Matthias Oesterheld[3], Volker Stümpflen[3], Arnaud Ceol[4], Andrew Chatr-aryamontri[4], John Armstrong[5], Peter Woollard[5], John J Salama[6], Susan Moore[6,7], Jérôme Wojcik[8], Gary D Bader[9], Marc Vidal[10], Michael E Cusick[10], Mark Gerstein[11], Anne-Claude Gavin[12], Giulio Superti-Furga[13], Jack Greenblatt[9], Joel Bader[14], Peter Uetz[15], Mike Tyers[16], Pierre Legrain[17], Stan Fields[18], Nicola Mulder[19], Michael Gilson[20], Michael Niepmann[21], Lyle Burgoon[22], Javier De Las Rivas[23], Carlos Prieto[23], Victoria M Perreau[24], Chris Hogue[6], Hans-Werner Mewes[3], Rolf Apweiler[1], Ioannis Xenarios[8], David Eisenberg[2], Gianni Cesareni[4] & Henning Hermjakob[1]

A wealth of molecular interaction data is available in the literature, ranging from large-scale datasets to a single interaction confirmed by several different techniques. These data are all too often reported either as free text or in tables of variable format, and are often missing key pieces of information essential for a full understanding of the experiment. Here we propose MIMIx, the minimum information required for reporting a molecular interaction experiment. Adherence to these reporting guidelines will result in publications of increased clarity and usefulness to the scientific community and will support the rapid, systematic capture of molecular interaction data in public databases, thereby improving access to valuable interaction data.

[1]European Molecular Biology Laboratory (EMBL) – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK. [2]UCLA–US Department of Energy Institute for Genomics & Proteomics, University of California, Los Angeles, California, USA. [3]Institute for Bioinformatics, Forschungszentrum für Umwelt und Gesundheit – National Research Center for Environment and Health, Neuherberg, Germany. [4]Department of Molecular Biology, University of Rome Tor Vergata, Rome, Italy. [5]GlaxoSmithkline R&D, Stevenage, UK. [6]Blueprint Initiative, Samuel Lunenfeld Research Institute, Ontario, Canada. [7]National University of Singapore, Clinical Research Centre, Singapore. [8]Merck Serono International S.A., Geneva, Switzerland. [9]Banting and Best Department of Medical Research, University of Toronto, Ontario, Canada. [10]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [11]Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut, USA. [12]EMBL Heidelberg, Germany. [13]CeMM Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. [14]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [15]Institute of Toxicology and Genetics, Leopoldshafen, Forschungszentrum Karlsruhe, Germany. [16]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada. [17]Commissariat à l'Energie Atomique, Institut de Biologie et de Technologie de Saclay, Gif sur Yvette, France. [18]Howard Hughes Medical Institute Department of Genome Sciences & Medicine, University of Washington, Seattle, Washington, USA. [19]Institute for Infectious Disease and Molecular Medicine, University of Cape Town, South Africa. [20]University of Maryland Biotechnology Institute, Rockville, Maryland, USA. [21]University of Giessen, Germany. [22]Toxicogenomic Informatics and Solutions, Lansing, Michigan, USA. [23]Cancer Research Center (Centro de Investigación del Cancer, University of Salamanca and Consejo Superior de Investigaciones Científicas), Salamanca, Spain. [24]Centre for Neuroscience, University of Melbourne, Victoria, Australia. Correspondence should be addressed to S.O. (orchard@ebi.ac.uk).

Published online 8 August 2007; doi:10.1038/nbt1324

Deciphering the molecular mechanisms of cell function relies to a large extent on tracing the multitude of interactions between the numerous components of living cells, and between these molecules and any entity or compound of interest to the scientist, such as pharmaceutical agents or environmental contaminants. Molecular interactions may be direct, with two molecules in contact with each other, or the molecules may be in the same affinity complex, purifying together without a physical interaction between them. Several public databases strive to capture the ever increasing amount of published molecular interaction data, which are generated by a broad range of biophysical, biochemical, genetic or predictive methods. During the process of manual curation, the raw data are extracted from a published paper or from a submitted manuscript and systematically transferred into a database.

Initially, interaction databases such as BIND[1] and DIP[2] worked in isolation and according to their own internal standards and data formats. Because no one database can achieve complete coverage of all known molecular interactions, the user may need to download and combine datasets from two or more databases to answer a specific question. Until recently, this could not be done without first transforming the data into a common format, using a different parser for each database. In 2004, however, several major databases jointly published a community-standard data model for the representation and exchange of protein interaction data[3]. This data model, developed by members of the Molecular Interaction (MI) group of the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO)[4], has already been adopted by major public interaction databases. Data sets can be downloaded from many of these databases in PSI-MI extensible markup language (XML) interchange format and further analyzed using a number of PSI-MI compatible tools, such as Cytoscape[5], ProViz[6] and PIMWalker[7].

Building on the PSI-MI standard, several public interaction databases have formed the International Molecular Interaction Exchange consortium (IMEx; http://imex.sf.net). The consortium, originally founded by BIND[1], DIP[2], IntAct[8], MINT[9] and MPact (MIPS)[10], has started to share the curation load and aims to regularly interchange data curated to the same common standards, in a manner similar to the well established pattern followed by the nucleotide sequence databases. However, the consortium's goal of achieving as near complete

Box 1 The MIMIx checklist

Example data are taken from ref. 21. This example may be seen as a completed checklist in **Supplementary Note 2** online and as a MIMIx-compatible submission in Excel, XML and HTML format at http://imex.sf.net. The full paper, annotated to the richer IMEx standard, can be viewed in the IntAct database (http://www.ebi.ac.uk/intact) using accession numbers EBI-958406, EBI-958452, EBI-958498 and EBI-959602.

- Submission
  A submission should contain the essential administrative information:
  - Contact email
  - Publication title
  - First author
  - Publication identifier (if manuscript is not yet submitted, authors, or the recipient, may substitute an internal tracker here)
- Experiment
  Each experimental setup should be described separately, with the following parameters:
  - Host system
    The host organism in which the interaction took place, identified by the NCBI taxonomy identifier
    Example: Yeast (TaxID:4932)Further specification of cell line or tissue is recommended
  - Interaction detection method
    The method by which the interaction was detected
    Root term MI:0001
    Example: two hybrid (MI:0018)
  - Participant identification method
    The method by which the interaction participants were determined
    Root term MI:0002
    Example: nucleotide sequence (MI:0078)
- Interaction
  - Participant list
    The list of all molecules participating in the interaction. The list can contain any number of elements. Each molecule should be characterized by:
    - Database
      Root term: MI:0444
      Example: UniProt (MI:0486)
    - Accession number from that database
      Example: P48551
    - Version number (optional)
    - Name
      The common name of the molecule used in the manuscript
      Example: IFN-αRβL
    - The species of origin for the molecule identified by NCBI taxonomy identifier
      Example: 9606
    - Biological role
      The biological role of the molecule in the interaction
      Root term: MI:0500
      Example: neutral component (MI:497)
    - Experimental role
      The experimental role of the molecule in the interaction
      Root term: MI:0495
      Example: bait (MI:0496)
  - Confidence (optional)
    A confidence value attributed to the interaction
    If a confidence value has been assigned, the confidence attribution system must be described in the manuscript. Ideally, the raw data for the confidence assignment should be available.

Controlled vocabularies are an essential part of the characterization of a molecular interaction in PSI-MI format. Elements of these controlled vocabularies are referred to as MI:xxxx above. The complete controlled vocabularies can be accessed at http://www.psidev.info/index.php?q=node/31 or interactively at http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI (ref. 18).

coverage as possible of interaction data in the literature is greatly hindered by inconsistencies and missing information in published papers. The absence of key pieces of information can lead both to misinterpretation of the paper by scientists and to a time-consuming, error-prone attempt to derive the missing information by a database curation team. Often, the reason for such information deficits is simply the lack of a community consensus on what information is required to appropriately describe a molecular interaction.

To address this issue, we have developed MIMIx as a basis for discussion. MIMIx represents a compromise between the depth of information necessary to describe all relevant aspects of an interaction experiment and the reporting burden placed on scientists who generate data. Its purpose is to ensure that the bench scientist has a checklist (**Box 1**) of the information to be supplied when describing experimental molecular interaction data in a journal article, displaying data on a website or depositing data directly into a public database (**Box 2**).

A MIMIx-compliant dataset is not intended to allow an interaction experiment to be reproduced from a database record but to enable database users to quickly assess and focus on data relevant to them and then link to the source publications for the full experimental context. On the other end of the complexity scale, the PSI-MI XML interchange format, which is adopted by all IMEx partners, provides for a much richer representation of a molecular interaction experiment than that required by MIMIx. IMEx partners also welcome data submissions that use the full complexity of the PSI-MI format.

Molecules

The single greatest source of data loss in transferring interaction data into a database is the use of ambiguous molecule identifiers, such as gene names. According to anecdotal estimates from database curators, as much as 70% of overall curation time is spent mapping molecule identifiers unambiguously to well characterized database entries. For example, a paper may not indicate both the gene name and the species from which the gene originated. This information is implicit in the molecule identifiers generated by the major databases. The description 'lck cloned in a mammalian expression vector' gives no indication as to whether the protein source is human, mouse, bovine or rat. 'Human $p56^{lck}$ protein' gives information about the species but not about the splice isoform, whereas both species and sequence are provided by the accession numbers UniProtKB[11] P06239

and RefSeq[12] NP_005347, and P06239-1 gives a full description of a specific isoform. 'Human PI3-kinase p85 subunit' may appear to be a unique reference, but does it refer to the alpha subunit (P27986) or the beta subunit (O00459), which are two distinct gene products? Such errors will almost certainly result in the paper in question not being added to a curated dataset and may also mislead the reader regarding the actual construction of the experiment. Similarly, it is important for authors to state whether an interaction described in one organism was modeled from an interaction detected between similar molecules in a related organism; for example, an interaction between a rat and a human protein being used to infer a human-human protein interaction. The constructs used, including the organism of origin of the sequence and the splice variant, should be clearly described.

We therefore request that all molecules be identified by a database accession number from a public database. For proteins, UniProt or RefSeq are strongly recommended; for genes, Ensembl[13] or Entrez Gene[14]; for chemical entities, PubChem[14] or ChEBI[15]. Nucleotide sequence database accession numbers (DDBJ, EMBL or GenBank, http://www.insdc.org) identify specific transcripts and give additional information as to the source and the class of nucleic acid under investigation. Where a molecule description is not available from these databases, identifiers from other public databases, such as model-organism databases, may be used. For a full list of recommended databases, please refer to the relevant section of the PSI-MI controlled vocabulary (see below), which also provides unified names for these resources.

An annotated protein or nucleic acid sequence may vary with time as the original submitters update their coding sequence prediction programs, frameshifts are identified, and correction or resequencing is undertaken. This may invalidate the mapping of specific sequence positions; for example, those where deletion mutants or binding domains are described. We therefore request the addition of version numbers, either of the molecule (for example, P06239.5) or of the database, to the MIMIx record.

Although the identification of molecules by accession number is precise, it may be unwieldy to refer to 'UniProt:P06239.5' instead of 'lck' in the text of a paper. To satisfy the need for both precision and readability, we recommend that the accession number and the molecule name used in the text be associated either in the submitted database record or at least at the first occurrence in the paper (for example, "…lck (UniProt: P06239.5)…").

A key element in the description of an interaction experiment is the role a molecule has in the interaction. MIMIx requests the classification of the molecule role in two ways: the biological role, for example, enzyme or enzyme target; and the experimental role, for example, bait or prey. For both of these, the PSI-MI standard defines a comprehensive controlled vocabulary, ensuring that the same term, rather than synonyms or alternative spellings, is used throughout a paper and that the interpretation of the meaning of that term remains fixed. A list of controlled vocabulary terms that describe the various methods used to detect molecular interactions, current as *Nature Biotechnology* went to press, is available in **Supplementary Note 1** online. This list undergoes continual revision as technologies evolve; the latest version is available at http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI.

Finally, it should be noted that databases describe the canonical form of a molecule. The actual participant in a molecular interaction may have been altered, either naturally by the cell (e.g., by cleavage of a bioactive peptide from a precursor protein) or by engineering (e.g., by addition of a tag or creation of a deletion mutant). Terms to describe the 'participant' in an interaction as a derivative of the 'molecule' in the database entry are available in the PSI-MI controlled vocabularies.

Box 2 Frequently asked questions about the MIMIx guidelines

To clarify when researchers should use MIMIx and some frequently asked questions about the guidelines, we provide select questions and answers below:

When should I use MIMIx?
- When describing any molecular interaction within a paper that you are writing
- When preparing your data for submission to any MIMIx-compliant database

How should I describe my experiment in a way that is clear to the reader?
- Use the appropriate terms from the PSI-MI controlled vocabulary; if you cannot find your technique listed, request a new term
- Add any additional detail to Materials and Methods as usual, or in a free-text description in a database submission

Why do I need to use accession numbers; surely a gene name is enough?
- Gene names often change with time, and their use may lead to ambiguities in the data
- Gene names give no indication as to originating species, and this important information is often missing from papers

What is meant by 'participant role'?
- In many experiments, a specific protein will be modified or specifically targeted by antibodies and then used to capture its interacting partners; this is a bait-prey relationship
- We can also infer that proteins are physically interacting if they have a biological relationship, such as enzyme–enzyme target
- In many cases, for example, cosedimentation, no such relationships can be assigned, and these interactions are described as 'neutral' in MIMIx terminology

Why should I submit my results to a MIMIx-compatible database; surely writing a paper is enough?
- Data deposition ensures that your data are available, in a downloadable format, to the entire interactome community
- It increases the visibility and readership (and thus potentially citations) of your paper
- It results in your data being available through many other routes, such as the UniProtKB database
- The database can help you provide data to the journal, give you accession numbers and ensure long-term storage of your data

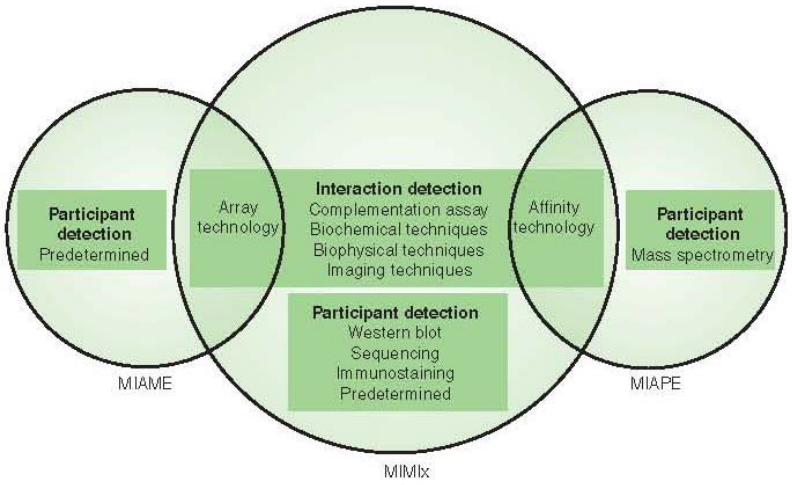


Figure 1 The relationship of MIMIx to two guidelines that may be relevant to molecular interaction studies, MIAME[17] and MIAPE[16]/MIAPE-MS. Almost all interaction data may be described using MIMIx; however, MIAME provides guidelines for describing a microarray experiment, and MIAPE allows the submitter to supply details of the peptides and underlying spectra when mass spectrometry has been used to identify protein participants.

Experiment

The MIMIx experiment description implements the core requirements of PSI's "minimum information about a proteomics experiment" (MIAPE) guidelines[16] (see p. 887) and aims to capture the aspects of an interaction experiment that are necessary to classify and critically assess the results and the interpretation of the results. It is likely to be further refined in the future as other technology-specific MIAPE modules evolve. The attributes we consider essential at present are as follows:

The 'host organism' describes the system in which the interactions were detected. The host organism should be described by a National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA) taxonomy identifier and should contain further specification, such as cell-line or tissue descriptors. When the experiment was performed *in vitro*, this should be described as free text.

The 'interaction detection method' describes the method by which the interaction was determined (for example, tandem affinity purification (MI:0676)).

The 'participant detection method' names the experimental procedure for the detection of the molecules participating in the interaction (for example, peptide mass fingerprinting (MI:0082)).

Beyond these essential requirements, we recommend that authors provide additional detail on molecule sources, sample preparation and further relevant experimental parameters using the detailed controlled vocabularies provided by the PSI-MI standard.

Interaction

The PSI-MI standard provides a formal frame for a detailed description of an interaction, including both qualitative parameters, such as details of mutations, and quantitative parameters, such as dissociation constants. However, these data are often not available, and, thus, MIMIx requires only one element for the description of an interaction: the list of molecules participating in it, characterized as above. If a quality assessment was carried out, the confidence value assigned to the interaction and the confidence attribution system must also be included in the manuscript. Particularly in large-scale experiments, interactions are usually assigned a quality score, which might be derived from data collected in the experiment itself or from additional data outside the experiment. Inclusion of interaction data in public databases requires that this reliability score be easily accessible. Ideally, not only the score but also the raw data used to derive the score should be reported so that users can perform alternative quality assessments.

Relationships to other biological standards

The MIMIx guidelines have been developed in close collaboration with related standards bodies within both the HUPO-PSI and the wider community, in consultation with contributors to the MIAME microarray standards[17]. MIMIx is one of a series of modules developed within the framework of the MIAPE guidelines[16]. When an interaction experiment encompasses experimental data that are more fully described by other modules, authors should refer to the relevant guidelines when preparing their data for submission to a journal or a database (**Fig. 1**). For example, identification of prey proteins in a tandem affinity purification (TAP) pulldown by mass spectrometry should be described according to the MIAPE-MS guidelines (C.F. Taylor *et al.*, unpublished).

Similarly, the HUPO-PSI and the Microarray Gene Expression Data (MGED) consortium are jointly working to provide guidelines for the annotation of array data that will ensure a smooth path for the annotation and submission of such data. Overarching all of these standards is the "functional genomics experiment" (FuGE) model[18], which can be used to provide protocols and data flow models should the user wish to annotate such detail. All these guidelines are being managed through a central repository of standards, as described in the "minimum information for biological and biomedical investigations" (MIBBI)[19], to ensure that they are complementary and nonoverlapping.

Data deposition

Curators of the main molecular interaction databases work to collect and archive data from journal publications. Although a systematic reporting of published interaction data according to the above guidelines would greatly increase the efficiency of the curation task, literature curation after publication is only a second-best option. We therefore recommend that all reported interaction data be deposited in a publicly available molecular interaction database before publication.

Data deposition has benefits for all parties involved. The databases will be able to work more efficiently and will have more direct access to the data producer to resolve unclear issues. The scientific community will benefit from more, and more precise, information in the databases, as database records can be checked directly by the data producer. Journals and data producers will benefit from consistently formatted database records, which can be included in the supplementary material of a publication. Accession numbers issued by a database and included in the journal publication will allow direct access to the data in the database and a quick connection to related data in the database, such as other records on the same molecules. Finally, data producers and journals will gain exposure for the publication through cross-references from the database records.

IMEx databases offer several options for data deposition (http://imex.sf.net/deposition.html). The submission of fully formatted PSI-MI

XML files is recommended for large-scale data producers, who usually have the data available in in-house databases anyway. For smaller-scale experiments, a preformatted Microsoft Excel spreadsheet file is available, with instructions on how to complete it. In addition to technical systems, such as the Ontology Look-up Service (OLS) browser[20] and a system for the automatic validation of PSI-MI XML files (http://www.ebi.ac.uk/intact/validator), database curation teams provide assistance in all stages of the data deposition process, for example, in the correct use of the detailed controlled vocabularies used to characterize an interaction. We particularly encourage early contacts with database curation teams, to embed appropriate data collection protocols into the experiment-planning stage.

In addition to the biological data, each data deposition must be accompanied by the minimal administrative data, namely contact email, publication title, first author and the publication identifier, usually a PubMed or Digital Object (http://www.doi.org) identifier. In the prepublication stage, a journal-specific identifier can be used to provide a unique identification of the manuscript accompanying the data deposition; before manuscript submission, the authors may use their own in-house identifier.

To optimize the use of public resources, IMEx partners have developed common curation guidelines and have agreed to synchronize their curation work and exchange all user-submitted data so as to build up a network of stable, well coordinated molecular interaction databases freely accessible to the community. Although accession numbers for deposited interactions will be issued within five working days of the provision of all necessary data, deposited data will be released only upon publication of the associated manuscript or at the request of the data provider.

Conclusion

The MIMIx guidelines presented here will not be static. They will evolve based on community requirements in the context of a rapidly developing science. This document has been assembled by a large number of experts and subjected to public review both on the PSI website and through *Nature Biotechnology* community review. At all stages, we have discussed input and fed it back into the document. The MIMIx guidelines, PSI-MI XML interchange format and the corresponding controlled vocabularies are all maintained and updated through the PSI-MI workgroup using mailing lists, issue trackers and annual workshops. If you wish to make specific comments on the MIMIx guidelines, please use the issue tracker at http://www.psidev.info/index.php?q=node/279 or, for a wider involvement, refer to the mailing lists at http://www.psidev.info/.

