TESIS DOCTORAL

DISEÑO Y DESARROLLO DE UNA PLATAFORMA BIOINFORMÁTICA PARA LA INTEGRACIÓN, GESTIÓN Y VISUALIZACIÓN DE REDES DE INTERACCIÓN DE PROTEÍNAS E INTERACTOMAS

DIEGO ALONSO LÓPEZ

DIRECTORES

DR. JAVIER DE LAS RIVAS SANZ

DR. RODRIGO SANTAMARÍA VICENTE

SALAMANCA, MARZO DE 2017

El Dr. Javier De Las Rivas Sanz, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Instituto de Biología Molecular y Celular del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Instituto y la Universidad de Salamanca (USAL).

Y el Dr. Rodrigo Santamaría Vicente, con D.N.I. 70879303L, Profesor Titular del Departamento de Informática y Automática de la Universidad de Salamanca (USAL), miembro del grupo de investigación VisUsal (Visualización de información y analítica visual) de dicha universidad y del grupo de investigación en Dinámica del Genoma y Epigenética del Instituto de Biología Funcional y Genómica (CSIC-USAL).

CERTIFICAN

que han dirigido esta Tesis Doctoral titulada "DISEÑO Y DESARROLLO DE UNA PLATAFORMA BIOINFORMÁTICA PARA LA INTEGRACIÓN, GESTIÓN Y VISUALIZACIÓN DE REDES DE INTERACCIÓN DE PROTEÍNAS E INTERACTOMAS" realizada por D. Diego Alonso López, alumno del programa de doctorado de Ingeniería Informática de la Universidad de Salamanca.

Y AUTORIZAN

la presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 17 de marzo de 2017

Dr. Javier De Las Rivas Sanz

Dr. Rodrigo Santamaría Vicente

AGRADECIMIENTOS

A mis directores de tesis, Javier De Las Rivas y Rodrigo Santamaría, gracias por darme la oportunidad de realizar este trabajo y facilitarme el camino con vuestros consejos y ayuda. Gracias, Javier, por ayudarme a crecer personal y profesionalmente.

A Henning Hermjakob y Sandra Orchard, por darme la oportunidad de visitar su grupo en el *European Bioinformatics Institute* (EMBL-EBI) y de colaborar con ellos a través de la organización HUPO PSI-MI. A Marine Dumousseau, por ser mi anfitriona en Cambridge y descubrirme los entresijos de JAMI.

A los miembros del grupo de Bioinformática y Genómica Funcional en el Instituto de Biología Molecular y Celular del Cáncer (CIC-IBMCC), por compartir sus conocimientos conmigo contribuyendo a mi formación en el campo de la Bioinformática y la Biología Computacional.

A todos los investigadores del CIC-IBMCC con los que he trabajado en distintos proyectos que me han enriquecido profesional y científicamente.

A Miguel Ángel, por su inestimable ayuda en el desarrollo de la aplicación y por su valiosa amistad.

A Sofía y Jorge

TABLA DE CONTENIDO

1 INTRODUCCIÓN Y ESTADO DEL ARTE

La identificación de los elementos que componen un sistema celular, como piezas de un gran puzle, ha sido el objetivo principal de innumerables estudios científicos en la historia reciente de la biología molecular. La secuenciación completa del genoma humano hace ya más de una década **(Human Genome Sequencing Consortium 2004; Venter et al. 2001)** supuso un hito científico histórico, pero también sirvió de piedra de toque para comprobar que la complejidad de la naturaleza biológica es abrumadora. De hecho, a día de hoy, numerosos grupos de investigación siguen trabajando para descifrar y entender el genoma humano **(ENCODE Project Consortium 2012)**. El gran reto científico del genoma es un ejemplo paradigmático de lo que hoy supone la investigación en biología molecular: partir de la identificación de todas las piezas de un sistema con la ayuda de tecnologías de gran escala, para continuar con la identificación de las relaciones entre dichas piezas como modo de descubrir grupos asociados y generar hipótesis específicas sobre la función o funciones biológicas en las que están implicados. Por ello, el mapeo completo de todos los elementos constitutivos y sus relaciones es clave para logar el primer descifrado de los sistemas biológicos a escala global (es decir a escala "ómica") y para avanzar en el conocimiento sobre su funcionamiento.

Este ciclo de generación de nuevo conocimiento tiene lugar hoy en día en diversos ámbitos ligados a tecnologías específicas, tales como la genómica, la proteómica o la metabolómica, cada una centrada en un aspecto de la biología molecular, pero todas siguiendo una misma estrategia: el aproximamiento global a una realidad compleja. Esta perspectiva global se materializa en forma de experimentos masivos (*high throughput*) que generan como resultado ingentes cantidades de datos. Es aquí cuando la biología requiere de la bioinformática como herramienta esencial para manejar, analizar e interpretar estos datos: solo a través de procesos analíticos el investigador podrá obtener información útil, a partir de

los datos brutos, que le permita hacer asunciones coherentes con el marco biológico del experimento. La bioinformática es, por tanto, una disciplina científica surgida al calor de la biología y las técnicas experimentales modernas que permiten medir, de forma cada vez más exhaustiva y global, numerosos procesos a nivel celular.