Note: Supplementary information is available on the Nature Biotechnology website.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
2. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32** (Database issue), D449–D451 (2004).
3. Hermjakob, H. *et al.* The HUPO-PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
4. Orchard, S. *et al.* Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, 4–6 September 2005. *Proteomics* **6**, 738–741 (2006).
5. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
6. Iragne, F., Nikolski, M., Mathieu, B., Auber, D. & Sherman, D. ProViz: protein interaction visualization and exploration. *Bioinformatics* **21**, 272–274 (2005).
7. Meil, A., Durand, P. & Wojcik, J. PIMWalker: visualising protein interaction networks using the HUPO PSI molecular interaction format. *Appl. Bioinformatics* **4**, 137–139 (2005).
8. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35** (Database issue), D561–D565 (2007).
9. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35** (Database issue), D572–D574 (2007).
10. Pagel, P. *et al.*, The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
11. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35** (Database issue), D193–D197 (2007).
12. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31**, 34–37 (2003).
13. Hubbard, T.J. *et al.* Ensembl 2007. *Nucleic Acids Res.* **35** (Database issue), D610–D617 (2007).
14. Wheeler, L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34** (Database issue), D173–D180 (2006).
15. de Matos, P. *et al.* ChEBI - Chemical Entities of Biological Interest, Database Summary Paper 646, *Nucleic Acids Res.* <http://www3.oup.co.uk/nar/database/summary/646> (2006).
16. Taylor, C.F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**, 887–893 (2007).
17. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
18. Jones, A.R., Pizarro, A., Spellman, P., Miller, M. & FuGE Working Group. FuGE: Functional Genomics Experiment object model. *OMICS* **10**, 179–184 (2006).
19. Taylor, C.F. *et al.* Promoting coherent minimum reporting requirements for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* (in the press).
20. Cote, R.G., Jones, P., Apweiler, R. & Hermjakob, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* [online] **7**, 97 (2006).
21. Croze, E. *et al.* Receptor for activated C-kinase (RACK-1), a WD motif-containing protein, specifically associates with the human type I IFN receptor. *J. Immunol.* **165**, 5127–5132 (2000).

- Hernández-Toro J, **Prieto C**, De Las Rivas J (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics 23(18)*: 2495-7.

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 18 2007, pages 2495–2497
doi:10.1093/bioinformatics/btm373

Systems biology

APID2NET: unified interactome graphic analyzer

Juan Hernandez-Toro, Carlos Prieto and Javier De Las Rivas*

Bioinformatics and Functional Genomics Research Group, Cancer Research Center (IBMCC-CIC, CSIC-USAL), Salamanca, Spain

Received on April 13, 2007; revised on June 7, 2007; accepted on July 11, 2007

Advance Access publication July 21, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Exploration and analysis of interactome networks at systems level requires unification of the biomolecular elements and annotations that come from many different high-throughput or small-scale proteomic experiments. Only such integration can provide a non-redundant and consistent identification of proteins and interactions. APID2NET is a new tool that works with Cytoscape to allow surfing unified interactome data by querying APID server (http://bioinfow.dep.usal.es/apid/) to provide interactive analysis of protein–protein interaction (PPI) networks. The program is designed to visualize, explore and analyze the proteins and interactions retrieved, including the annotations and attributes associated to them, such as: GO terms, InterPro domains, experimental methods that validate each interaction, PubMed IDs, UniProt IDs, etc. The tool provides interactive graphical representation of the networks with all Cytoscape capabilities, plus new automatic tools to find concurrent functional and structural attributes along all protein pairs in a network.

Availability: http://bioinfow.dep.usal.es/apid/apid2net.html

Contact: jrivas@usal.es

Supplementary information: Installation Guide and User's Guide are supplied at the Web site indicated above.

1 INTRODUCTION

The advancement of genome and proteome-wide experimental technologies have introduced modern biology in the high complexity of living cells, where thousands of biomolecules work together in an interactive way, with many short and long range cross-talks and cross-regulations. To achieve a first level of understanding of such cellular complexity we need to unravel the physical interactions that occur between all the proteins that integrate an active working cell. The compendium of all known protein protein interactions (PPIs) for a given cell or organism is called the *interactome* and in recent years several key publications have used high-throughput proteomic techniques to determine the interactome in model organisms: yeast *Saccharomyces cerevisiae* (Ito *et al.*, 2001; Uetz *et al.*, 2000), fly *Drosophila melanogaster* (Formstecher *et al.*, 2005; Giot *et al.*, 2003) and worm *C.elegans* (Li *et al.*, 2004). More recently several efforts have focused on the human interactome (Rual *et al.*, 2005; Stelzl *et al.*, 2005). The PPI data coming from many publications done with high-throughput proteomic techniques or with small-scale experimental methods are being collected and stored in several databases, like DIP (Xenarios *et al.*, 2002) and IntAct (Hermjakob *et al.*, 2004). Recently we have developed a bioinformatic web server called APID (Prieto and De Las Rivas, 2006), that integrates the interactome data of five main PPI databases unifying them in a common platform. At present, June-2007, APID includes 42 699 proteins from 15 organisms and a total of 156 688 interactions. It also includes information to help in the validation of the interactions, like the number of experimental methods that prove each interaction or the existence of structural domains that may interact according to iPfam. iPfam is a resource that describes domain domain interactions that are observed in PDB entries (Finn *et al.*, 2005).

The complete interactomes are rather complex systems so, despite that APID and other web tools include graphic displays to show sections of the interactome network, such graphical visualizations are quite limited when a specific study of a proteome subset or a protein family wants to be undertaken in detail. Moreover, many biomolecular research groups may demand specific PPIs selection to explore and analyze interactively such concrete network within the interactome landscape.

2 MINING THE INTERACTOME: CYTOSCAPE AND APID

Cytoscape is a bioinformatic software for visualizing molecular interaction networks and integrating these interactions with other biological data (http://www.cytoscape.org/). Cytoscape allows and promotes the integration of additional plugins that can provide network and profiling analyses, new layouts, connection with databases, etc. In recent years there have been several publications reporting stand-alone applications to investigate protein interaction networks, however most of them provide independent software that cannot be combined and whose development and maintenance sometimes may face uncertain future. The development of tools integrated in Cytoscape can provide useful and robust applications, since a large international community is collaborating writing advance software for this common and well-maintained platform.

Following these ideas we have developed a Java application called APID2NET that allows exploration, annotation and analysis of one or more subsets of the protein interactome. The tool is completely integrated in Cytoscape as a plugin and allows to query APID using a transparent Servlet interface and

*To whom correspondence should be addressed.

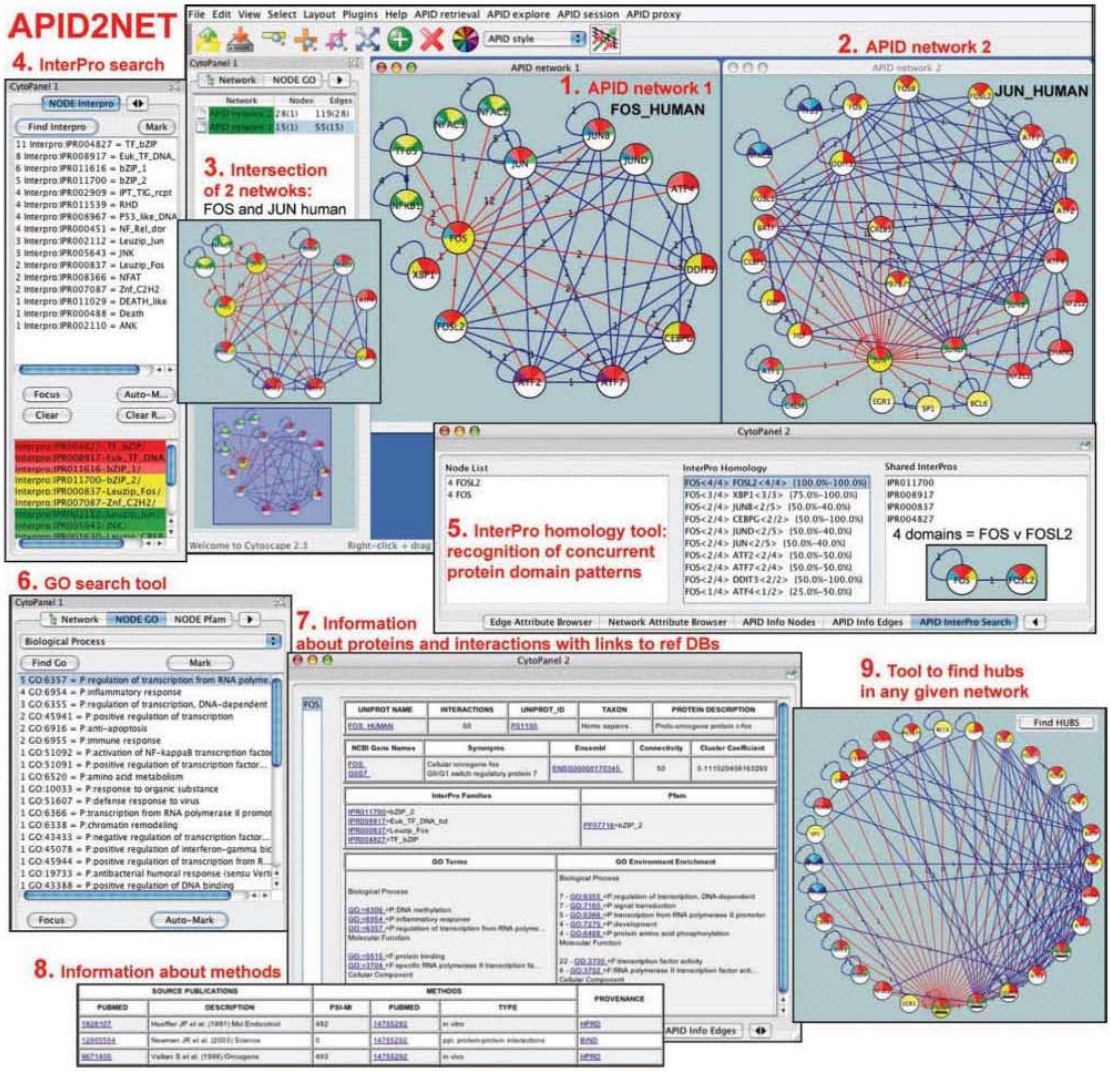


Fig. 1. Workflow of the application in nine consecutive windows. It includes an example of the protein interaction networks for two human proteins involved in transcription: FOS and JUN.

fetch the interaction network for a given protein or a given list of proteins. Once a network is downloaded it retains all the information present in APID, and includes several new features to explore and find co-occurrence of functional and structural annotations among the proteins.

3 FEATURES TO EXPLORE AND ANALYZE THE NETWORKS

APID2NET provides an interface within Cytoscape to allow on-line PPI data retrieval and further interactive analysis of nodes and edges. The main features included in the application are:

- 'APID retrieval' menu (to retrieve and visualize PPI networks):

(1) Search engine to retrieve PPI networks from APID server (i.e. query and fetch data). The search can retrieve the PPI network of a single protein or a protein list.

(2) The PPI networks can be retrieved using quality filters, like: a minimal number of experimental methods that

validate the interactions; the presence of iPfam predicted interacting domains.

(3) Simple clicking tool to expand the network from any selected node and retrieve new interactions.

- 'APID explore' menu (to explore functional and structural annotations and to locate hubs):

(1) Each network can be explored using several protein attributes: GO terms (within the three categories BP, MF or CC), Pfam sequence-based domains, InterPro domains and motifs. Each search displays the attributes ordered by the number of times that appear in the query interaction network ('CytoPanel-1').

(2) The information about each protein (node) and each interaction (edge) is displayed in tables ('CytoPanel-2'), including the experimental methods that validate each interaction with the PubMed literature references. The tables include links-out to the source files in public databases (UniProt, PubMed, GO, etc). The nodes and edges in the tables are interactive to allow easy location in the network window.

(3) 'Find' and 'Auto-Mark' (in 'CytoPanel-1' submenus): to find annotations and attributes about nodes or edges. Functional and structural attributes (GO, Pfam, InterPro) that are concurrent in at least one protein-pair can be found and color-coded using 'Auto-Mark'.

(4) 'Hubs' tool ('CytoPanel-1', 'Mark-NODE'): to locate the hubs present in a given network. The tool uses an algorithm that searches for the protein nodes that include a given percentage of the non-repetitive edges in the network (65% by default).

(5) InterPro homology tool ('CytoPanel-2', 'APID InterPro Search'): to find pair-wise protein domain pattern homology based on InterPro. The tool searches for all possible protein-pairs in a network and marks the domains and motifs from InterPro that are common to each pair. The number of common domains is counted providing a ranking from more to less similar pairs.

- 'APID session' menu (to save and reload data):

(1) Designed to save on the local PC a complete working session with the information and analyses that had been done by a user.

All the tools and features are explained in detail in the User's Guide available on-line.

4 EXAMPLE: EXPLORE AND COMPARE FOS AND JUN HUMAN NETWORKS

An example showing the main features of APID2NET is included in Figure 1, that presents the interaction networks of two human proteins involved in transcription: FOS_HUMAN (P01100) and JUN_HUMAN (P05412). The networks are filtered to retrieve only the proteins that have iPfam validation. In this way, FOS network includes 15 nodes and 28 edges; and JUN network includes 28 nodes and 119 edges. The FOS JUN interaction is well known and it is validated by 12 reported methods. The protein annotation co-occurrence discovery tool that uses InterPro, provides a clear way to find protein homologous. Window 5, in Figure 1, shows the interaction FOS FOSL2: FOSL2 is a FOS-related antigen and presents the same domain architecture as FOS. The InterPro co-occurrence analysis also shows that proteins JUN, JUNB and JUND have the same domains pattern, and in this way they are found putative paralogous proteins. The described co-occurrence tool can help to unravel functional protein relationships and to improve our understanding of the interactome modules and architecture.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges was provided by the Spanish Ministery of Health (MSyC, ISCIII, FIS grant reference PI061153) and Junta Castilla y Leon (grant reference CSI03A06).

Conflict of Interest: none declared.

REFERENCES

Finn,R.D. *et al.* (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Formstecher,E. *et al.* (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376–384.

Giot,L. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.

Hermjakob,H. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–455.

Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Li,S. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

Prieto,C. and De Las Rivas,J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.*, **34**, W298–W302.

Rual,J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Xenarios,I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

- **Prieto C** and De Las Rivas J (2006). APID: Agile Protein Interaction DataAnzalizer. *Nucleic Acids Res. 34 (Web Server issue)*: W298-302.

W298–W302 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl128

APID: Agile Protein Interaction DataAnalyzer

Carlos Prieto and Javier De Las Rivas*

Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CIC, CSIC/USAL), 37007 Salamanca, Spain

Received February 14, 2006; Revised February 23, 2006; Accepted March 14, 2006

ABSTRACT

Agile Protein Interaction DataAnalyzer (APID) is an interactive bioinformatics web tool developed to integrate and analyze in a unified and comparative platform main currently known information about protein–protein interactions demonstrated by specific small-scale or large-scale experimental methods. At present, the application includes information coming from five main source databases enclosing an unified sever to explore >35 000 different proteins and 111 000 different proven interactions. The web includes search tools to query and browse upon the data, allowing selection of the interaction pairs based in calculated parameters that weight and qualify the reliability of each given protein interaction. Such parameters are for the 'proteins': connectivity, cluster coefficient, Gene Ontology (GO) functional environment, GO environment enrichment; and for the 'interactions': number of methods, GO overlapping, iPfam domain–domain interaction. APID also includes a graphic interactive tool to visualize selected sub-networks and to navigate on them or along the whole interaction network. The application is available open access at http://bioinfow.dep.usal.es/apid/.

INTRODUCTION

Genome-wide and proteome-wide technologies on modern biochemistry and molecular biology provide vast and quickly increasing amounts of biological data that need to be stored, compared and organized using comprehensive and dynamic open access computational tools. One of the most productive areas is the one of protein–protein interactions and interactome data (1). The data about the interaction of two or more proteins come either from small-scale experimental work or from large-scale experimental methods. Both kind of data are being included in biological databases focus on protein interaction and several bioinformatic initiatives have been undertaken to this purpose [see reviews (1–3)]. However, several studies in recent years have reported comparative assessments of large-scale and high-throughput protein–protein interaction data (4,5) indicating that data quality is a critical problem in these datasets, that many times include a high proportion of false positive interactions due to low accuracy of the methods. Some bioinformatic and computational work has been done to assess the reliability of high-throughput observations and to gain confidence in the data (6–9). However, we consider that more efforts based on validated experimental information are essential to improve the quality of the protein–protein interaction data and therefore to improve the biological information that can be inferred from the interactome networks.

At present time, there are several major protein interaction databases [Biomolecular Interaction Network Database (BIND), Database of Interacting Proteins (DIP), InAct] (10–12) that are collecting the increasing amount of biological data produced in this area. Data about the interactions of two or more proteins are stored in many published scientific papers and the databases extract and integrate such information. However, each database has its own extraction, curation and storage protocols, and not all of them explore the same scientific papers. In fact, we have observed that the intersection and overlap between these source databases is small, and therefore in many cases their information is complementary and can be unified to increase and improve our knowledge about interactome networks. At the same time, the existence of several experimental evidences about many protein–protein interactions, reported by different literature references, allows to increase the number of methods that validate any given interaction. We consider that an integrative effort is essential to draw more clear maps about the protein interaction network and to explore sub-networks for specific proteins or protein families.

Keeping the key critical needs described above, i.e. (i) better assessing the quality of the protein–protein interaction data and (ii) more comprehensive integration of main currently known protein–protein interactions; we have developed an interactive bioinformatics web tool to integrate and analyze in a common and comparative platform main known protein interactomes. This web tool can be very helpful for

*To whom correspondence should be addressed. Tel.: +34 923 294819; Fax: +34 923 294743; Email: jrivas@usal.es

the research on a specific protein or protein family, because it includes some score parameters that weight the reliability and functional meaning of the interactions.

METHODS

Agile Protein Interaction DataAnalyzer (APID) design tries to be as simple and light as possible keeping the minimal information to provide a correct and easy access to all included data sets. This design follows the software engineering methodology named 'agile' (13), that embraces software development using lightweight and adaptable methods. In this way, agile methods demand the idea of evolutionary design and seek to assume changes, allowing them to occur along all the live cycle of a product. Changes are controlled and easy to implement and the attitude of the designer is to enable them. APID has been designed following this strategy to achieve the purpose of a useful and active integration of the protein–protein interaction source databases included.

All the work has been developed in Java programming language (http://java.sun.com/), and a J2EE architecture has been used to built the web interface and the applet graphic tool described below. For the parsing of source data we have used SAX and DOM Java programs to extract the information from the XML files, and JDBC programs to insert the processed data in the server. After the parsing efforts we still found problems to unify all the source data, being the main obstacle the heterogeneous and multiple protein identifiers given by the different sources, that many times cause false disjunction and incoherence in the data. To solve it we used the proteins sequences as the most unique and biological meaningful 'protein code', that allowed a good unification using algorithm BLAST2 (14) to find in UniProt each protein given by the source databases. Once a protein was recognized based on sequence alignment, we linked to it a univocal UniProt code. Together with the protein univocal code to obtain a coherent and uniform data, we also had to reach coherence about the experimental method or methods that validate any given interaction. The identification of the method also allows to find the existing consensus or agreement between the different databases for any given interaction. In this way, we have obtained a protocol able to store and unify protein interaction databases in a clear uniform structure, maintaining the integrity of the data and correcting some existing failures found in the original files.

Following the described strategy, the data unification has been done based on three key reference identifiers (IDs): (i) UniProt ID (i.e. UniProt accession number), to allow a specific identification of each protein and a direct link to its sequence and to the rest of the curated protein information included in UniProt (15); (ii) PSI-MI ID, to unify the experimental methods used in different publications to a common terminology developed by PSI-MI (16) (i.e. to a controlled vocabulary with standard identifiers); (iii) PubMed ID (PMID), to link each interaction validated by a given experimental method to a specific PubMed literature reference, and also to assign experimental method identifiers to the PubMed publications that describe each method. These main key identifiers constitute a simple information core that makes APID an agile tool to access and search through the interactomes.

At present, APID integrates data coming from five main source databases: BIND (10), DIP (11), HPRD (Human Protein Reference Database) (17), IntAct (Database system and analysis tools for protein interaction data) (12) and MINT (Molecular Interactions Database) (18). The data included in APID coming from these source databases correspond only to protein–protein interactions (i.e. not interactions of proteins with other ligands like DNA and the like) and the interactions have to be experimentally validated with a PubMed reference given. At the same time, as indicated above each protein has to be identified by its sequence and its UniProt code. In all cases, the web tool includes for each interaction links to the original files of the source databases, and to the PubMed references that validated each interaction. Finally, each protein includes links to the corresponding UniProt file and to other related databases [like InterPro, Pfam, Gene Ontology (GO), Ensembl, NCBI Gene].