Este aumento exponencial de la información experimental generada por la comunidad científica ha supuesto un reto también para las bases de datos públicas de información biológica, que se han visto obligadas a redimensionar sus infraestructuras y adaptarse a nuevos tipos de datos **(Leinonen et al. 2011; Kodama et al. 2012)**. Hoy en día, la comunidad científica produce, mantiene y explota decenas de bases de datos donde se catalogan todos los elementos conocidos correspondientes a diferentes entidades biológicas, como por ejemplo la base de datos UniProt **(The UniProt Consortium 2014)**, referencia mundial en lo que a proteínas se refiere.

Aunque estos catálogos de entidades biológicas son esenciales para el avance en el estudio de los sistemas celulares, es evidente que un enfoque reduccionista, donde el todo se define como cada una de sus partes, no es suficiente para entender la biología celular, sino que es necesario estudiar en profundidad las relaciones entre todos los componentes del sistema. Las proteínas, como elemento funcional principal en el estudio de los procesos biológicos que ocurren en la célula, son uno de los mejores ejemplos de sistema asociativo que a nivel global es mucho más amplio y complejo que la suma de sus partes **(Aloy and Russell 2006)**.

El estudio de la fisiología celular nos ha enseñado que las proteínas trabajan habitualmente en grupo - ya sea de forma puntual en un determinado proceso o de forma estable constituyendo un complejo proteico - y que de esta manera regulan la mayoría de los procesos biológicos **(Griffiths et al. 2002)**.

Es por esto que el estudio exhaustivo de todas las posibles interacciones entre proteínas que pueden tener lugar en un organismo vivo es hoy uno de los objetivos principales de la comunidad científica. Al mapa completo que forman estas interacciones se le conoce como *interactoma*, concepto que hoy tiene importancia similar a la que empezó a tener el *genoma* hace ya más de dos décadas.

El interactoma de un organismo es, en definitiva, la red de contactos global entre todas sus proteínas. Se trata, por tanto, de un claro ejemplo de conjunto de datos masivo sobre el que el investigador necesita cierta penetración o inspiración (*insight*). Sería imposible obtener una idea aproximada de la interacción entre las proteínas de un organismo a partir de una tabla de valores: es necesario un análisis adecuado, apoyado por herramientas de visualización, para explorar y comprender la información expresada por dicho conjunto de interacciones. Para conseguir esto debemos recurrir a la teoría de grafos y la visualización de redes ya que no solo necesitamos una correcta visualización para explorar los datos, sino que además esperamos que esta nos proporcione nuevo conocimiento a través del uso de diferentes técnicas estadísticas sobre la red representada **(Cline et al. 2007)**.

El trabajo de investigación que se expone en esta Tesis Doctoral se centra precisamente en ese ámbito, las interacciones entre proteínas y la definición global de los conjuntos de interacciones presentes en cada organismo, o interactomas, en forma de redes biomoleculares. Partiendo de las diferentes bases de datos públicas sobre interacciones entre proteínas se construye un sistema de integración de dichas interacciones y se generan interactomas con diferentes niveles de calidad en función del soporte experimental de las interacciones que contienen. Toda esta información se pone a disposición de la comunidad científica a través de una aplicación diseñada a tal efecto que, entre otras cosas, posibilita la visualización y anotación funcional de las redes de interacción generadas por el propio investigador.

1.1 Interacciones proteína-proteína

Conviene definir la idea de interacción entre proteínas antes de seguir avanzando en los diferentes conceptos relacionados con el estudio de estas. Tal y como se describe en **(De Las Rivas and Fontanillo 2010)**, se entiende por interacción el contacto de dos proteínas a nivel molecular (a través de una interfaz de interacción con enlaces específicos) que puede darse dentro o fuera de la célula y que normalmente sucede en un contexto biológico concreto. Este tipo de interacciones proteína-proteína, a las que denominaremos "PPIs", son siempre de contacto físico y pueden ser de dos subtipos: **(i)** *físicas directas*, cuando la proteína A se une a la proteína B por contacto directo (por ejemplo, una citoquina a su receptor específico: IGF e IGFR); o **(ii)** *físicas indirectas*, cuando la proteína A está unida a la proteína B por contacto indirecto formando parte de una estructura multi-proteína, donde sólo algunas de ellas entran en contacto directo, pero todas se ensamblan en un único complejo macromolecular convirtiéndose en subunidades del mismo (por ejemplo en el ribosoma, que está constituido por más de 50 proteínas, pero no todas se tocan directamente). Frente a este tipo de interacciones físicas, hay muchas otras formas posibles de definir o encontrar relaciones o asociaciones entre proteínas; pero es importante discernir y no mezclar conceptos de modo que si se encuentra, por ejemplo, una "asociación funcional" entre dos proteínas por participar en la misma vía metabólica esto no implica que ambas entren en contacto y no se consideran propiamente PPIs **(Chatr-aryamontri et al. 2008) (Mackay et al. 2007)**.