PROGRAM DESCRIPTION

Workflow

To illustrate the workflow and the different tools included in APID web server, we present in Figure 1 a schematic description of the steps usually given for a query. Each box included in the figure corresponds to a web window. Starting with box 1 a protein name, protein identifier, protein description or part of it is inserted in the general 'APID search' tool. As an example, CDC28 from yeast ('CDC28_YEAST') is the starting query. In box 2 the figure shows the result given by the search for 'CDC28_YEAST' that is a UniProt entry name. A simple table with only one row is presented because only one protein is found. This table includes six columns with information about the protein: the UniProt entry name, the number of interactions, the UniProt ID number, the taxon (NCBI Taxonomy ID), the protein name or description and a link with more information about the protein. Clicking on the link '+info_prot' a new window with more detailed information about the query protein is displayed, including links to other referred biomolecular databases. The '+info_prot' file also includes some calculated parameters about the protein interaction network (i.e. connectivity and cluster coefficient) and about the protein functional environment based on GO annotation (i.e. GO environment and GO environment enrichment). In this way, it can be seen that connectivity 229 corresponds to the number of proteins that interact with CDC28 from yeast. This is a big number of interactions that may include false positives. Clicking on 229 in box 2 a new window is displayed including a table with details about the 229 interactions that have been reported for CDC28_YEAST. This table (Figure 1, box 3) has five columns with information about: the interaction protein partners, the number of methods that validate each interaction, the provenance source databases (with links to them) and a final column with more information about the interaction: '+info_inter'. Clicking on any '+info_inter' a new window with more detailed information about the corresponding interaction protein pair is displayed, including marks in yellow that show GO terms overlapping and marks in green that show iPfam doamin–domain interaction. This is the case shown for protein pair CDC28_YEAST and SWI6_YEAST.

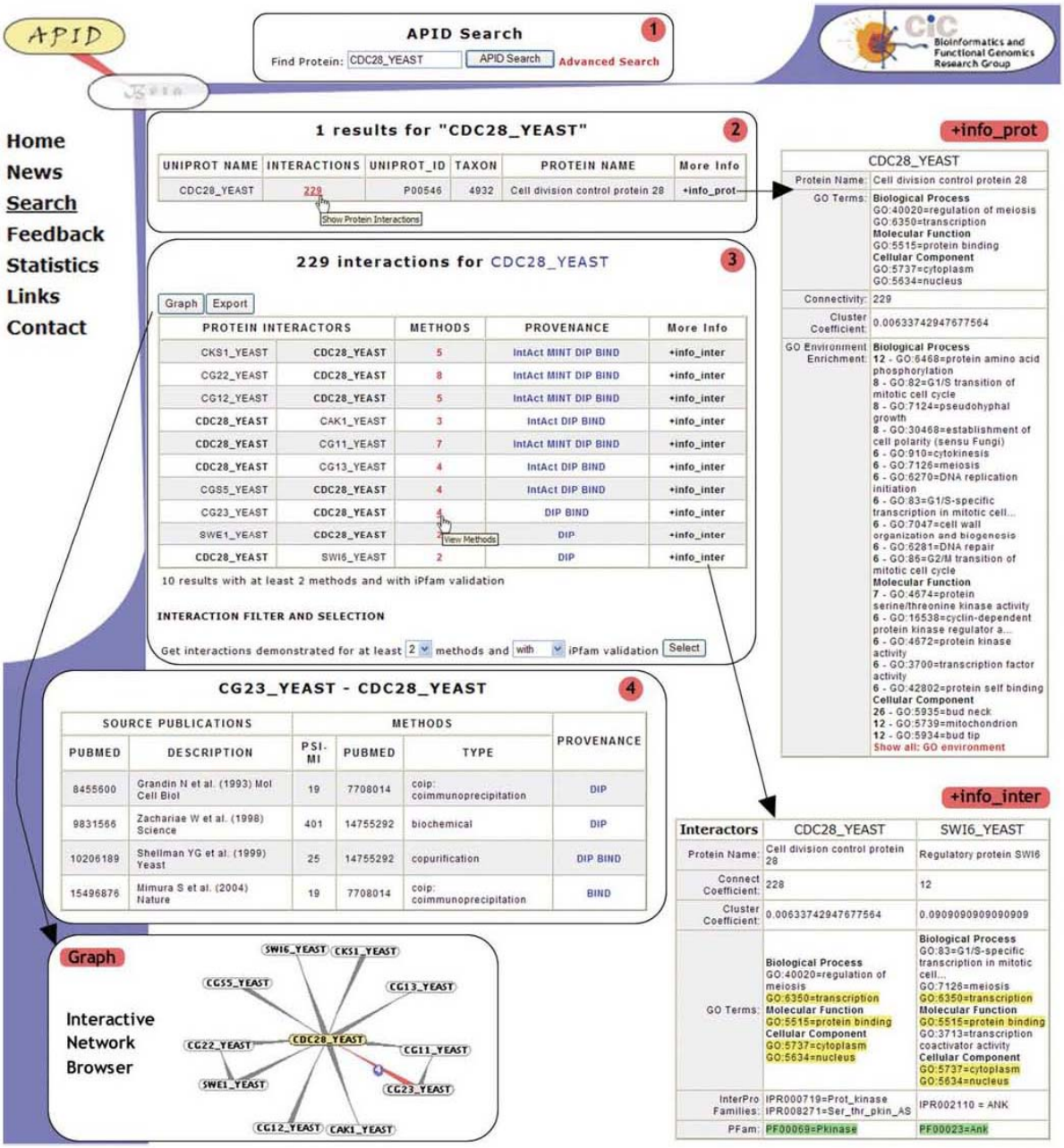


Figure 1. Schematic representation showing APID workflow example. Search query: 'cdc28_yeast' (box 1). Protein found with the text search (box 2) and its additional information (+info_prot). The found protein presents 229 protein partners and the filtered interactions (with iPfam validation and at least two methods) are show in next chart (box 3), that links to the graphical tool APIN where the corresponding interaction network can be visualized and explored in an interactive way. Each interaction also links to its additional information (+info_inter). The experimental methods that prove each interaction are indicated and the details about such methods for the protein pair CG23_YEAST and CDC28_YEAST are shown by clicking the corresponding number 4 (box 4). Each presented box corresponds to consecutive web pages in the APID website.

At the same time, as presented in box 3 a certain subset of protein interactions can be selected from the original 229 interaction using a filter that limitates the display to interaction pairs proven by two methods at least and that also show iPfam domain–domain interaction. Doing this the number of interaction partners for CDC28 is reduced to only 10 proteins (as seen in box 3). Finally, clicking on the number of methods APID displays another window with the information about all

the methods that validate any given protein–protein interaction (e.g. CG23_YEAST – CDC28_YEAST in box 4), presenting for each method: (i) the publications that describe and prove the interaction, linking to PubMed by the pubmed accession number (PMID) and including a description about the publication (i.e. first author, year, journal); (ii) the type of method (i.e. name given by the controlled vocabulary), the PMID of the publication that explains such experimental technique or method, and the PSI-MI method identifier; (iii) the source databases that include these data.

When any subset of protein interaction pairs is selected, as done in box 3, APID also includes a 'Graph' tool that opens a graphical interactive network browser, where the proteins are nodes and the interactions edges. This application tool visualizes dynamically the data, and allows interactive exploring and navigating along the network. The tool includes information about the proteins and the interactions that can be shown by opening windows with basic information and with links to the reference databases UniProt, PubMed and so on.

Statistics and overlap between source databases

At present time (February 2006) the 'statistics' section in APID web tool shows that the application includes >35 000 proteins from several organism and >111 000 interactions. The 'statistics' web page also presents the proteins and interactions per organism, the number of methods that report the same interaction and the detail numbers for the overlap and different intersections between the five source databases used. A more simple and graphical analysis of the overlapping of only three protein interaction databases (BIND, DIP and IntAact) is presented in Figure 2, that shows a Venn diagram with the number of interactions for the multiple intersections between these three databases. It is worthy of note that 62% of the overall protein interactions included in BIND, DIP and IntAct are presented in only one of these databases, i.e. they are exclusive to one of the protein interaction resources.

Protein interaction network assessment

High-throughput experimental technologies used to prove protein–protein interactions of complete proteomes, using the two-hybrid system (19) or mass spectrometry (20), have highly increased the data included in the protein interaction databases. However little overlap between the high-throughput datasets (6) and frequent disagreement with small-scale experiments jeopardize high-throughput interactions confidence. Several efforts have been undertaken to tackle this problem (4,5), but some critical steps to solve it are to achieve more comprehensive and integrated resources of the interactomic data and to include certain calculated parameters that weight the reliability of a given interaction between two proteins. These steps are the ones followed by APID application that is an integrated repository of interactions and includes some tools to assess the interactions:

- *Number of methods*: number of experimentally validated methods that prove a protein–protein interaction, given the PubMed reference and link.

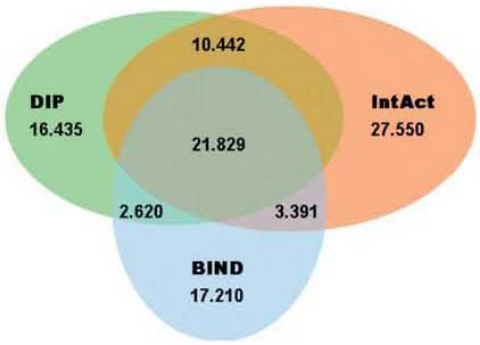


Figure 2. Venn diagram with the number of interactions for the multiple intersections between BIND, DIP and IntAct.

- *GO overlapping*: tool that shows the GO terms assigned to each protein pair and marks the ones that are common to both.
- *iPfam domain–domain interaction*: tool that identifies the Pfam domains of each protein pair and marks the ones that interact according to iPfam database.

At the same time, APID also infers data about proteins in the interaction network, since the web application measures graph parameters as the connectivity and the cluster coefficient of each node, and it also qualifies the functional environment around any given protein using GO:

- *GO environment*: tool that identifies and lists all the Gene Ontology (GO) terms that are assigned to the proteins directly interacting with a query protein.
- *GO environment enrichment*: tool that selects the most-represented and non-self GO terms assigned to the proteins interacting with a query protein.

The use of these quality parameters will allow to make functional predictions about the proteins based on the assumption that interacting proteins tend to have related functions or at least to be involved in common biological processes. Using protein neighbourhoods such biological processes can be explored and mapped on a more reliable interactome landscape.

ACKNOWLEDGEMENTS

We thank Alberto de Luís and Ángel Román for helpful discussions. We acknowledge the funding and support provided by the Spanish Ministerio de Sanidad y Consumo, ISCIII (research grant ref. PI030920), Junta de Castilla y Leon (research grant ref. SA104/03) and Fundación BBVA (Bioinformatics Grants Program). C.P. holds a research grant for PhD from Junta de Castilla y Leon (ref. BOCyL no. 119, EDU/777/2005).

Conflict of interest statement. None declared.

REFERENCES

1. Xenarios,I. and Eisenberg,D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.*, **12**, 334–339.

2. Legrain,P., Wojcik,J. and Gauthier,J.M. (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.*, **17**, 346–352.
3. De Las Rivas,J. and De Luis,A. (2004) Interactome data and databases: different types of protein interaction. *Comp. Funct. Genom.*, **5**, 173–178.
4. Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.
5. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
6. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.
7. Bader,J.S. (2003) Greedily building protein networks with confidence. *Bioinformatics*, **19**, 1869–1874.
8. Iossifov,I., Krauthammer,M., Friedman,C., Hatzivassiloglou,V., Bader,J.S., White,K.P. and Rzhetsky,A. (2004) Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, **20**, 1205–1213.
9. Bader,J.S., Chaudhuri,A., Rothberg,J.M. and Chant,J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, **22**, 78–85.
10. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
11. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
12. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
13. Cockburn,A. (2002) *Agile Software Development*. Addison-Wesly Longman, London, UK.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
16. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
17. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
18. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
19. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
20. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

- **Prieto C**., Rivas-Lopez M. J., Sánchez-Santos J. M., Lopez-Fidalgo J. and De Las Rivas J. (2006). Algorithm to find gene expression profiles of de-regulation and identify families of disease-altered genes. *Bioinformatics 22(9)*: 1103-10.

BIOINFORMATICS **ORIGINAL PAPER** Vol. 22 no. 9 2006, pages 1103–1110
doi:10.1093/bioinformatics/btl053

Gene expression

Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes

C. Prieto[1], M.J. Rivas[2], J.M. Sánchez[2], J. López-Fidalgo[2] and J. De Las Rivas[1,*]

[1]Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CIC USAL-CSIC) and [2]Department of Statistics, Faculty of Science (USAL), Salamanca, Spain

Received on July 21, 2005; revised on December 30, 2005; accepted on February 8, 2006
Advance Access publication February 24, 2006
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Alteration of gene expression often results in up- or down-regulated genes and the most common analysis strategies look for such differentially expressed genes. However, molecular disease mechanisms typically constitute abnormalities in the regulation of genes producing strong alterations in the expression levels. The search for such deregulation states in the genomic expression profiles will help to identify disease-altered genes better.

Results: We have developed an algorithm that searches for the genes which present a significant alteration in the variability of their expression profiles, by comparing an altered state with a control state. The algorithm provides groups of genes and assigns a statistical measure of significance to each group of genes selected. The method also includes a prefilter tool to select genes with a threshold of differential expression that can be set by the user *ad casum*. The method is evaluated using an experimental set of microarrays of human control and cancer samples from patients with acute promyelocytic leukemia.

Availability: The method is implemented in an R package called AlteredExpression available in http://bioinfow.dep.usal.es/Altered-Expression/ and will be included in the Bioconductor project.

Contact: jrivas@usal.es

1 INTRODUCTION

The development of genomic technology has provided the means to interrogate at once thousands of genes in biological samples. In this respect, the advance of high-density oligonucleotide microarrays has become one of the most powerful tools that allow testing the expression state of thousands of genes at a genome-wide scale. At present, a critical bottleneck for this kind of studies is the analysis of the huge amount of data that these microarrays provide (Marshall, 2004). Bioinformatic tools, based on robust computational and statistical studies, are needed to obtain correct and comparable expression profiles that characterize gene families or groups of genes involved in common biological functions.

Gene expression is a tightly regulated process, crucial for the proper functioning of a cell. In microarray data, common regulation is reflected by strong correlations between expression levels (Eisen *et al.*, 1998), and it is usual that genes of similar function yield similar expression profiles across a diverse range of conditions (Segal *et al.*, 2003; Stuart *et al.*, 2003). Molecular disease mechanisms typically constitute abnormalities in the regulation of genes producing strong alterations in the expression levels (Rhodes *et al.*, 2002). The resulting changes in expression profiles can help to identify disease-related genes (Golub *et al.*, 1999), and can also facilitate improved diagnosis and even prognosis of disease outcome (West *et al.*, 2001). Alteration of gene regulation often results in expression profiles with large variability. However, common analysis strategies (Tusher *et al.*, 2001; Efron *et al.*, 2001) look only for differentially expressed genes (overexpressed or represed), but do not explore the variability that appears in an altered state when compared with a control state.

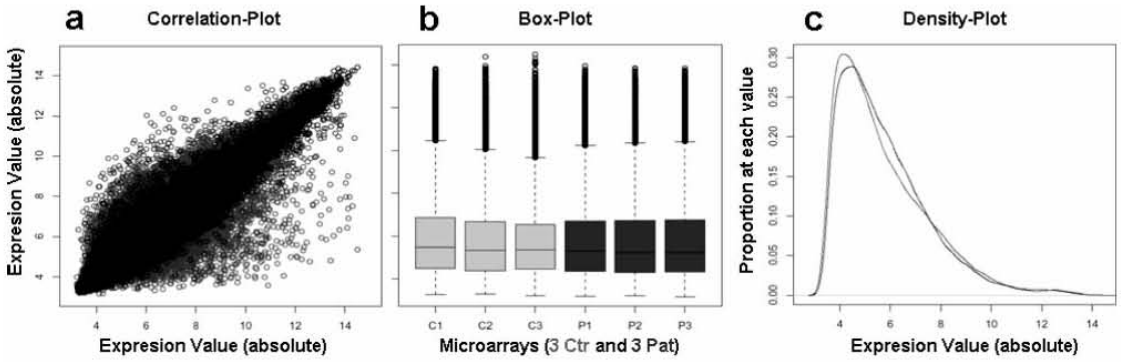
We have developed a new algorithm that is able to analyze and compare the expression profiles of genes from control samples versus disease altered samples and finds those genes that suffer a strong alteration in variability or de-regulation. The algorithm is specially adapted and developed for high-density oligonucleotide microarrays, particularly GeneChips manufactured by Affymetrix. It also includes a method to preselect differentially expressed genes. The method provides a list of well-defined groups of de-regulated genes assigning a statistical score of significance to each group. The method is evaluated in this paper using an experimental set of 16 microarrays: 6 controls and 10 cancer samples.

2 METHODS

2.1 Background correction and normalization

A key step before any analysis of gene expression between different samples is to achieve an adequate background correction and normalization of the expression signals done for all the microarrays that are going to be analyzed. Both background correction and normalization have to be done at internal level (intra-microarrays) and at comparative level (inter-microarrays). Several algorithms have been published to calculate the expression signal from raw data of Affymetrix microarrays. A statistical algorithm designed and recommended by Affymetrix called MAS5 (Liu *et al.*, 2002), a model-based algorithm called MBEI (Li and Wong, 2001) and a multiple average algorithm called RMA (Irizarry *et al.*, 2003a). Comparisons of the normalization methods used by these algorithms show that MAS5 does not achieve an adequate multiple chip normalization and RMA is the method that provides the best precision in signal detection, especially with the difficult genes that present low levels of expression (Irizarry *et al.*, 2003a; Barash *et al.*, 2004). RMA uses an efficient quantile normalization of the distribution of probe intensities for each array in a set of arrays (Bolstad *et al.*, 2003). Following these publications and using Bioconductor and R as computational tools, we use RMA for background correction, normalization and expression calculation of a set of 16 microarrays from clinical samples. The set corresponds

*To whom correspondence should be addressed.



Summary of statistical values for a Control (C3) and a Patient (P3):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.214	4.455	5.348	5.706	6.539	14.500
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.159	4.319	5.234	5.733	6.757	13.860

Fig. 1. Representation of the expression data from several normalized microarrays from control samples (C) and altered samples (P) from patients with APL. (a) Correlation plot of the expression values of three controls versus three patients (grey spots), three controls versus three other controls (black spots), one control against itself (middle line). (b) Box plots showing the data distributions of three controls (C1, C2 and C3) and three patients (P1, P2 and P3). (c) Density plots of the added distributions of three controls (black line) and three patients (grey line). The box below presents a summary of the statistical parameters calculated for the expression data of a control sample (C3, first row) and of a patient sample (P3, second row).

to samples coming from human bone marrow biopsies from 16 different individuals: 6 healthy persons and 10 patients with acute promyelocytic leukemia (APL). This is a multiple and complex set of microarrays that we used in this paper for the evaluation and validation of the algorithm presented. The complexity of the samples can be seen in Figure 1a that plots the correlation of expression values for each gene (i.e. each probeset in the microarrays) in a subset of six samples: three patients and three controls. The comparisons done are nine pairwise comparisons of three patients versus three controls (grey points); nine pairwise comparisons of three controls versus other three controls (black points) and one comparison of a control against itself (middle line). The black points reflect the high degree of biological variability between individuals in these clinical samples, but thanks to a correct normalization it is possible to see that the greater differences correspond to the comparisons of the expression values of patients versus controls. The spots in the grey region will be the ones that best define the differences between disease samples and control samples. Figure 1b and c present the box-plots and the density-plots of the distributions of expression values from the three control samples and three patient samples that were compared in Figure 1a. The numbers correspond to $\log_2$-transformed data. It can be observed in the density-plots that the distributions of expression values derived from Affymetrix microarrays do not follow a normal (i.e. Gaussian) distribution. Another observation is the small differences in the expression distributions between different samples, even between control samples and patient samples. This fact indicates a good normalization and makes it possible to proceed into the next step of the analysis to find de-regulated genes. The normalization shown for the subset of samples in Figure 1 was evenly done at once for all the 16 samples of the study.

2.2 Score function for altered expression and statistical adjustment to F distribution

The interest of the work is to find groups of genes that present a similar expression pattern in control samples but suffer a distinct alteration of their expression profiles in the query samples, i.e. the disease or altered samples. Each experimental set of microarrays, involving I genes from J samples, can be mathematically assigned to a $|I| \times |J|$-dimensional matrix $I \times J = M = (e_{ij})$, where e_{ij} is the intensity of the ith gene for the jth sample. In this way, each gene has an expression profile along the samples that defines its state. For a subset of genes $G \subset I$ and a subset of samples $S \subset J$, the submatrix $G \times S$ represents the intensities of those genes for those samples. Owing to the fact that the aim is to compare only two states within the samples, the standard or control state (i.e. control samples) versus the disease or altered state (i.e. altered samples), the samples set are always divided in two subsets: S_a (altered samples) and S_c (control samples); so, $|J| = |S_a| + |S_c|$.

For each expression value e_{ij}, once normalized and the background corrected, we can consider an additive model that conceptually divides this expression value in a set of four different additive elements: the gene effects (α_i), the sample effects (β_j), a constant value common to all the data in a set (γ) and the errors (ε_{ij}); so that

$$e_{ij} = \alpha_i + \beta_j + \gamma + \varepsilon_{ij}.$$

The usual assumption done in most of the algorithms that handle expression values is that the errors are independent and its variances are equal along genes and samples. Under this model the maximum-likelihood estimators of the additive elements described above will be for global common effects (γ) the global mean of all the expression values for samples and genes ($\bar{e}_{\bullet\bullet}$); for gene effects (α_i) the difference between the means of the expression values for different samples in a gene i ($\bar{e}_{i\bullet}$) and the global mean ($\bar{e}_{\bullet\bullet}$) and for sample effects (β_j) the difference between the means of the expression values for different genes in a sample j ($\bar{e}_{\bullet j}$) and the global mean ($\bar{e}_{\bullet\bullet}$). So the maximum-likelihood estimators for the errors $\varepsilon_{ij} = e_{ij} \quad \alpha \quad \beta_j \quad \gamma$ will be

$$e_{ij} \quad (\bar{e}_{i\bullet} \quad \bar{e}_{\bullet\bullet}) \quad (\bar{e}_{\bullet j} \quad \bar{e}_{\bullet\bullet}) \quad \bar{e}_{\bullet\bullet} = e_{ij} \quad \bar{e}_{i\bullet} \quad \bar{e}_{\bullet j} + \bar{e}_{\bullet\bullet}$$

Using these estimators, the aim is to compare all the expression values e_{ij} of the matrix M to find the variability and correlation between genes and samples. To find such variability of expression values, a statistical parameter of variability should be calculated. An obvious parameter is the standard correlation coefficient but several authors have shown that the residual variance (RV) allows a more efficient search for high-scoring gene sets (Cheng and Church, 2000; Kostka and Spang, 2004). As these authors,

we choose the **RV**(***G***, ***S***) as the parameter to measure the variability within a certain subset of genes ***G*** along a certain subset of samples ***S***. This parameter is defined as

$$\mathbf{RV}(G,S) = \frac{1}{(|G|-1)(|S|-1)} \sum_{G,S} (e_{ij} - \bar{e}_{i\bullet} - \bar{e}_{\bullet j} + \bar{e}_{\bullet\bullet})^2$$

From the linear model, the former summatory follows a Pearson's χ^2 distribution with $(|G|-1)$ and $(|S|-1)$ degrees of freedom (Montgomery, 2000; Draghici, 2003). In this way, the **RV** is the ratio between a Pearson's χ^2 distribution and its degrees of freedom, and it is calculated for the subset of genes ***G*** and samples ***S***. Next step in the algorithm is to implement the ability to compare **RV** for the two different and independent subsets of samples, altered and control (S_a and S_c), in a given subset of genes ***G***. For both subsets we calculate each $\mathbf{RV_c} = \mathbf{RV}(G, S_c)$ and $\mathbf{RV_a} = \mathbf{RV}(G, S_a)$, and the comparison can be done by the relative residual variance (**RRV**) that is the quotient or ratio between the $\mathbf{RV_c}$ of the controls and the $\mathbf{RV_a}$ of the altered samples:

$$\mathbf{RRV}(G,S) = \frac{(|G|-1)(|S_a|-1)}{(|G|-1)(|S_c|-1)} \frac{\sum_{G,S} (e_{ijc} - \bar{e}_{i\bullet c} - \bar{e}_{\bullet jc} + \bar{e}_{\bullet\bullet c})^2}{\sum_{G,S_a} (e_{ija} - \bar{e}_{i\bullet a} - \bar{e}_{\bullet ja} + \bar{e}_{\bullet\bullet a})^2}$$

The division of these two Pearson's χ^2 distributions follows a Fisher–Snedecor's distribution ($F_{n1,n2}$) with $(|G|-1)\cdot(|S_c|-1)$ and $(|G|-1)\cdot(|S_a|-1)$ degrees of freedom (i.e. ***n1*** and ***n2*** respectively) (Montgomery, 2000). The adjustment to a probability distribution allows constructing a statistical hypotheses test for each subset of genes ***G***, where the null hypothesis H_0 will be no difference in the parameter of variability (**RV**) of expression between the controls ($\mathbf{RV_c}$) and the altered samples ($\mathbf{RV_a}$) for a given subset of genes ***G***. In statistical terms, no such difference corresponds to the situation where the $\mathbf{RV_a}$ of altered samples is less than or equal to the $\mathbf{RV_c}$ of the controls. On the other side, the alternative hypothesis H_1 will be the existence of a real significant difference between the controls and the altered samples. **RV** is taken as the estimator of variance:

H_0: $\sigma_a^2 \leq \sigma_c^2$, i.e. the subset of genes ***G*** of altered versus control samples does not present difference in variability

H_1: $\sigma_a^2 > \sigma_c^2$, i.e. the subset of genes ***G*** of altered versus control samples does present difference in variability

This is a one-tail test where the rejection region is $\{F \leq k\}$ for a critical value k. In this way, a probability score can be calculated corresponding to the value of the ***F*** statistic, and such score is taken as an indicator of the difference in variability between the control and the altered samples of a subset of genes ***G***: F-score $= P\{F \leq F_{obs}\}$. The score is calculated for each subset of genes ***G***, and the program runs by iteration until achieving stabilization of such score for a group of genes. This allows a progressive selection of groups with increasing score from a minimal initial value that will correspond to the most distant from the null hypothesis.

2.3 Algorithm to search for de-regulation in gene expression: AlteredExpression

Following the score function defined above and the statistical parameters derived, we wrote in R the algorithm called AlteredExpression. A full scheme to show how it works is presented as a detailed flowchart in Figure 2. The algorithm is designed to run using as input a microarray expression dataset with ***I*** genes and ***J*** samples subdivided in two defined types: control samples S_c and altered samples S_a. The output provides each time an optimal group of genes identified with best **RRV** and minimal F-score. Once a first group of genes G_1 is selected, they are taken out from the whole gene dataset ($I - G_1$) and the algorithm runs all again selecting a second group (G_2). The process continues until the **RRV** of the groups generated do not change significantly, i.e. they arrive to a stable stage. The algorithm includes some variable parameters that are fixed in advance as input parameters and they can be adjusted depending of the type of samples analyzed. These parameters and their meaning are as follows:

(1) initialSize: indicates the number of genes included in the initial random sampling of the dataset. We explore several sizes and an initial value of 20 genes behaves well for most of the Affymetrix datasets or subsets (for a total number of genes between 2.000 and 20.000). The trials done with different initialSizes show that this parameter does not affect the resulting output groups.

(2) maxiter: indicates the maximum number of iterations that are allowed for the algorithm to work exploring the whole data matrix and find an optimal group. This parameter is just for security to avoid running without stopping, and it is fixed to a high number of 500 iterations to avoid stopping before the correct selection of a stable group.

(3) vSize: is a parameter introduced to modulate the effect of the group size. It can vary between 0 and 1, being close to 0 for higher size effect and close to 1 for lower size effect. The vSize is fixed to 0.5 by default.

(4) pctSubset: is a parameter, varying between 0 and 1, introduced to fix the percentage of in–out genes (ioGenes) that are randomly selected and changed in the genes set (***G***) every time the algorithm iterates in the external loop **1**. The ioGenes is the whole set of genes that individually have improved the **RRV** each time the algorithm iterates in the internal loop **2**. The pctSubset is fixed to 0.1 by default, meaning that only 10% of the ioGenes are changed in ***G*** each loop **1** run (Fig. 2).

2.4 Pre-selection of genes with differential expression

The main idea behind the method is to select the groups of genes that show a de-regulation, but also including the possibility of pre-selecting genes with a certain threshold of differential expression between control and altered state. The algorithm and score function described above are designed to look for a maximum variability in the altered state (altered samples, S_a) with a small variability in the control state (control samples, S_c). The algorithm can explore complete crude datasets from microarray experiments, but, in order to get a more correct and meaningful output it is better to avoid the mean expression of a certain gene to be the same between control and altered samples. To achieve this, the input data can be previously treated with an algorithm to look for differentially expressed genes. One of the most solid and useful algorithms for this purpose is Significant Analysis of Microarrays (SAM) (Tusher *et al.*, 2001) that uses permutations of the samples to estimate the percentage of genes identified by chance, and it also calculates a false discovery rate (FDR) which corresponds to a Type I error calculation (i.e. number of false positives; Benjamini and Hochberg, 1995).

A more simple way to start with a minimal differential expression between control and altered samples is to de-selected or delete from the crude input data the genes that do not present a minimal change in mean expression value between control and altered samples. This is a means pre-filter (ΔMF) that is included at the beginning of the algorithm and its threshold of expression difference can be fixed before the running according to the characteristics of the sample. As shown below in the results, we evaluate the use of SAM and/or the means pre-filter within the algorithm.

3 RESULTS AND DISCUSSION

3.1 Algorithm tuning with clinical experimental data.

For a correct development of the algorithm AlteredExpression, an evaluation of the performance was done using different sets of microarray data from clinical samples. The main data used were the set mentioned above corresponding to human samples from 16 different individuals: 6 healthy persons (controls) and 10 patients with APL (altered). This set of data has an important biological variability, as it is usually the case for clinical samples coming

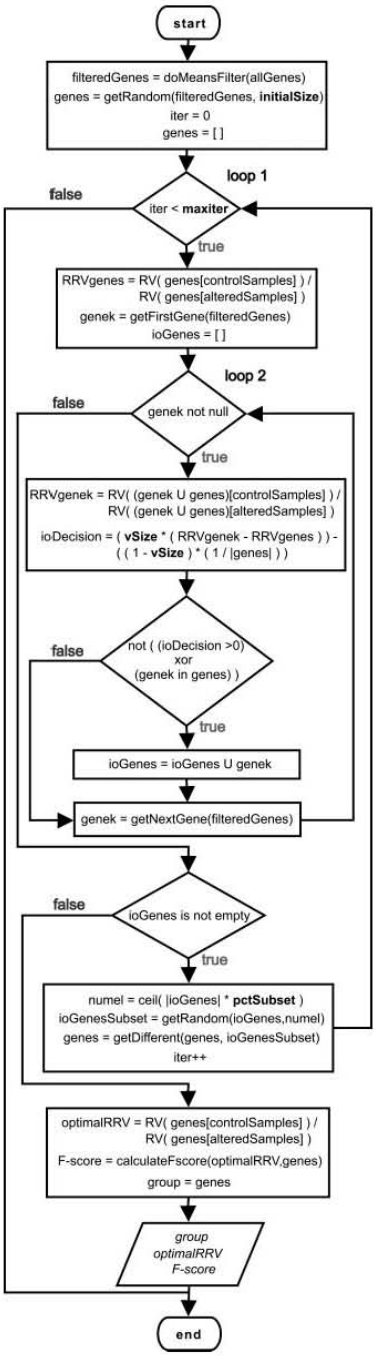


Fig. 2. Flowchart of the algorithm AlteredExpression showing the way it works to select a group of genes from the initial data matrix. Four parameters described in Methods are initialSize, maxiter, vSize and pctSubset. The output gives a stable group of genes, together with the RRV calculated for such group (optimalRRV) and its *F*-score.

from different persons. We observed that adequate inter-microarray normalization was needed before any further data analysis, because without such normalization the performance of the algorithms was critically diminished. For these reasons, in all our analyses the microarray expression datasets have been processed using RMA to calculate the signal (Irizarry *et al.*, 2003).

Another observation learned in the trials, done to tune the algorithm efficiency, was that the search for altered genes performed by the algorithm was improved when the data introduced as input were previously filtered by differential expression. Figure 3 shows these effects presenting the variation of gene content of the first group given by the algorithm. The raw data correspond to an input of the complete set of expression values for all genes in the microarrays without any filter (22.283 probesets) (Fig. 3a and d). The filtered data correspond to the ones obtained using ΔMF (Fig. 3g) or using SAM with or without ΔMF (Fig. 3b, c ,e, f, h and i). As seen in Figure 3 scaled graphs, using a ΔMF of expression value 1.0 (that corresponds to 8–9% of the expression range) a big number of genes, that do not show means change between control and altered samples (called 'flat' genes), are cleared (Fig. 3g, h and i). The use of both ΔMF and SAM gives the algorithm the best performance in the search for altered genes. This is a stringent strategy that pushes the method to find genes that clearly fulfil the criteria of both differential expression and alteration, in a way to avoid *ad maximum* false positives. The value of using ΔMF together with SAM is shown in Figure 3, where the graphs illustrate that only SAM, even when using a quite stringent FDR of 0.001, is not enough to filter out all 'flat' genes (Fig. 3e and f). These genes are all eliminated adding a ΔMF of value 1.0 (Fig. 3h and i). To better demonstrate that ΔMF, implemented in our method, improves the simple use of SAM, Figure 4 shows that using SAM with FDR 0.01 still lets some genes that are 'flat' genes and do not fulfil the criteria of a minimal difference in expression means that we are looking for. Such genes are taken by SAM because they have a very small standard deviation in the control samples but larger deviation in the altered samples, therefore they can be taken as significant in statistical tests involving the mean and the standard deviation as it is SAM (that is a modified *t*-test). In contrast, the means-filter ΔMF is most efficient to filter-out such genes and this allows obtaining more consistent results in the performance of the AlteredExpression algorithm.

To give an orientation for the use of ΔMF we have done a numerical correlation between the relative number of genes that are filtered at a certain value of ΔMF and at a certain value of SAM FDR. The obtained correlation (data not shown) indicates that for the dataset used in this work a $\Delta MF = 0.8 - 1.0$ was equivalent to a $FDR = 0.01$ (i.e. a 99% of statistical specificity).

The genes pre-filtered with ΔMF are in many cases low expression genes, but not always because the criteria used is a 'difference in expression means' independently of the absolute expression value of a gene. This point is quite important because other selective algorithms, like MAS5, apply a simple cutoff of low expression values clearing or ignoring all the genes that are below a certain signal value. A filter based on a simple cutoff follows the criteria that the noise and error accumulates at low expressions and that a major part of the genes in the transcriptome do not change in expression in a certain biological state. However, this approach is frequently inadequate for microarray data, because many critical genes involved in cellular regulation have

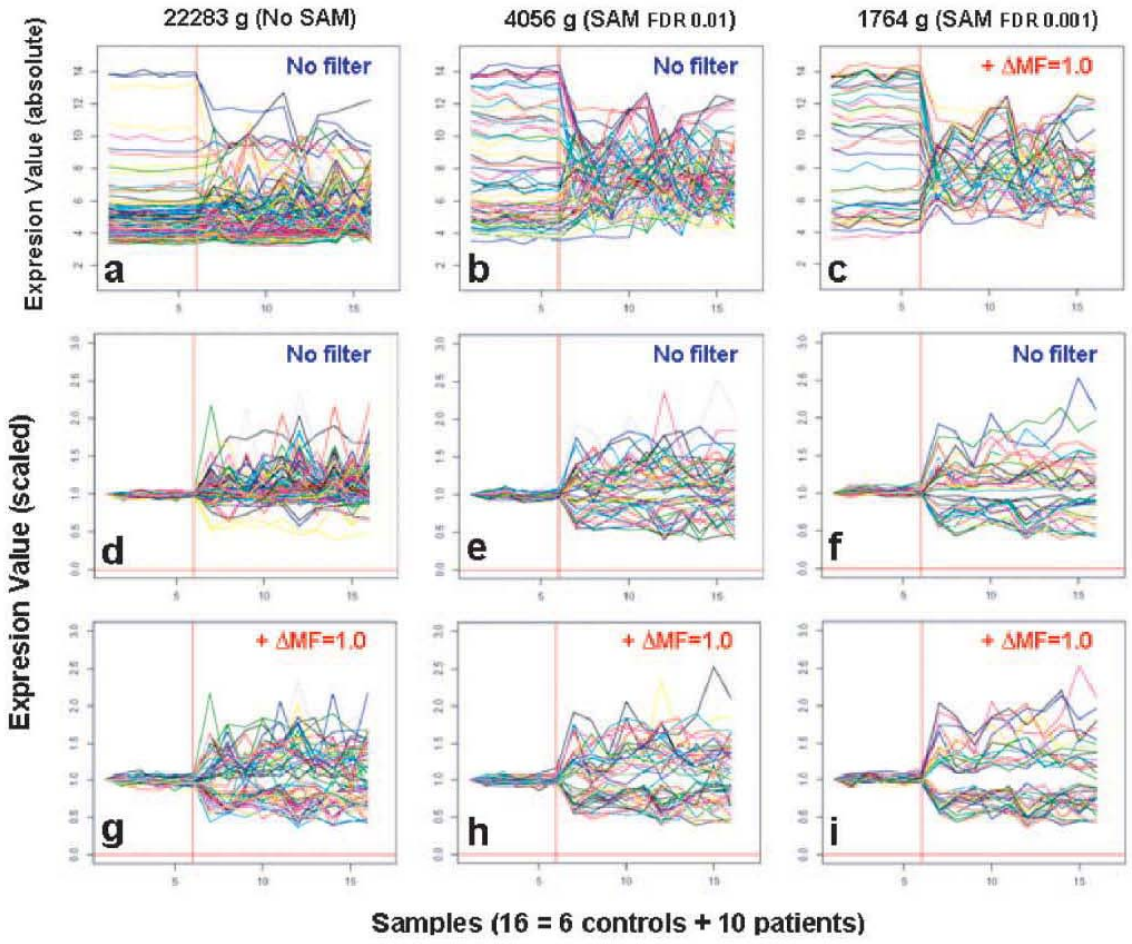


Fig. 3. Graphs presenting the expression values of genes in 16 different samples: 6 control and 10 altered samples (i.e. from patients with APL) separated by a red line. The genes presented are the ones included in the first groups and each graph corresponds to different runs of the algorithm using different sets of initial genes: all 22.283 genes of the microarrays (**a,d,g**); 4056 genes obtained using SAM at FDR 0.01 (**b,e,h**); 1764 genes obtained using SAM at FDR 0.001 (**c,f,i**). The use of a means-filter of 1.0 is also indicated in the graphs. Each gene is represented as a color line that includes the 16 values of expression in absolute numbers (a,b,c) or 16 values of expression scaled by making value 1.0 for the expression of the first sample and the rest relative to it (d,e,f,g,h,i).

low expression values (e.g. genes that are at the beginning of signal transduction cascades), and also because these type of genes many times show small fold changes when they are over or underexpressed.

Behaviour of the *F*-score in consecutive groups: finding the most significant groups of genes

As described, AlteredExpression algorithm explores the data matrix of expression values searching for groups of genes that have minimal variation in the control samples but show clear variability in the disease samples. In this way, the algorithm produces a series of consecutive stable groups starting from a minimal *F*-score for the first group, followed by increasing scores for the next groups found. In all cases, before each run of the algorithm, a group of 20 genes was randomly selected (initialSize = 20). The usual size of the final groups found was about 30–50 genes using the default values of the parameters described above. The behaviour of the *F*-scores for the groups selected after running the algorithm is shown in Figure 5, that represents the scores in log10 scale for each group (Fig. 5a), and the adjusted probability scores, using the Bonferroni method (Bonferroni, 1937; Hsu, 1996) for each group (Fig. 5b). These graphs indicate that the scores present an asymptotic tendency and the initial groups include most of the change. A 75% of the total change in score is achieved between the 1st group and the 10th, the change of first three groups being remarkable. The groups in Figure 5a and b are the ones selected starting with the total set of 1764 genes (obtained with SAM FDR 0.001 and ΔMF 1.0). Using the set of 4056 genes (obtained with SAM FDR 0.01 and ΔMF 1.0)

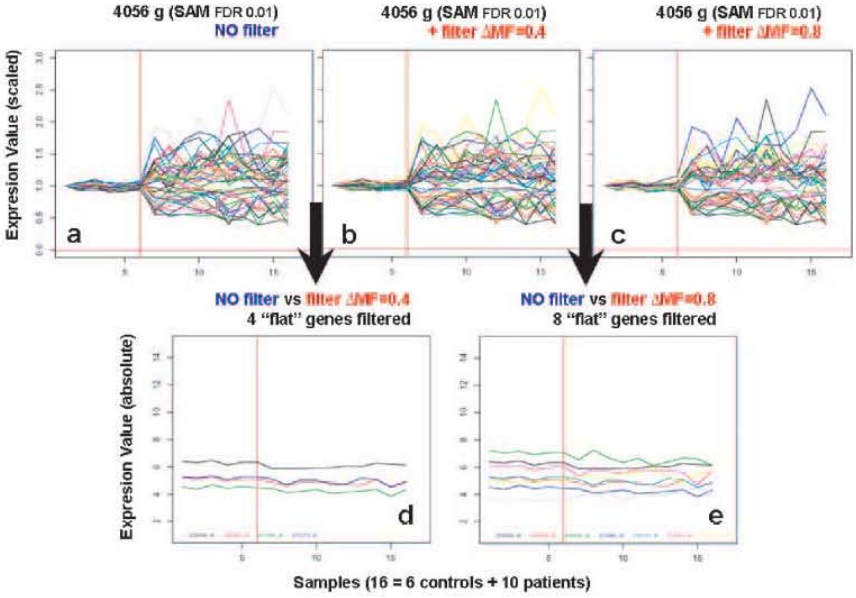


Fig. 4. Graphs presenting the expression values of genes as Figure 3. The genes are the ones included in the first groups using different conditions as initial set: (a) 4056 genes (SAM at FDR 0.01) without ΔMF; (b) with ΔMF = 0.4; (c) with ΔMF = 0.8. Graphs (d) and (e) present the 'flat' genes that are filtered out, from (a) and (b) graphs respectively, when using the ΔMF.