Como se ha indicado, cualquier contacto que se considere una interacción debe ser específico y para ello hay que tener en cuenta que estos también se pueden dar en los procesos de construcción, plegamiento o degradación de las proteínas ya que la mayoría de estas en algún momento de su ciclo de vida interactuarán, por ejemplo, con chaperonas o con proteínas pertenecientes a la maquinaría de degradación (como el proteosoma) **(De Las Rivas and Fontanillo 2010)**.

Por otra parte, el contacto que se produce en una interacción puede ser permanente o temporal. Un ejemplo de contacto permanente es el de las interacciones presentes en cualquier complejo macromolecular, como el complejo ATP sintasa (**Figura 1**), una enzima

transmembrana que cataliza la síntesis de ATP a partir de ADP y que está formada por ocho proteínas diferentes en mamíferos **(Yoshida et al. 2001)**. Como ejemplo de contacto temporal transitorio (*transient interacction*) resulta válida cualquier asociación de proteínas formada puntualmente para llevar a cabo una función concreta, como por ejemplo la activación de la expresión de un gen mediante la unión de factores de trascripción y activadores a la región de DNA promotora de dicho gen.

Finalmente, como ya se ha mencionado, otro aspecto muy importante a tener en cuenta en las interacciones de proteínas es el contexto biológico en el que estas tienen lugar. No todas las posibles interacciones ocurren en cualquier célula y en cualquier momento: que una interacción se produzca dependerá de factores como el tipo celular, la fase del ciclo celular, el estado de diferenciación, las condiciones del medio, la presencia de cofactores, etc.

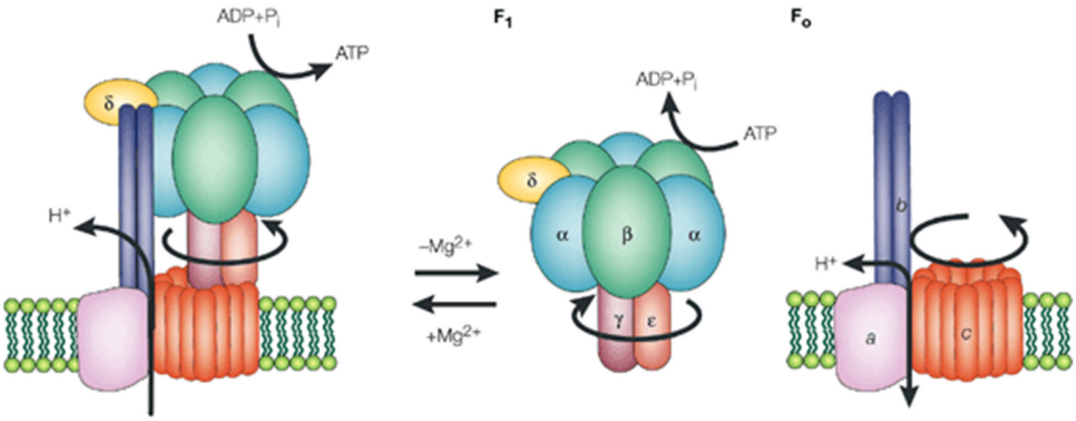


Figura 1. *Representación gráfica de la estructura del complejo ATP sintasa, formado a su vez por los complejos F0 y F1, y ensamblado con un total de 8 proteínas diferentes en el caso de los mamíferos. Extraída de* **(Yoshida et al. 2001)**.

1.2 Métodos experimentales de detección de interacciones proteína-proteína

La determinación experimental de las posibles interacciones entre proteínas es una tarea tan compleja como necesaria para la comprensión de la biología celular. Existen varias estrategias para llevar a cabo dichos experimentos, pero pueden englobarse en dos grandes grupos: **(i)** las técnicas que miden interacciones físicas entre dos proteínas únicas o interacciones binarias (*binary methods*) y **(ii)** las técnicas que miden todas las interacciones entre un grupo específico de proteínas, también conocidas como técnicas co-complejo (*co-complex methods*) **(Yu et al. 2008)**.

Los métodos binarios son técnicas experimentales cuyo objetivo es la identificación de parejas de proteínas interactuantes. Entre ellos, el de uso más común es el experimento denominado análisis de doble híbrido (Y2H).

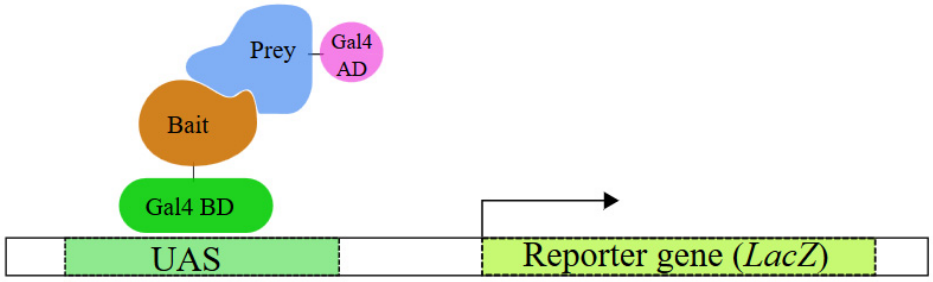


Figura 2. *Ilustración esquemática del funcionamiento del análisis de doble híbrido donde se representan los dominios de enlace (BD) y de activación (AD) unidos a las proteínas cebo y presa respectivamente (Extraído de Wikimedia Commons bajo licencia CC BY-SA 3.0).*

El análisis de doble híbrido está basado en la organización modular de algunos factores de transcripción eucarióticos, que son capaces de mantener su función aunque estén separados en fragmentos y estos se unan de forma indirecta.

Esta característica se aprovecha para dividirlos en dos fragmentos: **(i)** un dominio de enlace (*binding domain o BD)*, que reconoce la secuencia regulatoria de activación (*UAS*) y **(ii)** un dominio activador (*activating domain* o *AD)* que activa la maquinaria de transcripción. A estos dominios se fusionan las dos proteínas que se quieren analizar y que normalmente son denominadas *cebo (bait)* en el caso de aquella que se une al DNA a través del dominio BD y *presa (prey)* en el caso de la que se une al dominio activador AD. De esta manera, la trascripción del gen *delator (reporter)* se llevará a cabo solo si las dos proteínas estudiadas forman un complejo **(Young 1998)**.

Los métodos de tipo *co-complex* están más orientados a la búsqueda de complejos y miden las relaciones tanto directas como indirectas entre varias proteínas. Entre ellos, cabe destacar el análisis de afinidad de purificación seguido de espectrometría de masas (TAP-MS) y los métodos de co-inmunoprecipitación (CoIP). Las técnicas de tipo TAP-MS se basan en etiquetar una proteína con un marcador molecular, a la que se suele denominar *bait*, para después poder aislarla mediante técnicas *pull-down* junto a sus interactores, normalmente denominados *prey* **(Berggård et al. 2007)**. Por su parte, los métodos de co-inmunoprecipitación están basados en el uso de anticuerpos: la proteína de interés se aísla mediante el uso de un anticuerpo específico y las moléculas que interaccionan con ella se identifican después mediante la técnica *western blot* **(Mackay et al. 2007)**.

La diferencia fundamental entre las dos grandes aproximaciones experimentales para la detección de interacciones, métodos binarios y *co-complex*, reside en la detección de relaciones indirectas cuando se usan estos últimos.

Los resultados experimentales obtenidos mediante técnicas de tipo *co-complex* no pueden ser interpretados directamente de forma binaria ya que describen una interacción múltiple pero no especifican las interacciones binarias a partir de las cuales se forma dicho complejo.

Para poder interpretar este resultado, y extraer a partir de él las interacciones en formato binario, debe aplicarse un algoritmo de expansión que convierta la interacción múltiple en varias interacciones simples. Básicamente, existen dos alternativas en este tipo de algoritmos, el modelo matriz, donde se interpreta que todas las proteínas interactúan de forma binaria con el resto, o el modelo *spoke*, donde se asume que la proteína de interés (*bait*) interactúa con todas las demás (*prey*). Como se explicará en la sección de resultados, el método desarrollado en este trabajo de investigación utiliza esta segunda aproximación ya que genera menos falsos positivos **(Hakes et al. 2007)**.