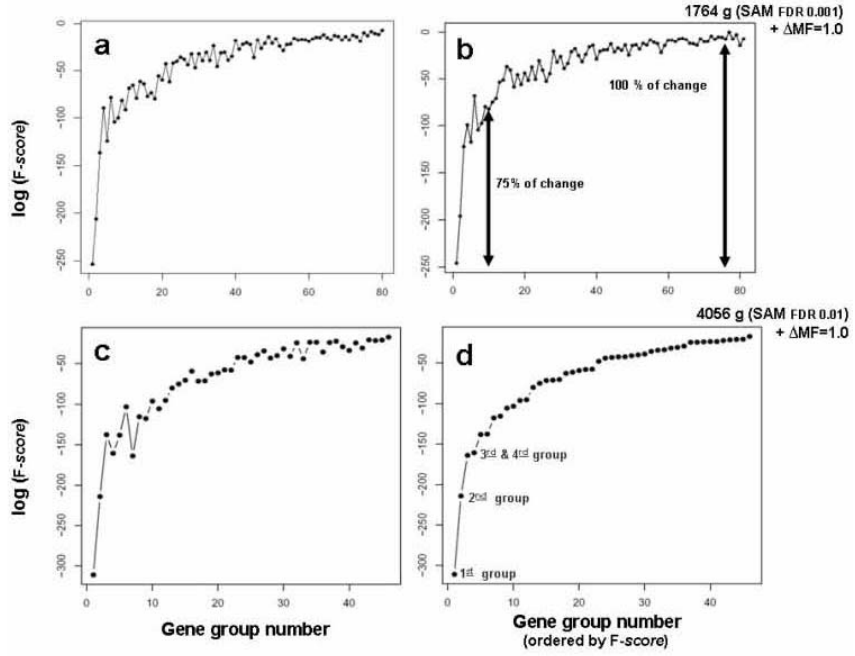


Fig. 5. Graphs presenting the change in *F*-scores (log10 scale) given for each group of genes generated. Groups obtained using initial set of (a) 1764 genes (SAM at FDR 0.001) and ΔMF = 1.0; or using (c) 4056 genes (SAM at FDR 0.01) and ΔMF = 1.0 . Graph (b) presents the same data as (a) but with adjusted *F*-scores by Bonferroni method. Graph (d) presents the same data as (c) but ordering the groups by *F*-scores instead of ordering by appearance.

the behaviour of the scores is similar (Fig. 5c and d) and again the initial groups include the most significant changes. The general tendency to increase the scores does not always keep the order of the most stable groups found (Fig. 5a, b and c). To avoid such fluctuations the groups are ordered by increasing the F-scores in Figure 5d, that shows better continuous growth of this statistical parameter and allows an adequate selection of the most significant groups of genes.

3.3 Stability of the algorithm in repetitive runs, biological meaning of the groups and comparison with other methods

A critical point in any algorithm that selects groups of genes is how reproducible it is, or in other words, which is the consistency, coherence and stability of the groups selected with maximal significance. To check this we use the initial set of 4056 genes (obtained with SAM FDR 0.01 and ΔMF 1.0) and the algorithm was run 10 times completely, obtaining 10 different sets of gene groups. Then, all the genes included in each group of each run were compared with the list of genes of all the other groups in other runs. The data obtained are represented in Figure 6 as percentage of identity of each group along 10 runs. The graph indicates that from first to fifth group the reproducibility of the runs is very high because the groups show an identity >85%. The consistency is high, ~80%, for Groups 6–9, and it starts to be lower after Group 10. This result also indicates that there is a correlation between the high coherence of the initial groups and the low F-scores that these groups present, observed in Figure 5.

The strong consistency and the high F-scores of the initial groups selected by the algorithm are two good indicators that the genes included on these groups keep some kind of biological relationships that make them to cluster. This suggestion agrees with the basic hypothesis that an altered biological state, in cells or tissue from a disease sample, will include groups of genes with altered expression. The algorithm proposed is able to find such groups of altered expression with respect to the control samples. A further question is if such groups of altered expression include genes with similar or related biological functions.

A final analysis was done to check the biological consistency of each group by assignment of the genes included in the group to functional terms given by the gene ontology (GO) annotation (Gene Ontology Consortium, 2005). The result of this assignment for the example used (set of 4056 genes) is shown in Figure 7. The data indicate that the functional annotation of the groups obtained is quite different; for example the first group includes some genes related to functions that are not found at all in the other groups (Groups 2–5): blood vessel development, endothelial cell differentiation, myoblast differentiation, etc. This result shows consistency and functional coherence for the groups obtained with the samples used and such consistency was only dependent on the variability observed between the set of altered samples and the control samples. In any case, the detail about the way the algorithm performs showed in this paper is a clear guide to make a good use of it, knowing that for any given study on some specific samples some tuning runs will be needed to explore the characteristics of such samples.

A final point is to consider how similar is the algorithm presented to others previously published and publicly available. As far as

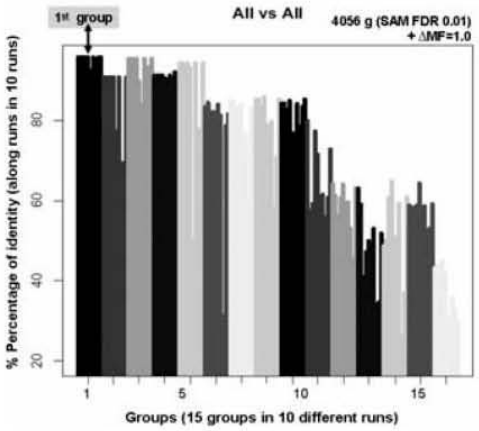


Fig. 6. Comparison of the list of genes present in each group obtained in 10 different complete runs of the algorithm. Each group of genes is compared in percentage of identity with all the other groups obtained in the other nine runs. In this way, the 10 groups most similar are selected making 15 groups ordered by percentage of identity. Each colour bar includes 10 groups one from each run.

we could find out there are not many algorithms that address the issue of statistically significant variability in gene expression in the comparison of two sets of samples. As we said above the core idea of the algorithm presented is to use a residual variance and this idea was previously proposed (Cheng and Church, 2000; Kostka and Spang, 2004). In the case of Kostka and Spang (2004) they wrote an algorithm (dcoex) that tries to find groups of genes with differences in the covariance structure. AlteredExpression follows a similar strategy but also incorporates information on differential expression to find disease-altered groups of genes.

There are in the literature many algorithms that search for groups of genes that show significant co-expression and co-regulation on microarray data, being probably k-means (Hartigan, 1975) and SOM (Kohonen, 1995) the most widely used to generate clusters of genes in a divisive way. These algorithms are unsupervised and they are not designed for class comparison, however such is a key point of our study where control and altered samples are well defined. Other methods perform supervised clustering of genes including the information of sample classes (Dettling and Buhlmann, 2002). This kind of methods usually use as grouping process the partial least squares procedure, to build a covariance matrix that tries to be maximized (Hartigan, 1975). These algorithms are again mostly designed to find co-regulated genes and they do not perform class comparison. The algorithms most used for class comparison are the ones that search for significant differential expression, like SAM algorithm (Tusher *et al*., 2001) or EBA algorithm (Efron *et al*., 2001). These methods do not explore the variability in gene expression that occurs between a control and an altered set of samples, because their objective is to find significant expression mean differences. For these reasons, the proposed algorithm AlteredExpression can be a good complement for a more comprehensive search of the changes that occur between two sets of samples, control versus altered ones.

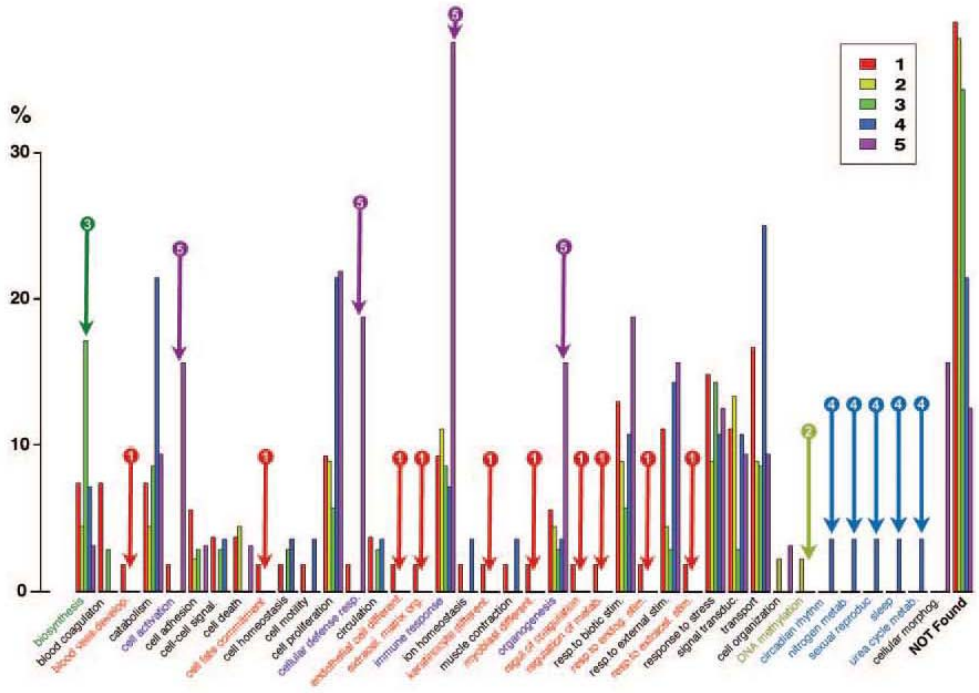


Fig. 7. Assignment of the genes included in the first to fifth groups to functional terms given by the GO annotation. The assignment is done to GO category biological_process at level >3. The scale represents the percentage of total genes that are assigned to each GO term knowing that each gene can be assigned to more than one term.

ACKNOWLEDGEMENTS

The authors acknowledge the funding and support provided by the Spanish *Ministerio de Sanidad y Consumo, ISCIII* (research grant ref. PI030920), *Junta de Castilla y Leon* (research grant ref. SA104/03) and *Fundación BBVA* (Bioinformatics Grants Program 2003).

Conflict of Interest: none declared.

REFERENCES

Barash,Y. *et al.* (2004) Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, **20**, 839–846.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bonferroni,C.E. (1937) Teoria statistica delle classi e calcolo delle probabilita. In *Volume in Onore di Ricarrdo dalla Volta*. Universita di Firenza, Italy, pp. 1–62.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Dettling,M. and Buhlmann,P. (2002) Supervised clustering of genes. *Genome Biol.*, **3**, RESEARCH0069.

Draghici,S. (2003) *Data Analysis Tools for DNA Microarrays*. Chapman and Hall/CRC, London, UK.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Hartigan,J.A. (1975) *Clustering Algorithms*. Wiley, New York.

Hsu,J.C. (1996) *Multiple Comparisons Theory and Methods*. Chapman and Hall, London, UK.

Irizarry,R.A. *et al.* (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A. *et al.* (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kohonen,T. (1995) *Self-Organizing Maps*. Springer, Berlin.

Kostka,D. and Spang,R. (2004) Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, **20** (Suppl. 1), I194–I199.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Liu,W.M. *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.

Marshall,E. (2004) Getting the noise out of gene arrays. *Science*, **306**, 630–631.

Montgomery,D.C. (2000) *Design and Analysis of Experiments*. 5th edn. John Wiley and Sons Inc., New York, NY.

Rhodes,D.R. *et al.* (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.

Segal,E. *et al.* (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19** (Suppl. 1), i273–i282.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Tusher,V.G. *et al.* Significance analysis of microarrays applied to the ionizing radiation response. [Erratum (2001), *Proc. Natl Acad. Sci. USA*, 98 10515.]. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

West,M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

APÉNDICE 2: POSTERS

- Congreso: 7th European Conference On Computational Biology (ECCB2008)
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Comparison and assessment of structural domain-domain intercations to improve protein-protein interaction data**"
- Congreso: Targeting and Tinkering with Interaction Networks
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Functional exploration of a curated human gene coexpression network and relations with human and protein interactome network**"
- Congreso: 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2007)
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Functional coupling of human gene coexpression and protein interaction networks**"

- Congreso: 5th European Conference On Computational Biology (ECCB2006)
  Autores: Prieto C. y De Las Rivas J.
  Título: "**APID, an Integrated Web Platform to Explore and Evaluate Protein-Protein Interaction Networks**"

- Congreso: VI Jornadas Nacionales de Bioinformática
  Autores: Prieto C., Rosón B. y De Las Rivas J.
  Título: "**Defined groups of deregulated genes found in cancer samples by AlteredExpression algorithm**"

- Congreso: : Interactome Networks
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Integration of protein protein interaction data in a unified platform with quality assessment parameters**"

- Congreso: 4th European Conference On Computacional Biology
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Interactome unified database for protein network mining**"

- Congreso: V Jornadas de Bioinformática jbi2004
  Autores: Prieto C., De Luis A. y De Las Rivas J.
  Título: "**Integration and comparison of interactome databases**"

- Congreso: 7th European Conference On Computational Biology (ECCB2008)
  Autores: Prieto C. y De Las Rivas J.
  Título: "**Comparison and assessment of structural domain-domain intercations to improve protein-protein interaction data**"

- Congreso: 7th European Conference On Computational Biology (ECCB2008)
  Fecha: 22-26 Septiembre de 2008 Lugar: Cagliari (Italia)
  Autores: Prieto C. y De Las Rivas J.
  Título: " Comparison and assessment of structural domain-domain intercations to improve protein-protein interaction data"

Comparison and assessment of structural domain-domain interactions to improve protein-protein interaction data

Prieto C. & De Las Rivas J.
e-mail: cprietos@usal.es

Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC-IBMCC, USAL/CSIC)
Salamanca, Spain

ABSTRACT

The assessment of the reliability of protein interaction data is a critical issue to work in a correct way with interactomes. To address this problem the integration of structural domain domain interactions could help to improve the reliability of protein interactomes. With this goal we have compared seven resources that define domain-domain interactions (ddi) to probe if they could validate protein-protein interactions (ppi) and if they are complementary.

Domain-Domain Interaction Datasets

The Database of 3D interacting Domains: 3did

"3did computed physical interactions by requiring at least five contacts (hydrogen bonds, electrostatic or van de Waals interactions) between the two domains, and removed those that lack a significant interface."

Conserved binding mode analysis: CBM

"Two domains qualify an interacting domain pair to be interacting if there are at least five residue–residue contact pairs made between their residues. Residue contacts are counted between residues of one interacting domain and any other residue of another interacting domain whose Cα–Cα distances are within 8Å. To define the conserved binding modes we first collect all structure queries that correspond to the same interacting domain pair. Then we apply the Vector Alignment Search Tool (VAST) to obtain the structure-structure alignments between the queries. We then cluster all interacting domain pairs based on their interface similarity using single linkage clustering. At the end, each cluster corresponds to a binding mode, and clusters with more than one nonredundant query are defined as conserved binding modes."

Domain pair exclusion analysis: DPEA

"Expectation maximization (EM) is a numerical method for obtaining a maximum likelihood estimate of some parameters of a model given incomplete data. We extend the use of EM for estimating probabilities of each kind of potential domain interaction as a starting point for our analysis of the change in likelihood of a set of observed protein interactions, when a potential underlying domain interaction is excluded from the model."

PFAM Domain Interaction: iPFAM

"To identify the interactions between residues, we calculate all bonds (van der Waals, side chain and main chain H-bonds, salt bridge and disulphide) between all pairs of residues in different domains. These bonds can be between domains within the same chain or between domains in different chains in the structure."

Database of structurally defined protein interfaces: PIBASE

"The list of binary interfaces was generated by a three-step procedure. (1) Interatomic distances were calculated for all structures using a user specified distance cutoff. A cutoff of 6.05Å was chosen unless specified otherwise, to allow contacts made via water molecules. (2) The interatomic distances were then combined with the domain definitions to create a list of all domain pairs that share at least one interatomic distance below the specified distance threshold. This list of interacting domain pairs serves as the core of PIBASE. (3) Buried solvent accessible surface area was also computed for each interacting domain pair and a minimum cutoff on the burial was imposed to yield the list of interfaces. Unless specified otherwise, acutoff of 300Å^2 was used."

Database of protein structural interactome map: PsiBASE

"The basic mechanism to check interactions between any two domains or proteins is the calculation of the Euclidean distance in order to see if they are within a certain distance threshold. PSIMAP checks every possible pair of structural domains in a protein to see if there are at least five residue contacts within a 5Å distance (5–5 rule)."

Interfaces and Alignments for PPI: SNAPPI

"Interactions between domains are determined based on distance. Atoms are considered to interact if the distance between them was less than the sum of their van der Waals radii +0.5 Å. Two domains are considered to be interacting if there are >10 interacting residue pairs between the domains"

Heterogeneity of DDI Datasets

	3did	CBM	DPEA	iPFAM	PIBASE	PSIbase	SNAPPI
3did	**2.385**	562	56	1.559	1.039	1.047	974
CBM	562	**5.335**	25	553	908	794	460
DPEA	56	25	**2.767**	48	70	66	42
iPFAM	1.559	553	48	**13.041**	1.063	1.055	1.078
PIBASE	1.039	908	70	1.063	**9.742**	6.471	789
PSIbase	1.047	794	66	1.055	6.471	**6.978**	820
SNAPPI	974	460	42	1.078	789	820	**1.134**

REFERENCES

- Stein, A., Russell, R.B., and Aloy, P. 2005. 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 33: D413-417.
- Shoemaker, B.A., Panchenko, A.R., and Bryant, S.H. 2006. Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci 15: 352-361.
- Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. 2005. Inferring protein domain interactions from databases of interacting proteins. Genome Biol 6: R89.
- Finn, R.D., Marshall, M., and Bateman, A. 2005. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21: 410-412.
- Davis, F.P. and Sali, A. 2005. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 21: 1901-1907.
- Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H., Cho, Y., Han, K., Lee, S. et al. 2005. PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics 21: 2541-2543.
- Jefferson, E.R., Walsh, T.P., Roberts, T.J., and Barton, G.J. 2007. SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 35: D580-589.

METHODS & TOOLS

http://bioinfow.dep.usal.es/apid

APID (Agile Protein Interaction DataAnalyzer)

APID is an interactive bioinformatic web-tool that has been developed to allow exploration and analysis of main currently known information about protein-protein interactions integrated and unified in a common and comparative platform.
The analytical and integrative effort done in APID provides an open access frame where all known experimentally validated protein-protein interactions (BIND, BioGRID, DIP, HPRD, IntAct and MINT) are unified in a unique web application.

Reference Sets of PPI

Agile Protein Interaction DataAnalizer (APID)

All the protein interactions stored in APID.

Statistics: - **241.204** Protein-Protein Interactions.
- **51.873** Proteins.

Exhaustive Experimental Validation (Methods)

APID subset which have been checked with 2 or more experiments.

Statistics: - **37.918** Protein-Protein Interactions.
- **15.266** Proteins.

Same Biological Pathway (KEGGs)

APID subset where the interacting proteins are annotated to the same KEGGs identifier.

Statistics: - **12.315** Protein-Protein Interactions.
- **2.290** Proteins.

Similar Cellular Location (Location)

APID subset where the interacting proteins are present in only one of these locations: Nucleus, Cytoplasm or Mitochondrion; and have the same cellular location.

Statistics: - **6.014** Protein-Protein Interactions.
- **3.220** Proteins.

Coverage Analysis of DDI

		3did	CBM	DPEA	iPFAM	PIBASE	PSIbase	SNAPPI
APID	True Positives	7.061	2.500	2.691	7.050	9.354	8.260	5.245
	All Positives	164.357	164.357	164.357	164.357	164.357	164.357	164.357
	Coverage	4,30	1,52	1,64	4,29	5,69	5,03	3,19
METHODS	True Positives	2.614	777	645	2.697	3.035	2.818	1.996
	All Positives	29.095	29.095	29.095	29.095	29.095	29.095	29.095
	Coverage	8,98	2,67	2,22	9,27	10,43	9,69	6,86
KEGGs	True Positives	1.097	258	125	1.204	1.266	1.233	930
	All Positives	5.485	5.485	5.485	5.485	5.485	5.485	5.485
	Coverage	20,00	4,70	2,28	21,95	23,08	22,48	16,96
LOCATION	True Positives	312	161	53	373	336	305	263
	All Positives	4.319	4.319	4.319	4.319	4.319	4.319	4.319
	Coverage	7,22	3,73	1,23	8,64	7,78	7,06	6,09

Assessment of DDI to improve PPI

		3did	CBM	DPEA	iPFAM	PIBASE	PSIbase	SNAPPI
APID	True Positives	7.061	2.500	2.691	7.050	9.354	8.260	5.245
	Avg Random Positives	30,89	86,44	43,41	77,91	96,85	89,30	33,30
	Sd Random Positives	28,63	50,23	32,12	44,26	66,88	57,13	24,58
	Z-Score	245,54	48,05	82,43	157,54	138,84	143,02	212,02
METHODS	True Positives	2.614	777	645	2.697	3.035	2.818	1.996
	Avg Random Positives	24,85	90,74	47,98	74,73	97,26	94,43	31,89
	Sd Random Positives	21,65	49,28	34,74	43,58	60,27	57,50	22,67
	Z-Score	119,57	13,93	17,19	60,17	48,74	47,37	86,62
KEGGs	True Positives	1.097	258	125	1.204	1.266	1.233	930
	Avg Random Positives	32,00	85,70	44,07	77,43	90,19	86,25	28,14
	Sd Random Positives	26,42	46,84	30,51	42,09	62,90	57,28	23,05
	Z-Score	40,31	3,68	2,65	26,77	18,69	19,91	39,13
LOCATION	True Positives	312	161	53	373	336	305	263
	Avg Random Positives	24,91	75,62	46,75	76,10	98,46	94,70	27,84
	Sd Random Positives	22,43	50,78	30,73	45,03	68,03	59,96	23,33
	Z-Score	12,80	1,68	0,20	6,59	3,49	3,51	10,08

ACKNOWLEDGEMENTS

Carlos Prieto acknowledge his travel fellowship funding by Embrace and his research grant for PhD from Junta de Castilla y Leon (ref. BOCyL no. 119, EDU/777/2005).

The research group acknowledge the funding and support by Spanish Ministerio de Sanidad y Consumo (ISCIII FIS ref. PI061153) and by Junta de Castilla y Leon (ref. CSI03A06).

- Congreso: Targeting and Tinkering with Interaction Networks
  Fecha: 14-16 Abril de 2008 Lugar: Barcelona (España)
  Autores: Prieto C. y De Las Rivas J.
  Título: "Functional exploration of a curated human gene coexpression network and relations with human and protein interactome network"

Functional exploration of a curated human gene coexpression network and relations with human and protein interactome network

Prieto C. & De Las Rivas J.
e-mail: cprietos@usal.es
Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC, USAL/CSIC)
Salamanca, Spain

HTTP://BIOINFOW.DEP.USAL.ES/APID

DESCRIPTION

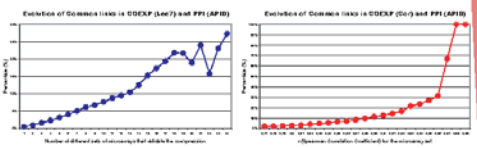
Well-curated integration of proteome-wide interaction and genome-wide coexpression data is essential to Systems Biology. This is based on the principle that interacting proteins may have similar gene expression and regulation profiles, but this link will occur only in some specific cellular processes. We explore when the functional correlation between protein interaction networks (PPI) and gene co-expression (COEXP) networks is satisfied.

PPI & COEXP NETWORKS

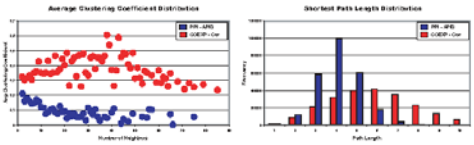
Studied Networks

	Number of Proteins	Number of Links	Connectivity
PPI - APID Agile Protein Interaction DataAnalyzer Human Interactome	12.047	58.232	5,67
COEXP - Lee7 (Lee et al. N>=7) Genes correlated (Pearson r>0.76) in more than 6 microarrays sets	827	4.399	10,64
COEXP - Cor Crossvalidated Coexpression data in 48 heterogeneous microarrays	3.052	12.670	8,30

Correlation between PPI and COEXP data



Main differences between PPI & COEXP networks

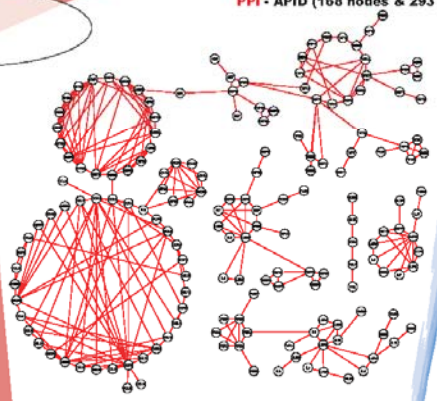


Network Parameters	Cluster Coefficient	Network Diameter	Characteristic Path Length	Avg Number of Neighbors
PPI - APID	0,129	19	4.092	5,87
COEXP - Lee7	0,332	21	6.216	10,64
COEXP - Cor	0,224	16	5.813	7,61

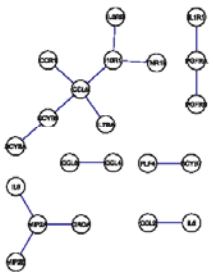
TOPOLOGICAL FEATURES

Citokines Receptors Networks

PPI - APID (168 nodes & 293 links)



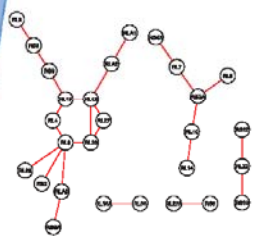
COEXP - LEE7 (21 proteins & 15 links)



Lack of coexpression between citokines & receptors

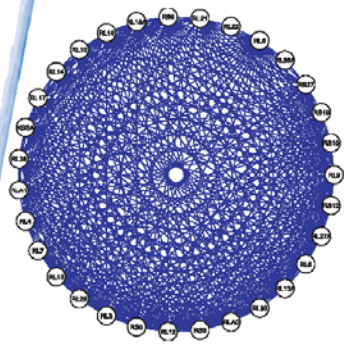
Ribosome Networks

PPI - APID (28 nodes & 25 links)



Lack of PPI coverage about human ribosome

COEXP - LEE7 (31 proteins & 380 links)

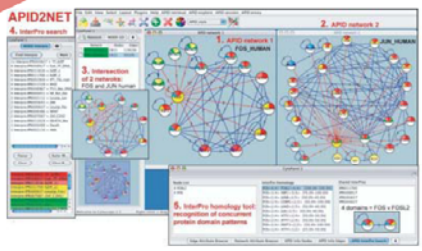


METHODS & TOOLS

http://bioinfow.dep.usal.es/apid

APID2NET (interactome graphic analyzer)

APID2net is a new a Cytoscape plugin for APID. This plugin is designed to visualize dynamically the nodes (proteins), the edges (interactions) and all the quality information that is associated to PPIs in APID. APID2net also provides GO terms annotation for interactive analysis to locate the proteins in the network that are assigned to specific functions. All these features, joined to CytoScape capabilities, make APID2net a helpful tool for the research community interested in exploring the interactomes.



APID (Agile Protein Interaction DataAnalyzer)

APID is an interactive bioinformatic web-tool that has been developed to allow exploration and analysis of main currently known information about protein-protein interactions integrated and unified in a common and comparative platform. The analytical and integrative effort done in APID provides an open access frame where all known experimentally validated protein-protein interactions (BIND, BioGRID, DIP, HPRD, IntAct and MINT) are unified in a unique web application

REFERENCES

Prieto C and De Las Rivas J.
APID: Agile Protein Interaction DataAnalyzer.
Nucleic Acids Res. 34 (2006):W298-302.

H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin and P. Pavlidis.
Coexpression analysis of human genes across many microarray data sets. Genome Res. 14 (2004): 1085–1094.

ACKNOWLEDGEMENTS

Carlos Prieto acknowledge his research grant for PhD from Junta de Castilla y Leon (ref. BOCyL no. 119, EDU/777/2005).

The research group acknowledge the funding and support by Spanish Ministerio de Sanidad y Consumo (ISCIII FIS grants ref. PI030920 and PI061153) and by Junta de Castilla y Leon (grant ref. CSI03A06).

Junta de Castilla y León

Instituto de Salud Carlos III

Targeting and Tinkering with Interaction Networks, Barcelona (Spain)

- Congreso: 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2007)
  Fecha: 21-25 Julio de 2007 Lugar: Viena (Austria)
  Autores: Prieto C. y De Las Rivas J.
  Título: "Functional coupling of human gene coexpression and protein interaction networks."

Functional coupling of human gene coexpression and protein interaction networks

Prieto C. & De Las Rivas J.

e-mail: cprietos@gmail.com

Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC, USAL/CSIC)
Salamanca, Spain

HTTP://BIOINFOW.DEP.USAL.ES/APID

DESCRIPTION

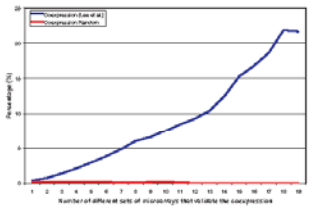
Well-curated integration of proteome-wide interaction and genome-wide coexpression data is essential to Systems Biology. This is based on the principle that interacting proteins may have similar gene expression and regulation profiles, but this link will occur only in some specific cellular processes. We explore when the functional correlation between protein interaction networks (PPI) and gene co-expression (COEXP) networks is satisfied.

PPI & COEXP NETWORKS

Studied Networks	Number of Proteins	Number of Links	Connectivity
PPI - APID Agile Protein Interaction DataAnalyzer Human Interactome	11.611	49.888	8,59
COEXP - Lee7 (Lee et al. N°=7) Genes correlated (Pearson r>0.75) in more than 6 microarray sets	827	4.399	10,64
COEXP - Cor5 Cross-validated Coexpression data in 43 heterogeneous microarrays	2.071	17.702	17,10
COEXP - Cor75 Genes correlated (Pearson r>0.75) in 43 heterogeneous microarrays	9.414	259.934	55,22

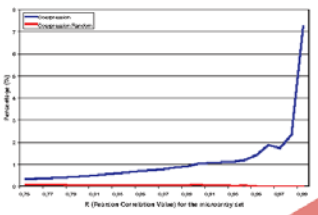
PPI & COEXP Correlation

Evolution of common links in COEXP and PPI networks



Lee & APID

Evolution of common links in COEXP and PPI networks

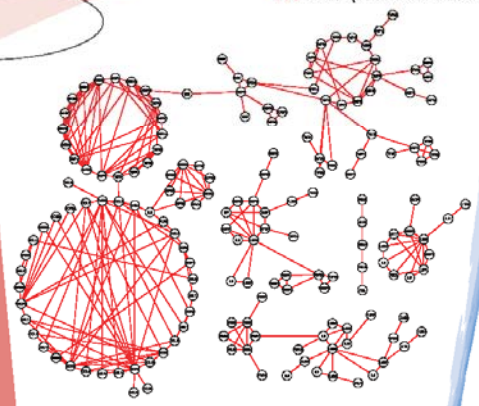


Cor75 & APID

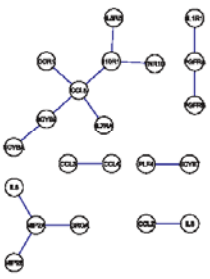
TOPOLOGICAL FEATURES

Citokines Receptors Networks

PPI - APID (168 nodes & 293 links)



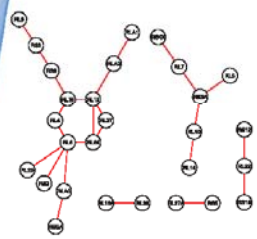
COEXP - LEE7 (21 proteins & 15 links)



Lack of coexpression between citokines & receptors

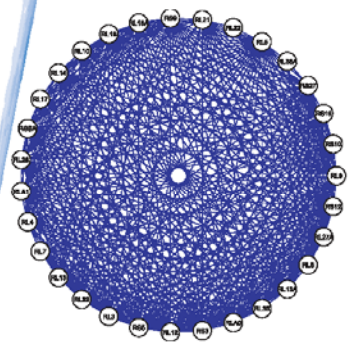
Ribosome Networks

PPI - APID (28 nodes & 25 links)



Lack of PPI coverage about human ribosome

COEXP - LEE7 (31 proteins & 380 links)



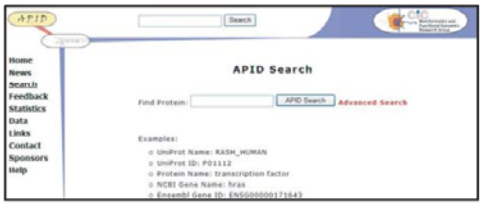
METHODS & TOOLS

APID (Agile Protein Interaction DataAnalyzer)

http://bioinfow.dep.usal.es/apid

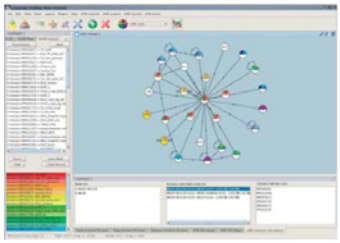
APID is an interactive bioinformatics web tool developed to integrate and analyze in a unified and comparative platform main currently known information about protein–protein interactions demonstrated by specific small-scale or large-scale experimental methods.

The web includes search tools to query and browse upon the data, allowing selection of the interaction pairs based in calculated parameters that weight and qualify the reliability of each given protein interaction. Such parameters are for the proteins: connectivity, cluster coefficient, Gene Ontology (GO) functional environment, GO environment enrichment; and for the interactions: number of methods, GO overlapping, IPfam domain-domain interaction.



APID2NET (new interactome graphic analyzer)

APID2net is a new a Cytoscape plugin for APID. This plugin is designed to visualize dynamically the nodes (proteins), the edges (interactions) and all the quality information that is associated to PPIs in APID. APID2net also provides GO terms annotation for interactive analysis to locate the proteins in the network that are assigned to specific functions. All these features, joined to CytoScape capabilities, make APID2net a helpful tool for the research community interested in exploring the interactomes.



REFERENCES

Prieto C and De Las Rivas J.
APID: Agile Protein Interaction DataAnalyzer.
Nucleic Acids Res. 34 (2006):W298-302.

H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin and P. Pavlidis.
Coexpression analysis of human genes across many microarray data sets. Genome Res. 14 (2004): 1085–1094.

ACKNOWLEDGEMENTS

Carlos Prieto acknowledge the travel fellowship funding by the Sixth Framework Programme of the European Union (Biosapiens project) and his research grant for PhD from Junta de Castilla y Leon (ref. BOCyL no. 119, EDU/777/2005).

The research group acknowledge the funding and support by Spanish Ministerio de Sanidad y Consumo (ISCIII FIS grants ref. PI030920 and PI051153) and by Junta de Castilla y Leon (grant ref. CSI03A06).

BIOSAPIENS NETWORK · Junta de Castilla y León · Instituto de Salud Carlos III

ISMB/ECCB 07, Vienna (Austria)

- Congreso: 5th European Conference On Computational Biology (ECCB2006)
  Fecha: 21-24 Enero de 2007 Lugar: Eilat (Israel)
  Autores: Prieto C. y De Las Rivas J.
  Título: "APID, an Integrated Web Platform to Explore and Evaluate Protein-Protein Interaction Networks. "

APID, an integrated web platform to explore and evaluate protein-protein interaction networks

Prieto C., Hernández-Toro J. & De Las Rivas J.

e-mail: cprietos@gmail.com

Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC, USAL/CSIC)
Salamanca, Spain

APID USE

http://bioinfow.dep.usal.es/apid

HTTP://BIOINFOW.DEP.USAL.ES/IAPID

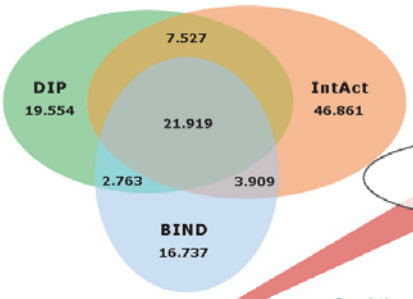
DESCRIPTION

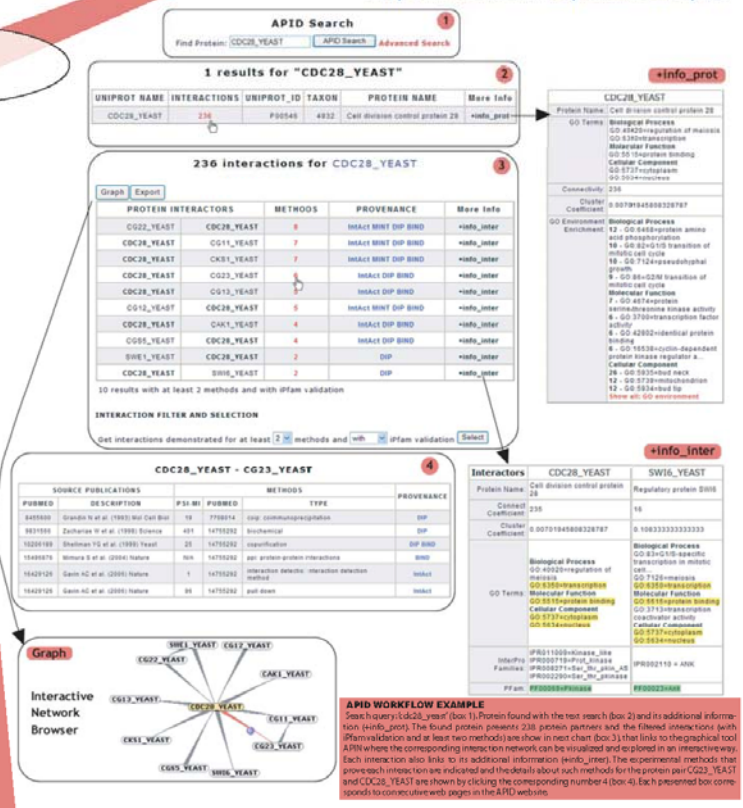
The use of biotechnological high-throughput experimental methods and genome and proteome-wide computational methods has flood the field of protein interactions and interactomes. Large-scale methods are producing huge amounts of data and many databases and bioinformatics tools are being used to store and explore the interactomes drafted. However, low quality of high-throughput data and low overlapping of interactome databases are clear concerns in the field. Therefore, before getting many inductive and general conclusions we need strategies to better **evaluate**, **validate** and **integrate** protein-protein interaction information as a way to obtain correct and useful scientific knowledge. A first step to achieve this is to improve protein interaction data integration and to provide means that allow objective evaluation of the data quality and biological reliability. We address this aim in APID (Agile Protein Interaction DataAnalyzer) where we integrate and unify 5 protein interaction source databases and provide a set of parameters to allow characterization of the key elements of an interactome network. For the proteins, parameters derived are associated to protein function (GO environment and GO environment enrichment), structure (domain composition) or location in the network (connectivity, cluster coefficient). For the interactions, parameters derived are associated to the **methods** that validate an interaction, i.e. the number of biochemical methods including publications reference; the overlapping of functional **GO terms** assigned to each protein-pair; the existence of structural domains that are known to interact because they are included in the **IPfam** domain-domain interaction database. We show that these parameters can be used to unravel and better explore the biological information present in complex protein interaction networks.

In this way APID is a unified web server that allows to evaluate the interactions as well as the proteins in the protein interaction network. APID also allows functional predictions based on the assumption that the linking proteins tend to have related functions or be involved in similar biological processes.

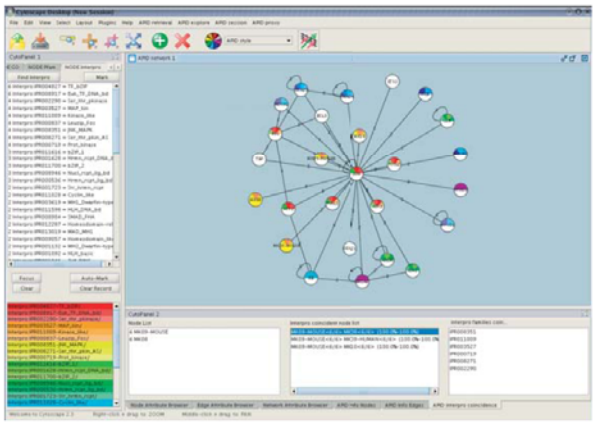
STATISTICS

Number of Proteins	42.699
Number of Interactions	156.668
Interactions validated with iPfam	15.250
Interactions with B.P. GO term overlap	14.834
Interactions with M.F. GO term overlap	43.098
Interactions with C.L. GO term overlap	20.170
Interactions validated in 2 or + methods	15.250





APID2NET



Description:

APID2net is a new a Cytoscape plugin for APID. This plugin is designed to visualize dynamically the nodes (proteins), the edges (interactions) and all the quality information that is associated to PPIs in APID. APID2net is open access and provides an intuitive and clear interface within Cytoscape to allow interactive on-line data subsetting from APID and to navigate on them, focusing on some areas or on some specific proteins, applying restrictions to queried data and linking out for more information. APID2net also provides GO terms annotation for interactive analysis to locate the proteins in the network that are assigned to specific functions. All these features, joined to CytoScape capabilities, make APID2net a helpful tool for the research community interested in exploring the interactomes.

Features:

- Interactive Visualization of Interaction Networks from APID.
- The interaction network can be retrieved using several interaction quality filters.
- The network can be explored using different protein and interaction attributes.
- Automatic expand to new interactions from any node.
- All the APID information is displayed in enclosed panels.
- GO Terms surfing: Proteins with significant GO terms can be selected and marked with a custom color pie portion.
- Pfam domains surfing: Proteins with a query Pfam can be selected and marked with a custom color pie portion.
- InterPro surfing: Proteins with a InterPro identifier can be selected and marked with a custom color pie portion.

REFERENCE

Prieto C and De Las Rivas J. (2006).
APID: Agile Protein Interaction DataAnalyzer.
Nucleic Acids Res. 34:W298-302.