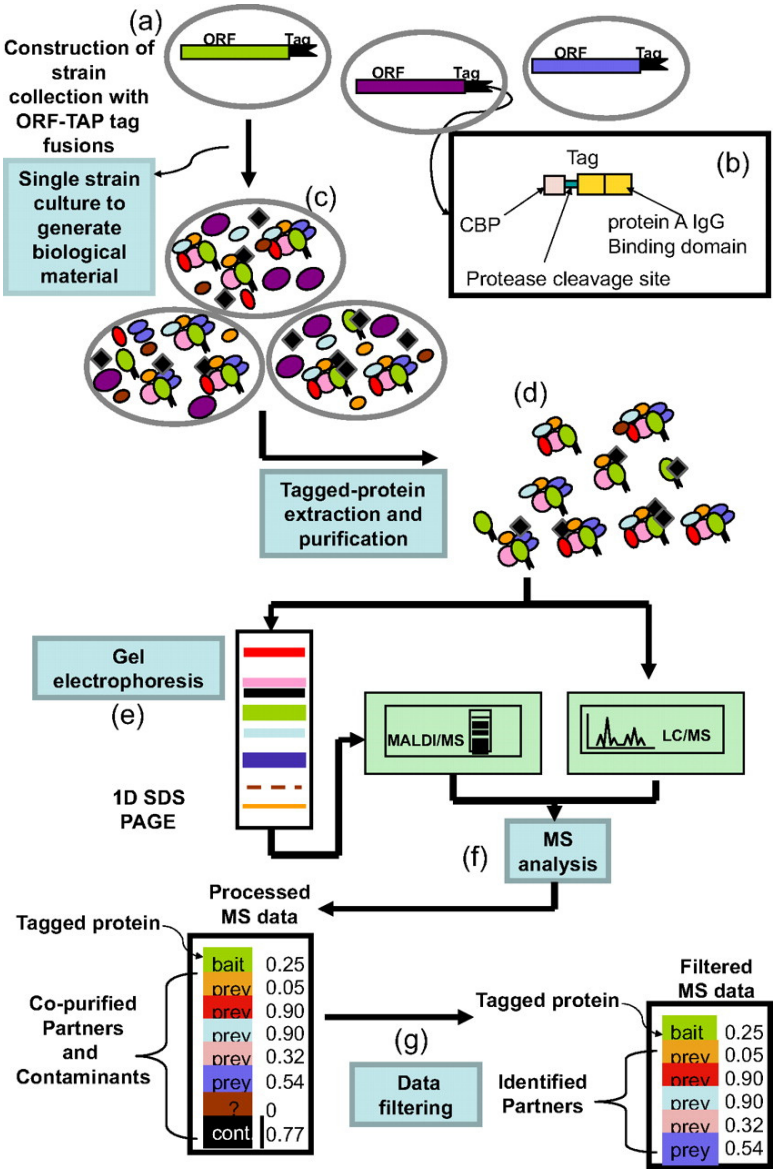


Figura 3. *Flujo de trabajo esquematizado de las técnicas de tipo TAP-MS a partir de las cuales se obtienen como resultado interacciones múltiples de tipo co-complex. Extraído de* **(Wodak et al. 2009)**.

1.2.1 Métodos computacionales de predicción de interacciones proteína-proteína

Aunque este trabajo se centra en los datos generados a partir de métodos experimentales de detección de interacciones de proteínas, existen también diferentes aproximaciones computacionales orientadas a la predicción de posibles interacciones entre proteínas. Una buena revisión al respecto fue publicada por Valencia y Pazos en 2008 **(Valencia and Pazos 2008)**. Algunos de estos métodos, brevemente enunciados, son: **(i)** predicción de interacción por similitud entre los árboles filogenéticos de proteínas **(Pazos and Valencia 2001)**, **(ii)** predicción de interacción por identificación de mutaciones correlacionadas entre parejas de proteínas **(Pazos et al. 1997)**, **(iii)** predicción de interacción basada en estructuras 3D por exploración de *docking* físico entre parejas de proteínas **(Wass et al. 2011)** y **(iv)** predicción de interacción por conservación de genes vecinos en genomas y por eventos de fusión génica **(Dandekar et al. 1998; Enright et al. 1999)**.

En estas aproximaciones se aplican diferentes métodos computacionales que, a partir de postulados teóricos sobre el conocimiento experimental existente, construyen modelos estadísticos orientados a la predicción de nuevas interacciones. Entre el tipo de técnicas computacionales aplicadas caben destacar los métodos bayesianos **(Jansen et al. 2003)** o las técnicas de tipo *Support Vector Machines (SVM)* **(Ben-Hur and Noble 2005)**. Normalmente, este tipo de aproximaciones predictivas se pueden usan para tratar de elaborar grandes mapas de interacciones, o incluso interactomas completos para determinadas especies como la predicción global hecha para el interactomas humano en **(Zhang et al. 2012)**.

1.3 Representación de datos y estándares

El almacenamiento y uso de la información es clave en la generación de nuevo conocimiento. Se trata, además, de un proceso iterativo donde dicho nuevo conocimiento debe ser descrito y almacenado para pasar a formar parte de la base sobre la que se construirán las próximas hipótesis.

Los datos experimentales sobre interacciones de proteínas, al igual que cualquier otro experimento llevado a cabo por un grupo de investigación, son puestos a disposición del resto de investigadores a través de la literatura científica. Durante la última década, la comunidad científica ha trabajado intensamente, no solo en la identificación masiva de interacciones biomoleculares, sino también en asegurar la calidad de los datos obtenidos y organizar su almacenamiento en bases de datos **(Kerrien et al. 2007; del-Toro et al. 2013; Orchard et al. 2012)**.

La iniciativa de estándares en proteómica (PSI, *Protein Standards Initiative*) (1), perteneciente a la organización del proteoma humano (HUPO, *Human Proteome Organization*) (2), es la institución encargada de establecer los diferentes estándares en el área de la proteómica. Dentro de ella existen diferentes grupos de trabajo entre los que se encuentra el que se ocupa de las interacciones moleculares (MI, *Molecular Interactions*).