ACKNOWLEDGEMENTS

Carlos Prieto acknowledge the travel fellowship funding by BioSapiens and his research grant for PhD from Junta de Castilla y Leon (ref. BOCyL no. 119, EDU/777/2005).

The research group acknowledge the funding and support by Spanish Ministerio de Sanidad y Consumo (ISCIII FIS grants ref. PI030920 and PI061153) and by Junta de Castilla y Leon (grant ref. CSI03A06).

BIOSAPIENS NETWORK · Junta de Castilla y León · Instituto de Salud Carlos III

ECCB 2006, Eilat (Israel)

- Congreso: VI Jornadas Nacionales de Bioinformática
  Fecha: 20-23 Noviembre de 2006 Lugar: Zaragoza (España)
  Autores: Prieto C., Rosón B. y De Las Rivas J.
  Título: "Defined groups of deregulated genes found in cancer samples by AlteredExpression algorithm."

Defined groups of deregulated genes found in cancer samples by AlteredExpression algorithm

Carlos Prieto, Beatriz Rosón and Javier De Las Rivas
e-mail: cprietos@usal.es

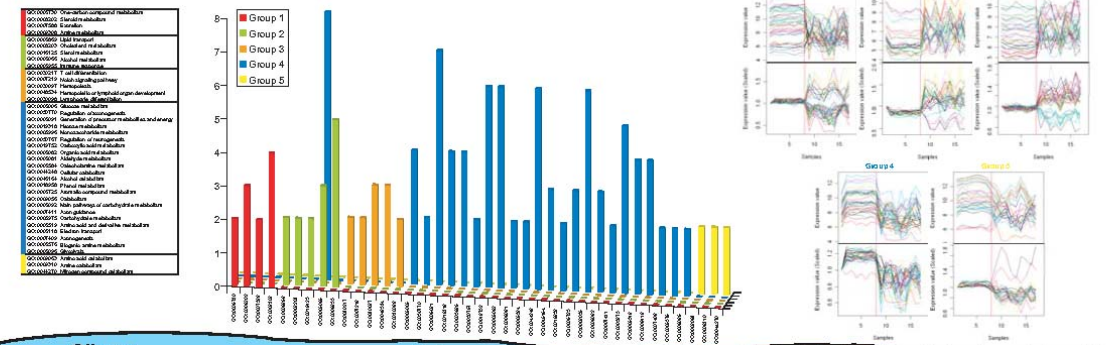
Bioinformatics and functional genomics research group
(IBMCC-CiC, CSIC-USAL) - E37007 Salamanca – Spain.

Genome-wide expression microarrays technology allows to interrogate at once the transcription state of thousand of genes and can be done for many biological samples to investigate the **expression profiles** of genes at a genomic scale. Gene expression is a tightly regulated process, crucial for the proper functioning of a cell. In data from expression microarrays, common regulation is reflected by strong correlation between expression levels, and it is usual that genes of similar function yield similar expression profiles across a diverse range of conditions.

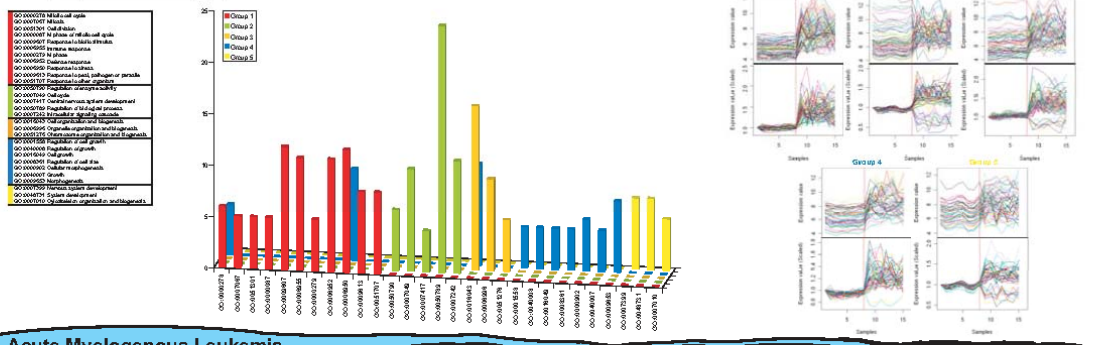
Molecular disease mechanisms typically constitute **abnormalities** in the regulation of genes producing **strong alterations** in the expression levels. The resulting changes in expression profiles can help to identify **disease-related genes**, and can also facilitate improved diagnosis and even prognosis of disease outcome. Alteration of gene regulation often results in expression profiles with large **variability**. However, common analysis strategies look only for differentially expressed genes (overexpressed or repressed), but do not explore the **variability** that appears in an **altered state** when compared with a **control state**. These gene expression profiles should be common in cancer diseases that produce **lack of control** of lots of cellular processes and, therefore, the **gene expression chaos**. For this reason, is quite interesting the search of genes that have a coexpression pattern in normal states and lost this pattern when the biological samples have a disease. So, recently we developed a new algorithm called **AlteredExpression** (http://bioinfow.dep.usal.es/AlteredExpression/), appropriate to find genes which present a significant alteration in the variability of their expression profiles when comparing a sample set in altered state versus a sample set in control state. This algorithm, based on the calculation of relative residual variances, provides several groups of genes and assigns a statistical measure of significance to each group of genes selected (Prieto et al., 2006).

Once, a new **algorithm** in bioinformatics has been developed it is quite important to investigate about its real biological usefulness and efficiency when applied upon a large spectrum of real experimental data. This is the only way to validate the practical meaning of an algorithm. Based on this statement, in the present work we investigate how is the consistency and biological functional meaning of the groups of altered-genes found by the **AlteredExpression algorithm** in 4 different sets of human cancer samples (acute leukemia, osteosarcoma, renal carcinoma and glioma) and their respective normal controls. The genome-wide expression data of all these biological samples was obtained using human Affymetrix microarrays (hgu133a). The results show that in all cases several groups of significant genes were obtained for each dataset. The groups found had high confidence scores and were stable under several runs of the algorithm. Moreover, the functional analyses of these groups done through Gene Ontology (GO) show that there are statistically consistent differences in biological function between the different groups selected. Biological annotation also indicates that the most deregulated groups have genes that are functionally related with the diseases. In this way, **we prove the existence of clear patterns of deregulation at least in cancer samples**.

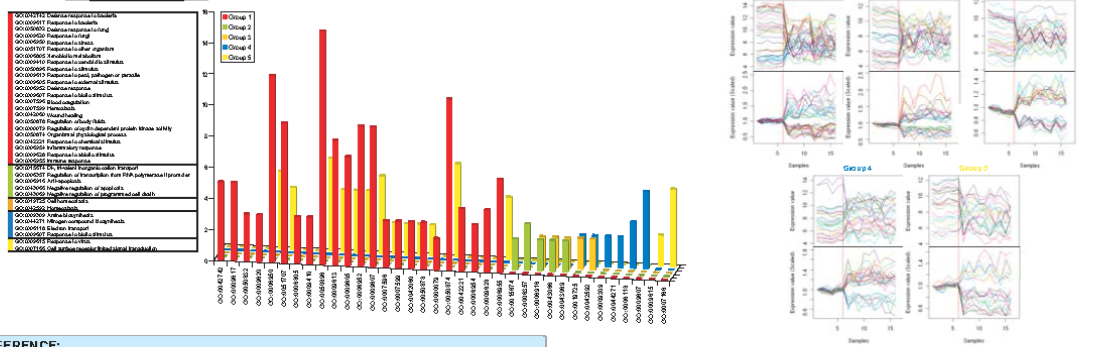
Renal Carcinoma



Glioma



Acute Myelogenous Leukemia



REFERENCE:
C. Prieto , M.J. Rivas , J.M. Sánchez , J. López-Fidalgo , and J. De Las Rivas (2006)
Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes.
Bioinformatics 2006 22(9):1103-1110.

Jdb 2006

- Congreso: Interactome Networks
  Fecha: 1-3 Septiembre de 2006 Lugar: Hinxton (UK)
  Autores: Prieto C. y De Las Rivas J.
  Título: "Integration of protein protein interaction data in a unified platform with quality assessment parameters. "

APID

Agile Protein Interaction DataAnalizer

Integration of protein-protein interaction data in a unified platform with quality assessment parameters

Prieto C. & De Las Rivas J.
e-mails: cprietos@usal.es & jrivas@usal.es

Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC, USAL/CSIC)
Salamanca, Spain

DESCRIPTION

http://bioinfow.dep.usal.es/apid

INTRODUCTION

At present time one of the most productive areas of biological data is protein-protein interactions. The data about the interaction of two or more proteins are stored in published scientific papers where the information is difficult to manage and compute. For this reason several bioinformatic initiatives have been undertaken to store, in biological databases, information about protein interactions. These initiatives tend to extract and integrate experimental knowledge about interacting proteins from scientific journals in their database. Each initiative has its own extraction, curation and storage protocols. This is the reason why the intersection and overlap between these source databases is small, and therefore in many cases their information is complementary and can be unified to increase our knowledge about interactomes of different species.

APPLICATION DATA INTEGRATION AND UNIFICATION

The data unification has been done based on three key reference identifiers: (i) UniProt ID entry name, to allow a specific identification of each protein; (ii) PSI-MI ID, to unify the experimental methods used in different publications to a common terminology developed by PSI-MI; (iii) PubMed ID, to attach each interaction validated with a given experimental method to a specific PubMed literature reference. These three key identifiers allow only to get additional data about the proteins or about the interactions by link to other biological data sources. In this way, we have obtained a protocol able to store and unify protein interaction databases in a clear uniform structure, maintaining the integrity of the data and correcting some existing failures found in the original files.

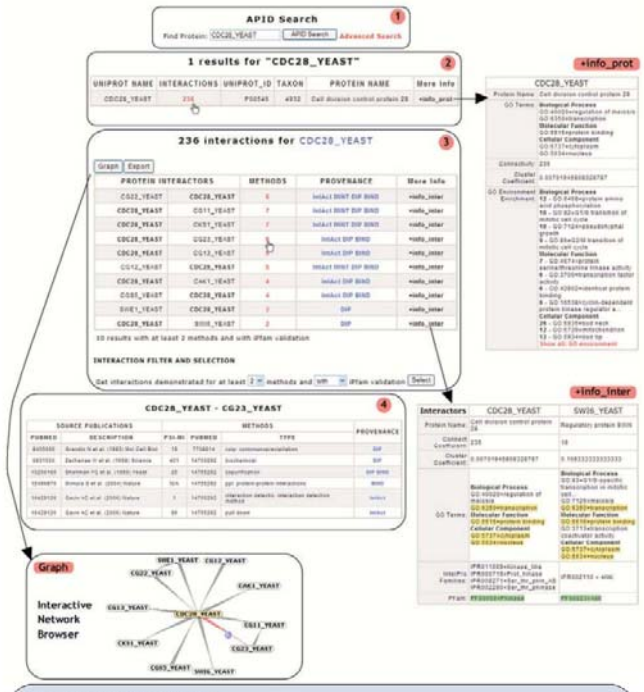
INTERACTION ASSESSMENT

Large-scale methods are producing huge amounts of data and many databases and bioinformatics tools are being used to store and explore the interactomes drafted. However, low quality of high-throughput data and low overlapping of interactome databases are clear concerns in the field. Therefore, before getting many inductive and general conclusions we need strategies to better evaluate, validate and integrate protein-protein interaction information as a way to obtain correct and useful scientific knowledge. A first step to achieve this is to improve protein interaction data integration and to provide means that allow objective evaluation of the data quality and biological reliability. We address this aim in APID (Agile Protein Interaction DataAnalyzer) where we integrate and unify 5 protein interaction source databases and provide a set of parameters to allow characterization of the key elements of an interactome network: proteins (nodes) and interactions (edges). For the proteins, parameters derived are associated to protein function (GO environment and GO environment enrichment), structure (domain composition) or location in the network (connectivity, cluster coefficient). For the interactions, parameters derived are associated to the methods that validate an interaction, i.e. the number of biochemical methods including publications reference (with a distinction between large-scale and small-scale methods and between experimental and predicted methods); the overlapping of functional GO terms assigned to each protein-pair; the existence of structural domains that are known to interact because they are included in the iPfam domain-domain interaction database. We show that these parameters can be used to unravel and better explore the biological information present in complex protein interaction networks.

In this way APID is a unified web server that allows to evaluate the interactions as well as the proteins in the protein interaction network. APID also allows functional predictions based on the assumption that the linking proteins tend to have related functions or be involved in similar biological processes.

For more information APID can be explored via web (**http://bioinfow.dep.usal.es/apid/**) and it includes a graphic interactive tool to visualize and browse the interaction network.

APID USE



APID WORKFLOW EXAMPLE

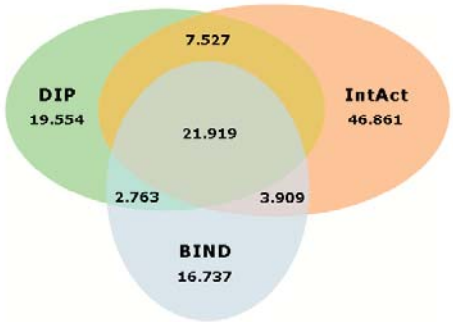
Search query: 'cdc28_yeast' (box 1). Protein found with the text search (box 2) and its additional information (+info_prot). The found protein presents 238 protein partners and the filtered interactions (with iPfam validation and at least two methods) are show in next chart (box 3), that links to the graphical tool APIN where the corresponding interaction network can be visualized and explored in an interactive way. Each interaction also links to its additional information (+info_inter). The experimental methods that prove each interaction are indicated and the details about such methods for the protein pair CG23_YEAST and CDC28_YEAST are shown by clicking the corresponding number 4 (box 4). Each presented box corresponds to consecutive web pages in the APID website.

STATISTICS

Number of Interactions:	128.703
Interactions with pfam annotation:	87.802
Interactions validated with iPfam:	10.984
Interactions validated in 2 or + methods:	11.971
Interactions with iPfam & 2 or + methods:	2.633

DIAGRAM

Venn diagram with the number of interactions for the multiple intersections between BIND, DIP and IntAct. 62% of the overall protein interactions included in BIND, DIP and IntAct are presented in only one of these databases, i.e. they are exclusive to one of the protein interaction resources.



REFERENCE

- **Prieto C. and De Las Rivas (2006). APID: Agile Protein Interaction DataAnnalyzer. Nucl. Acids Res. 2006 34: W298-W302.**

Acknowledgements:
We acknowledge the financial support from the Spanish Ministry of Health (FIS, ISCIII, Spain), from Fundacion BBVA (Spain) and from Junta de Castilla y Leon (JCyL).

- Congreso: 4th European Conference On Computacional Biology (ECCB05)
  Fecha: 28 Septiembre – 1 Octubre de 2005 Lugar: Madrid (España)
  Autores: Prieto C. y De Las Rivas J.
  Título: "Interactome unified database for protein network mining. "

APID

Agile Protein Interaction Database

Prieto C. & De Las Rivas J.*
*e-mail: jrivas@usal.es

Bioinformatics and Functional Genomics Research Group
Cancer Research Center (CIC, USAL/CISIC)
Salamanca, Spain

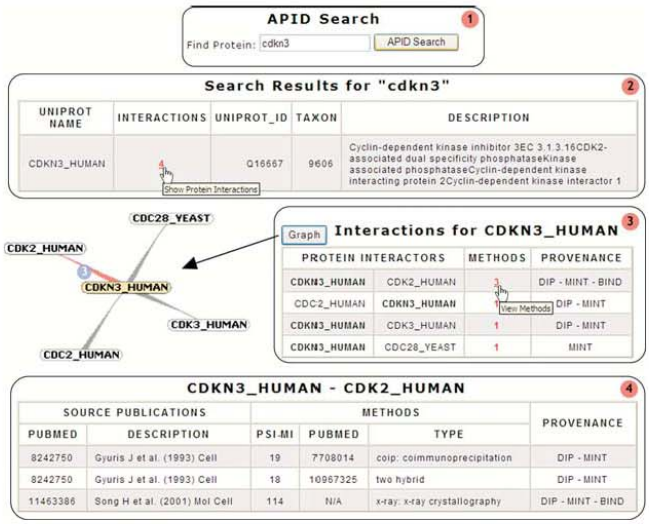
Description

http://bioinfow.dep.usal.es/apid

Agile Protein Interaction Database (APID) is a unified relational database that integrates in a common and comparative platform current public available protein-protein interaction data, keeping a lightweight and adaptable structure. At present, the included interaction data come from five main source databases: **DIP***, **BIND**, **IntAct, MINT** and **HPRD***. These source provide different data formats not synchronized, and their comparison indicates that the data sets are heterogeneous and have **little overlap** (see statistics). In APID the comprehensive integrative effort done provides a bioinformatic resource where all known protein interactions are unified in a unique common data structure. The database puts strong emphasis in the experimental evidences that validate each interaction with clear links to the publications and databases that report such evidences. Moreover, the database uses **UniProt** (names, IDs, etc) as main keys to identify each protein, allowing link to each protein sequence and to other protein features. APID can be explored via web (**http://bioinfow.dep.usal.es/apid**) and it includes a graphic interactive tool to visualize and browse the interaction network.

* At present it can't be accessed for license terms.

APID USE



Scheme showing APID search-flow example.

Search query: "cdkn3" (chart 1). Proteins found with the text search: CDKN3_HUMAN (chart 2). The found protein presents four interactions and it's protein partners are show in next chart (chart 3), that links to the graphical tool APIN where the corresponding interaction network can be visualize and explored in an interactive way. The experimental methods that prove each interaction are indicated and the details about such methods for the protein pair CDKN3_HUMAN and CDK2_HUMAN are shown by clicking in the corresponding number 3 (chart 4). Each chart presented corresponds to consecutive web pages in the APID web site.

STATISTICS

DATABASES	INTERACTIONS
DIP & IntAct & MINT & BIND	21471
BIND	20403
IntAct	14601
HPRD	11203
DIP	8345
DIP & IntAct & MINT	7510
DIP & MINT	6532
MINT	5788
IntAct & MINT	5299
DIP & BIND	3179
IntAct & BIND	1862
IntAct & MINT & BIND	1759
BIND & HPRD	958
DIP & IntAct	840
DIP & IntAct & BIND	553
DIP & MINT & BIND	552
MINT & HPRD	469
MINT & BIND	415
DIP & HPRD	259
IntAct & HPRD	195
IntAct & BIND & HPRD	117
MINT & BIND & HPRD	114
DIP & BIND & HPRD	75
DIP & MINT & HPRD	61
IntAct & MINT & HPRD	44
IntAct & MINT & BIND & HPRD	35
DIP & MINT & BIND & HPRD	33
DIP & IntAct & MINT & BIND & HPRD	28
DIP & IntAct & MINT & HPRD	26
DIP & IntAct & BIND & HPRD	11
DIP & IntAct & HPRD	8

Acknowledgements: We acknowledge the financial support from the Spanish Ministry of Health (FIS, ISCIII, Spain) and from Foundation BBVA (Spain) to develop our scientific research on bioinformatics and functional genomics.

ECCB 05, Madrid, Spain

- Congreso: V Jornadas de Bioinformática jbi2004
  Fecha: 29 Noviembre – 30 Diciembre de 2004 Lugar: Barcelona (España)
  Autores: Prieto C., De Luis A. y De Las Rivas J.
  Título: "Integration and comparison of interactome databases"

Bioinformatics & Functional Genomics Research Group Cancer Research Center (CIC)

Prieto C., De Luis A. & De Las Rivas J.
Cancer Research Center (CIC, CSIC / USAL)
CSIC / University of Salamanca
Salamanca, Spain (cprietos@usal.es)

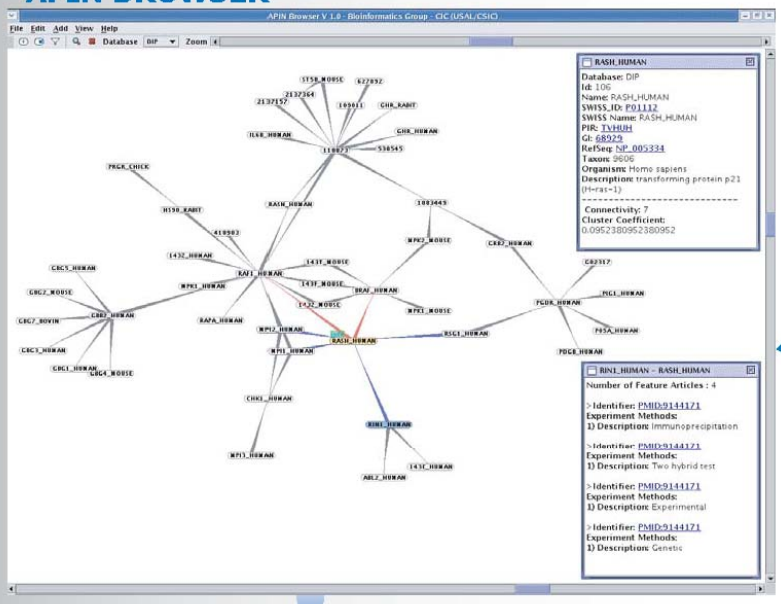
APIN: Integration and comparison of interactome databases (http://bioinfow.dep.usal.es/apin/)

APIN (Agile Protein Interaction Network) is a bioinformatics package designed to view, analyze, explore and browse the nodes (proteins) and edges (interactions) of an integrated protein interaction database, which includes the information coming from several public protein interaction databases.