HUPO PSI-MI es, por tanto, un grupo de trabajo perteneciente a una organización científica abierta a la comunidad y especializado en interacciones moleculares. HUPO establece, de forma general, cuatro objetivos fundamentales referidos en cada grupo de trabajo a su ámbito de conocimiento particular:

1. Establecer conjuntos de especificaciones o estándares en proteómica con la información mínima necesaria para describir un resultado experimental en el dominio de conocimiento correspondiente a cada grupo.
2. Diseñar un formato de datos que permita el intercambio de este tipo de resultados entre múltiples plataformas y grupos.
3. Definir vocabularios controlados y ontologías aplicables a datos proteómicos.
4. Dar soporte a la comunidad científica para la adopción y el uso de estos estándares.

En el caso específico del grupo PSI-MI, actualmente existen una serie de especificaciones mínimas para experimentos de interacción molecular denominadas MIMIx **(Orchard et al. 2007)**, varios formatos de intercambio de datos con vocabularios controlados **(Hermjakob et al. 2004)** y algunas herramientas para el uso de dichos estándares por parte de bases de datos públicas y otros proyectos científicos **(Kerrien et al. 2007; Aranda et al. 2011)**. Además, el grupo PSI-MI creó en 2005 un consorcio internacional denominado IMEx **(Orchard et al. 2012)** (3) con el objetivo de establecer una serie de políticas generales para los procesos de extracción de información sobre interacciones de proteínas a partir de la literatura científica.

A continuación, se describen los proyectos del grupo HUPO PSI-MI que más importancia han tenido para el desarrollo de este trabajo de investigación.

1.3.1 Formatos de intercambio de datos

Uno de los estándares más importantes establecido por el grupo PSI-MI es el formato informático para el intercambio de datos de interacciones de proteínas. Este formato tiene dos versiones, una más sencilla con ficheros de texto tabulados y otra más completa con ficheros en formato XML **(Kerrien et al. 2007)**.

El formato tabulado, PSI-MI MITAB, se encuentra actualmente en su versión 2.7 y define el uso de archivos de texto con 42 columnas para diferentes tipos de datos (especificaciones completas en (4)). Este formato de texto plano tabulado tiene la ventaja de que puede leerse sin necesidad de software especializado y a su vez es sencillo de procesar de forma automatizada. Sin embargo, no es capaz de representar información compleja como, por ejemplo, las interacciones múltiples.

Para representar información más compleja se creó el formato PSI-MI XML. Esta especificación, basada en el uso del formato XML, cuenta con un esquema XSD (*XML Schema Definition*) completo que contiene diferentes tipos de etiquetas que permiten modelar gran cantidad de datos relacionados con los experimentos de detección de interacciones moleculares y la interpretación de sus resultados (especificaciones completas en (5)). Actualmente, el formato PSI-MI XML es ampliamente usado en su versión 2.5 pero, en los

últimos años, el grupo de trabajo ha estado definiendo las nuevas especificaciones 3.0, que permitirán modelar aún más información sobre determinados fenómenos biológicos. Estas especificaciones, en cuya definición se ha colaborado durante este trabajo de investigación, se publicarán oficialmente en los próximos meses **(Dumousseau et al. 2017)**.

La existencia de un formato estándar para el almacenamiento e intercambio de datos de interacción de proteínas es esencial para la comunidad científica. Gracias a esto, un investigador puede descargar información de diferentes bases de datos o proyectos y analizarla en otros contextos mediante el uso de herramientas software de visualización y análisis de redes. Este formato es también el punto de partida para proyectos como el desarrollado en este trabajo de investigación, pues permite establecer una política de mínimos entre las diferentes bases de datos procesadas para después tratar de integrar todos los datos recopilados.