Nowadays, bioinformatics tools are not able to display and correlate interactome data from different databases. We have designed and built a database able to store the main information contained in the most important public protein interaction databases. In this way we have obtained DIP, BIND, INTACT and MINT data stored in a new uniform structure, maintaining the integrity of their data and correcting some existing failures in the files that jeopardized the structure and the coherence of the protein interaction data. In order to access to our integrated database we have developed a program based on a JAVA Applet that is the specific visualization tool included in APIN package, the applet can be executed like a Web application (http://bioinfow.dep.usal.es/apin/) from our Website. The tool can visualize dynamically the information contained in the interactome databases, do an agile navigation by the showed interaction network, select/deselect some proteins and get the information about the proteins and interactions of interest. Moreover the tool has an advanced search engine which allows specific pre-selection of a group of proteins and interactions to visualize them. Finally, we have calculated for each node (each protein) some key graph parameters that reflect the degree of connectivity and clustering of each protein.

APIN BROWSER

Fig 1



APIN Browser
Navigation tools
Interactive Navigation
Advanced Search
Search Filters
Edition Mode
Analytical tools
Graph Parameters
Shortest Path

APIN Browser use an intuitive and clear interface to navigate into complex interactome databases, the browser have several key features that will allow you to view, analyze, explore and browse interaction networks. Within **APIN** you can use:

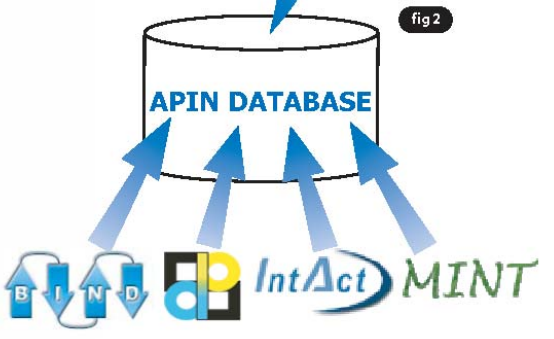
NAVIGATION TOOLS

- Easy access to protein and interaction information (including the experimental methods that probe the interactions).
- **Advanced search** in our integrate protein interaction database with a set of filters that allows to apply restrictions to the queried data.
- Dynamic and **interactive visualization** of protein interaction networks.
- **Modify** the protein network displayed (**edition mode**).

ANALYTICAL TOOLS

- Show precalculated graph parameters like **connectivity** value and **cluster coefficient** for each node that will allow better exploration of the interactome features.
- Calculate the **shortest path** between each two proteins in the interactome network within Dijkstra algorithm.

fig 2



JBI 2004, BCN, SPAIN

REFERENCIAS WEB

[1] http://news.bbc.co.uk/1/hi/sci/tech/2940601.stm

[2] http://www.nlm.nih.gov

[3] http://thunder.biosci.umbc.edu/classes/biol414/spring2007/index.php/Co-immunoprecipitation

[4] http://www.protocolonline.org/prot/Molecular_Biology/Protein/Immunoprecipitation__IP_/Co-Immunoprecipitation__Co-IP_

[5] http://www.piercenet.com/Products/Browse.cfm?fldID=F3FD3612-415F-42A5-8922-736F9FDD36FB

[6] http://cshprotocols.cshlp.org/cgi/content/full/2007/6/pdb.top2

[7] http://imex.sourceforge.net

[8] http://www.blueprint.org/bind

[9] http://www.thebiogrid.org

[10] http://dip.doe-mbi.ucla.edu

[11] http://www.hprd.org

[12] http://www.hprd.org/PhosphoMotif_finder

[13] http://www.netpath.org

[14] http://www.ebi.ac.uk/intact

[15] http://mint.bio.uniroma2.it/mint

[16] http://pubchem.ncbi.nlm.nih.gov

[17] http://www. sciencemag.org/cgi/content/full/296/ 5569/827

[18] http://www.psidev.info/index.php?q=node/60

[19] http://java.sun.com

[20] http://java.sun.com/javaee

[21] http://java.sun.com/applets

[22] http://www.saxproject.org

[23] http://www.w3.org/DOM/

[24] http://java.sun.com/products/jdbc/overview.html

[25] http://www.mysql.com

[26] http://www.pubmed.gov

[27] http://medialab-prado.es

[28] http://sourceforge.net/projects/touchgraph

[29] http://chianti.ucsd.edu/cyto_web/plugins/plugindownloadstatistics.php

[30] http://www.gen-es.org/02_cono/docs/Microarrays.pdf

[31] http://www.affymetrix.com/

[32] http://en.wikipedia.org/wiki/Correlation

[33] http://www.garban.org/factory

[34] http://www.r-project.org

BIBLIOGRFAFÍA

Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M.J., Dumontier, M.R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J.P., Parker, B., Pintilie, G., Pirone, R., Salama, J.J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B.F. y Hogue, C.W. (2005). The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic acids research*, **33**, D418-424.

Aloy, P. y Russell, R.B. (2006). Structural systems biology: modelling protein interactions, *Nature reviews*, **7**, 188-197.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. y Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research*, **25**, 3389-3402.

Alwine, J.C., Kemp, D.J. y Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes, *Proc Natl Acad Sci U S A*, **74**, 5350-5354.

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. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bader, G.D. y Hogue, C.W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources, *Nature biotechnology*, **20**, 991-997.

Bader, G.D. y Hogue, C.W. (2003). An automated method for finding molecular

complexes in large protein interaction networks, *BMC Bioinformatics*, **4**, 2.

Bader, J.S. (2003). Greedily building protein networks with confidence, *Bioinformatics (Oxford, England)*, **19**, 1869-1874.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. y Yeh, L.S. (2005). The Universal Protein Resource (UniProt), *Nucleic acids research*, **33**, D154-159.

Barabasi, A.L. y Albert, R. (1999). Emergence of scaling in random networks, *Science*, **286**, 509-512.

Barabasi, A.L. y Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Benjamini, Y. y Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. Ser. B,*, **57**, 289–300.

Bolstad, B.M., Irizarry, R.A., Astrand, M. y Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics (Oxford, England)*, **19**, 185-193.

Bonferroni, C.E. (1937) Teoria statistica delle classi e calcolo delle probabilita. In, *In Volume in Onore di Ricarrdo dalla Volta*. Universita di Firenza, Italy, 1–62.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O. y Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution, *Genome Biol*, **5**, R35.

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K. y Tyers, M. (2008). The BioGRID Interaction Database: 2008 update, *Nucleic acids research*, **36**, D637-640.

Breitkreutz, B.J., Stark, C. y Tyers, M. (2003). Osprey: a network visualization system, *Genome Biol*, **4**, R22.

Ceol, A., Chatr-Aryamontri, A., Licata, L. y Cesareni, G. (2008). Linking entries in protein interaction database to structured text: the FEBS Letters experiment, *FEBS letters*, **582**, 1171-1177.

Cockburn, A. (2001) *Agile software development*. Addison-Wesley, Boston, Mass.

Cusick, M.E., Klitgord, N., Vidal, M. y Hill, D.E. (2005). Interactome: gateway into systems biology, *Hum Mol Genet*, **14 Spec No. 2**, R171-181.

Chang, L.W., Fontaine, B.R., Stormo, G.D. y Nagarajan, R. (2007). PAP: a

comprehensive workbench for mammalian transcriptional regulatory sequence analysis, *Nucleic acids research*, **35**, W238-244.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. y Cesareni, G. (2007). MINT: the Molecular INTeraction database, *Nucleic acids research*, **35**, D572-574.

Cheng, Y. y Church, G.M. (2000). Biclustering of expression data, *Proc Int Conf Intell Syst Mol Biol*, **8**, 93-103.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. y Botstein, D. (1998). SGD: Saccharomyces Genome Database, *Nucleic acids research*, **26**, 73-79.

Davis, F.P. y Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces, *Bioinformatics (Oxford, England)*, **21**, 1901-1907.

de Borst, M.H., Benigni, A. y Remuzzi, G. (2008). Primer: strategies for identifying genes involved in renal disease, *Nat Clin Pract Nephrol*, **4**, 265-276.

Deane, C.M., Salwinski, L., Xenarios, I. y Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations, *Mol Cell Proteomics*, **1**, 349-356.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M. y Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic acids research*, **36**, D344-350.

Draghici, S. (2003) *Data analysis tools for DNA microarrays*. Chapman & Hall/CRC, Boca Raton, Fla.

Eisen, M.B., Spellman, P.T., Brown, P.O. y Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, **95**, 14863-14868.

Eisenberg, E. y Levanon, E.Y. (2003). Human housekeeping genes are compact, *Trends Genet*, **19**, 362-365.

Enright, A.J., Van Dongen, S. y Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families, *Nucleic acids research*, **30**, 1575-1584.

Falvo, J.V., Parekh, B.S., Lin, C.H., Fraenkel, E. y Maniatis, T. (2000). Assembly of a functional beta interferon enhanceosome is dependent on ATF-2-c-jun heterodimer orientation, *Mol Cell Biol*, **20**, 4814-4825.

Finn, R.D., Marshall, M. y Bateman, A. (2005). iPfam: visualization of protein-protein

interactions in PDB at domain and amino acid resolutions, *Bioinformatics (Oxford, England)*, **21**, 410-412.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. y Bateman, A. (2008). The Pfam protein families database, *Nucleic acids research*, **36**, D281-288.

Fodor, S.P., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P. y Adams, C.L. (1993). Multiplexed biochemical assays with biological chips, *Nature*, **364**, 555-556.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. y Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis, *Science*, **251**, 767-773.

Futschik, M.E., Chaurasia, G. y Herzel, H. (2007). Comparison of human protein-protein interaction maps, *Bioinformatics (Oxford, England)*, **23**, 605-611.

Gardner, T.S. y Faith, J.J. (2005). Reverse-engineering transcription control networks, *Physics of Life Reviews*, **2**, 65 - 88.

Giorgini, F. y Muchowski, P.J. (2005). Connecting the dots in Huntington's disease with protein interaction networks, *Genome Biol*, **6**, 210.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. y Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H., Cho, Y., Han, K., Lee, S., Choi, H., Lappe, M., Holm, L., Kim, S., Oh, D. y Bhak, J. (2005). PSIbase: a database of Protein Structural Interactome map (PSIMAP), *Bioinformatics (Oxford, England)*, **21**, 2541-2543.

Graeber, T.G. y Eisenberg, D. (2001). Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles, *Nat Genet*, **29**, 295-300.

Griffith, O.L., Pleasance, E.D., Fulton, D.L., Oveisi, M., Ester, M., Siddiqui, A.S. y Jones, S.J. (2005). Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses, *Genomics*, **86**, 476-488.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma,

A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. y Apweiler, R. (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data, *Nature biotechnology*, **22**, 177-183.

Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D. y Cherry, J.M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods, *Nucleic acids research*, **36**, D577-581.

Hope, I.A. y Struhl, K. (1986). Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast, *Cell*, **46**, 885-894.

Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P., Weng, Z., Mutter, G.L., Frosch, M.P., Macdonald, M.E., Milford, E.L., Crum, C.P., Bueno, R., Pratt, R.E., Mahadevappa, M., Warrington, J.A., Stephanopoulos, G., Stephanopoulos, G. y Gullans, S.R. (2001). A compendium of gene expression in normal human tissues, *Physiol Genomics*, **7**, 97-104.

Hu, Z., Snitkin, E.S. y DeLisi, C. (2008). VisANT: an integrative framework for networks in systems biology, *Brief Bioinform*, **9**, 317-325.

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. y Flicek, P. (2009). Ensembl 2009, *Nucleic acids research*, **37**, D690-697.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale,

D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. y Yeats, C. (2009). InterPro: the integrative protein signature database, *Nucleic acids research*, **37**, D211-215.

Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J.S., White, K.P. y Rzhetsky, A. (2004). Probabilistic inference of molecular networks from noisy data sources, *Bioinformatics (Oxford, England)*, **20**, 1205-1213.

Iragne, F., Nikolski, M., Mathieu, B., Auber, D. y Sherman, D. (2005). ProViz: protein interaction visualization and exploration, *Bioinformatics (Oxford, England)*, **21**, 272-274.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. y Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data, *Nucleic acids research*, **31**, e15.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. y Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci U S A*, **98**, 4569-4574.

Jefferson, E.R., Walsh, T.P., Roberts, T.J. y Barton, G.J. (2007). SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions, *Nucleic acids research*, **35**, D580-589.

Kafatos, F.C., Jones, C.W. y Efstratiadis, A. (1979). Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure, *Nucleic acids research*, **7**, 1541-1552.

Kaiser, R.J., MacKellar, S.L., Vinayak, R.S., Sanders, J.Z., Saavedra, R.A. y Hood, L.E. (1989). Specific-primer-directed DNA sequencing using automated fluorescence detection, *Nucleic acids research*, **17**, 6087-6102.

Kandasamy, K., Keerthikumar, S., Goel, R., Mathivanan, S., Patankar, N., Shafreen, B., Renuse, S., Pawar, H., Ramachandra, Y.L., Acharya, P.K., Ranganathan, P., Chaerkady, R., Keshava Prasad, T.S. y Pandey, A. (2009). Human Proteinpedia: a unified discovery resource for proteomics research, *Nucleic acids research*, **37**, D773-781.

Kanehisa, M. (2002). The KEGG database, *Novartis Found Symp*, **247**, 91-101; discussion 101-103, 119-128, 244-152.

Keegan, L., Gill, G. y Ptashne, M. (1986). Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein, *Science*, **231**, 699-704.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C.,

Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. y Hermjakob, H. (2007). IntAct--open source resource for molecular interaction data, *Nucleic acids research*, **35**, D561-565.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J.J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stumpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M.E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R. y Hermjakob, H. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions, *BMC Biol*, **5**, 44.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. y Pandey, A. (2009). Human Protein Reference Database--2009 update, *Nucleic acids research*, **37**, D767-772.

Kostka, D. y Spang, R. (2004). Finding disease specific alterations in the co-expression of genes, *Bioinformatics (Oxford, England)*, **20 Suppl 1**, i194-199.

Kypriotou, M., Beauchef, G., Chadjichristos, C., Widom, R., Renard, E., Jimenez, S.A., Korn, J., Maquart, F.X., Oddos, T., Von Stetten, O., Pujol, J.P. y Galera, P. (2007). Human collagen Krox up-regulates type I collagen expression in normal and scleroderma fibroblasts through interaction with Sp1 and Sp3 transcription factors, *J Biol Chem*, **282**, 32000-32014.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. y Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets, *Genome research*, **14**, 1085-1094.

Li, C. y Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc Natl Acad Sci U S A*, **98**, 31-36.

Lim, W.K., Wang, K., Lefebvre, C. y Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks, *Bioinformatics (Oxford, England)*, **23**, i282-288.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. y Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays, *Nat Genet*, **21**, 20-24.

Liu, W.M., Mei, R., Di, X., Ryder, T.B., Hubbell, E., Dee, S., Webster, T.A., Harrington, C.A., Ho, M.H., Baid, J. y Smeekens, S.P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms, *Bioinformatics (Oxford, England)*, **18**, 1593-1599.

Loong, T.W. (2003). Understanding sensitivity and specificity with the right side of the brain, *Bmj*, **327**, 716-719.

Magee, C., Nurminskaya, M., Faverman, L., Galera, P. y Linsenmayer, T.F. (2005). SP3/SP1 transcription activity regulates specific expression of collagen type X in hypertrophic chondrocytes, *J Biol Chem*, **280**, 25331-25338.

Maglott, D., Ostell, J., Pruitt, K.D. y Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI, *Nucleic acids research*, **35**, D26-31.

Magwene, P.M. y Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence, *Genome Biol*, **5**, R100.

Maskos, U. y Southern, E.M. (1992). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ, *Nucleic acids research*, **20**, 1679-1684.

Maskos, U. y Southern, E.M. (1992). Parallel analysis of oligodeoxyribonucleotide (oligonucleotide) interactions. I. Analysis of factors influencing oligonucleotide duplex formation, *Nucleic acids research*, **20**, 1675-1678.

Montgomery, D.C. (2008) *Design and analysis of experiments*. Wiley.

Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J. y Hermjakob, H. (2007). Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition, *Proteomics*, **7 Suppl 1**, 28-34.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J.J., Moore, S., Wojcik, J., Bader, G.D., Vidal, M., Cusick, M.E., Gerstein, M., Gavin, A.C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V.M., Hogue, C., Mewes, H.W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G. y Hermjakob, H. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx), *Nature biotechnology*, **25**, 894-898.

Panne, D., Maniatis, T. y Harrison, S.C. (2004). Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer, *Embo J*, **23**, 4384-4393.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. y Fodor, S.P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis, *Proc Natl Acad Sci U S A*, **91**, 5022-5026.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A. y Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms, *BMC Bioinformatics*, **6 Suppl 4**, S21.

Poree, B., Kypriotou, M., Chadjichristos, C., Beauchef, G., Renard, E., Legendre, F., Melin, M., Gueret, S., Hartmann, D.J., Mallein-Gerin, F., Pujol, J.P., Boumediene, K. y Galera, P. (2008). Interleukin-6 (IL-6) and/or soluble IL-6 receptor down-regulation of human type II collagen gene expression in articular chondrocytes requires a decrease of Sp1.Sp3 ratio and of the binding activity of both factors to the COL2A1 promoter, *J Biol Chem*, **283**, 4850-4865.

Prieto, C. y De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer, *Nucleic acids research*, **34**, W298-302.

Prieto, C., Rivas, M.J., Sanchez, J.M., Lopez-Fidalgo, J. y De Las Rivas, J. (2006). Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes, *Bioinformatics (Oxford, England)*, **22**, 1103-1110.

Pruitt, K.D., Tatusova, T. y Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research*, **35**, D61-65.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. y Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification, *Methods*, **24**, 218-229.

Ramirez, F., Schlicker, A., Assenov, Y., Lengauer, T. y Albrecht, M. (2007). Computational analysis of human protein interaction networks, *Proteomics*, **7**, 2541-2552.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. y Chinnaiyan, A.M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, *Cancer Res*, **62**, 4427-4433.

Saiki, R.K., Walsh, P.S., Levenson, C.H. y Erlich, H.A. (1989). Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes, *Proc Natl Acad Sci U S A*, **86**, 6230-6234.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. y Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update, *Nucleic acids research*, **32**, D449-451.

Schena, M., Shalon, D., Davis, R.W. y Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

Segal, E., Wang, H. y Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data, *Bioinformatics (Oxford, England)*, **19 Suppl 1**, i264-271.

Segal, E., Yelensky, R. y Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression, *Bioinformatics (Oxford, England)*, **19 Suppl 1**, i273-282.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. y Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome research*, **13**, 2498-2504.

Shoemaker, B.A. y Panchenko, A.R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases, *PLoS Comput Biol*, **3**, e42.

Shoemaker, B.A., Panchenko, A.R. y Bryant, S.H. (2006). Finding biologically relevant protein domain interactions: conserved binding mode analysis, *Protein Sci*, **15**, 352-361.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L. y Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature biotechnology*, **25**, 1251-1255.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis, *J Mol Biol*, **98**, 503-517.

Stein, A., Panjkovich, A. y Aloy, P. (2009). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure, *Nucleic acids research*, **37**, D300-304.

Stelzl, U. y Wanker, E.E. (2006). The value of high quality protein-protein interaction networks for systems biology, *Curr Opin Chem Biol*, **10**, 551-558.

Stuart, J.M., Segal, E., Koller, D. y Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules, *Science*, **302**, 249-255.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R. y Hogenesch, J.B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*, **101**, 6062-6067.

Suojanen, J.N. (1999). False false positive rates, *N Engl J Med*, **341**, 131.

Suzuki, R. y Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics (Oxford, England)*, **22**, 1540-1542.

Teorey, T.J. (1999) *Database modeling & design*. Morgan Kaufmann, San Francisco, Calif.

Tirosh, I., Weinberger, A., Carmi, M. y Barkai, N. (2006). A genetic signature of interspecies variations in gene expression, *Nat Genet*, **38**, 830-834.

Tukey, J.W. (1977). Some thoughts on clinical trials, especially problems of multiplicity, *Science*, **198**, 679-684.

Tusher, V.G., Tibshirani, R. y Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci U S A*, **98**, 5116-5121.

Uetz, P. y Finley, R.L., Jr. (2005). From protein networks to biological systems, *FEBS letters*, **579**, 1821-1827.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. y Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, **403**, 623-627.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. y Bork, P. (2007). STRING 7--recent developments in the integration and prediction of protein interactions, *Nucleic acids research*, **35**, D358-362.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. y Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Watson, J.D. y Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature*, **171**, 737-738.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Jr., Marks, J.R. y Nevins, J.R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles, *Proc Natl Acad Sci U S A*, **98**,

11462-11467.

Wu, Z. y Irizarry, R.A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays, *J Comput Biol*, **12**, 882-893.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. y Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic acids research*, **30**, 303-305.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabasi, A.L., Tavernier, J., Hill, D.E. y Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network, *Science*, **322**, 104-110.