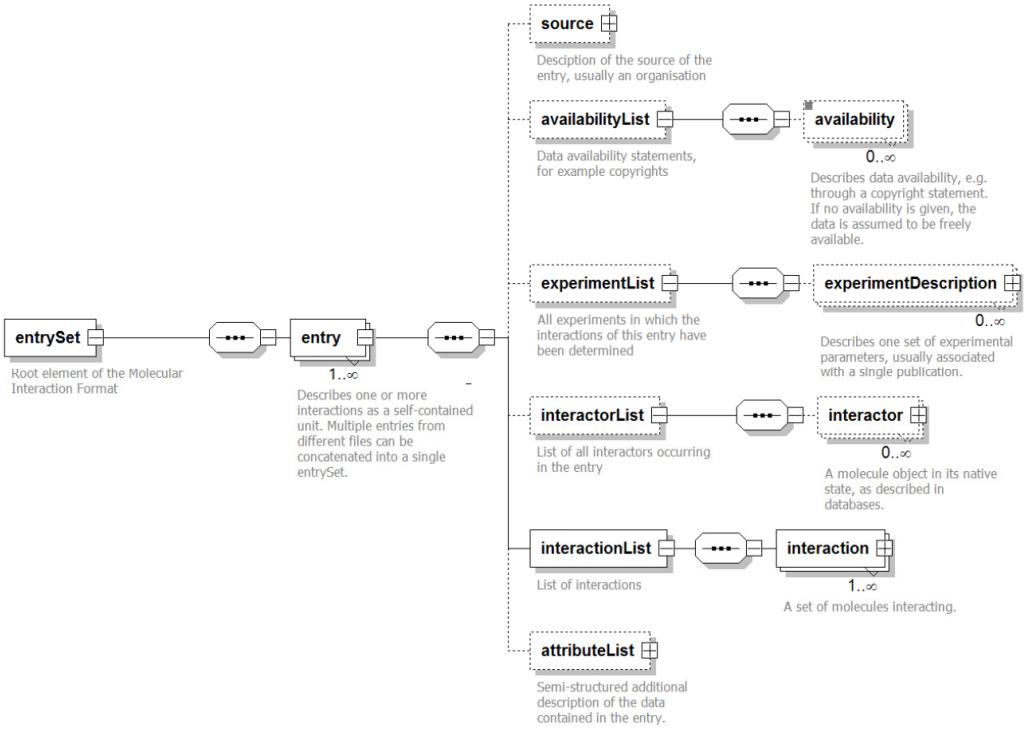


Figura 4. *Representación gráfica de la estructura del formato PSI-MI XML 2.5 donde pueden verse los tres objetos fundamentales: las listas de experimentos, interactores e interacciones. Extraído de* **(Kerrien et al. 2007)**.

1.3.2 Ontologías y vocabularios controlados

La representación de la información relativa a un experimento de detección de interacciones moleculares requiere de una estructura adecuada y común que viene definida por los formatos antes mencionados. Pero esto no es suficiente ya que también debe existir consenso en cuanto al posible contenido de algunas de las entidades definidas en dichas estructuras. Esto es, no solo se necesita establecer una etiqueta XML o una columna para, por ejemplo, la identificación del método de detección de una interacción, sino que también debe

establecerse qué contenido puede almacenarse en ella y con qué texto literal debe especificarse. Esto es importante para evitar ambigüedades relacionadas con el uso de sinónimos o problemas derivados de posibles errores a la hora de insertar los datos en texto libre. Dichos problemas son los que tratan de resolver las ontologías y sus vocabularios controlados.

Una ontología es un vocabulario específico acerca de un dominio de conocimiento donde se establecen términos y relaciones entre estos para representar dicho conocimiento. Estos vocabularios son utilizados en los formatos PSI-MI para estandarizar el significado de los diferentes objetos de datos que contienen las estructuras previamente definidas. El objetivo es proporcionar a cualquier investigador, o técnico que extrae información de la literatura, un conjunto de términos cuyo significado tenga siempre la misma interpretación.

Estos términos, además, están conectados a través de una estructura jerárquica de tal manera que un término general se desglosa en diferentes términos con mayor nivel de especificidad. Esto permite al usuario elegir el nivel de detalle adecuado en cada momento y anotar con términos de mayor o menor nivel de profundidad.

Los vocabularios controlados se usan, por ejemplo, para definir el tipo de participantes en una interacción o para especificar el método de detección de esta. El mantenimiento de las ontologías es llevado a cabo por el grupo PSI-MI que, a lo largo de las diferentes versiones del formato de intercambio de datos, ha ido añadiendo y modificando términos y relaciones en los diferentes vocabularios controlados.

Estas ontologías se almacenan y mantienen en un formato de ficheros denominado OBO (*Open Biomedical Ontologies*) **(Osumi-Sutherland 2010; Smith et al. 2007)**. Existe también un servicio web denominado OLS (*Ontology Lookup Service*) (6) que permite la búsqueda de términos en todas estas ontologías, así como la visualización de los árboles que las relaciones entre estos determinan **(Côté et al. 2008; Côté et al. 2010)**.

1.3.3 PSICQUIC

El último de los objetivos establecidos por HUPO para sus grupos de trabajo está orientado a dar soporte a la comunidad científica para la adopción y el uso de los estándares propuestos por cada uno de ellos. En el caso del grupo PSI-MI, además de las especificaciones y formatos ya mencionados, existe un proyecto específico para facilitar el acceso a los datos de interacciones moleculares, contenidos en diferentes bases de datos, de forma automatizada.

Este proyecto se denomina *Proteomics Standard Initiative Common QUery InterfaCe* (PSICQUIC) y básicamente proporciona un servicio web con un listado de métodos bien definidos y un lenguaje de consulta, denominado MIQL (*Molecular Interactions Query Language*), que permite al investigador especificar formalmente el conjunto de datos con el que han de operan dichos métodos **(Aranda et al. 2011; del-Toro et al. 2013)**.

Este proyecto trabaja con los datos de interacciones disponibles en diversas bases de datos, entre las que se encuentra la desarrollada en esta Tesis Doctoral, y permite obtener los

resultados disponibles en todas aquellas bases de datos que el investigador especifique. Esto hace que, desde cualquier software compatible con la interfaz de PSICQUIC, se pueda acceder a dichos datos tal y como se muestra en **Figura 5**.

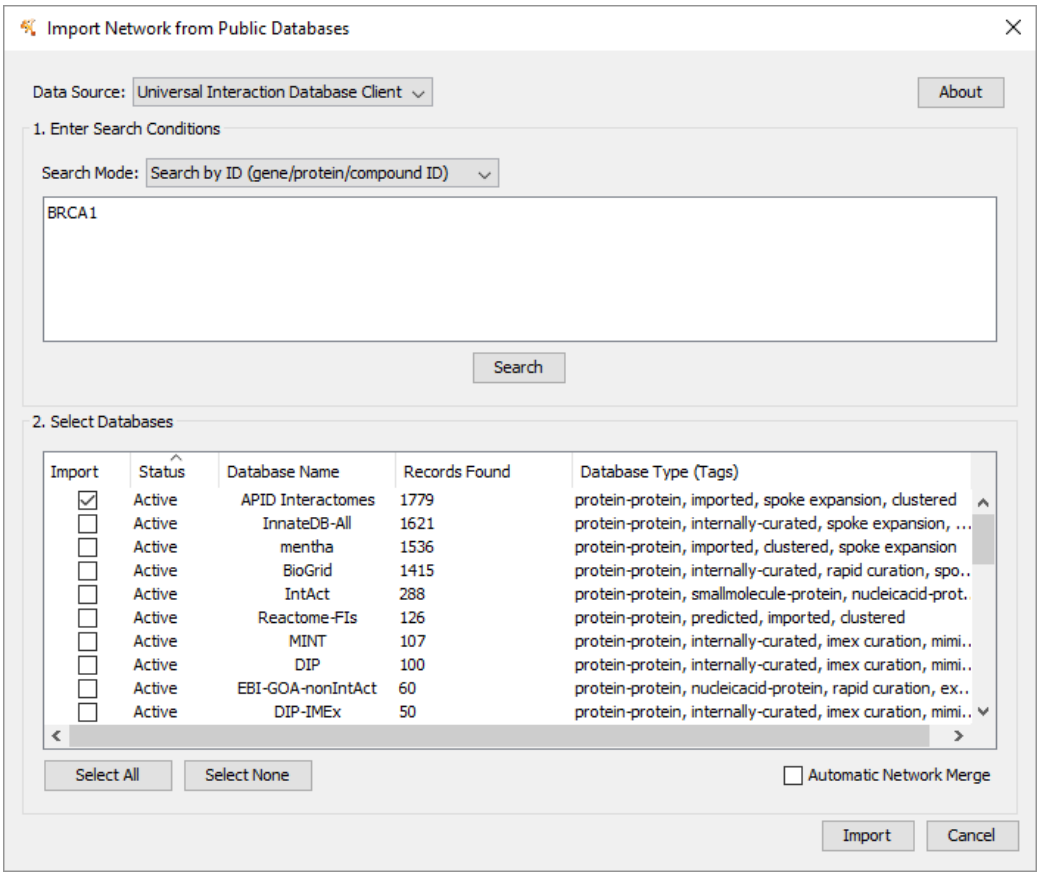


***Figura 5**. Interfaz de búsqueda de interacciones a través de los servicios web de PSICQUIC que ofrece el software Cytoscape* **(Shannon et al. 2003)**. *Se puede observar que la base de datos construida en esta Tesis Doctoral (denominada APID Interactomes, como se explicará en el próximo capítulo) es la que más información contiene para la proteína BRCA1.*

Los resultados que PSICQUIC ofrece al investigador se generan a partir de ficheros de tipo PSI-MI TAB, que cada base de datos genera expresamente para su uso en PSICQUIC. Este mismo formato es en el que el investigador obtiene los resultados de su consulta, lo cual tiene una serie de ventajas relacionadas con su simplicidad, pero presenta también ciertas limitaciones tal y como se mencionó anteriormente. Es por esto que, como se explicará después, los métodos propuestos en esta Tesis Doctoral usan como origen de datos los ficheros originales en formato PSI-MI XML de cada base de datos primaria.

1.3.4 Framework JAMI

Otro de los proyectos relacionados con el objetivo de facilitar el uso de los estándares PSI-MI es la librería de desarrollo o *framework* JAMI (*Java Framework for Molecular Interactions*). Inicialmente se desarrollaron librerías JAVA independientes para leer los ficheros PSI-MI TAB y XML pero, con el paso del tiempo, cada vez se hizo más patente la necesidad de una

plataforma general para la gestión de los diferentes procesos relacionados con la información que estos ficheros almacenan.

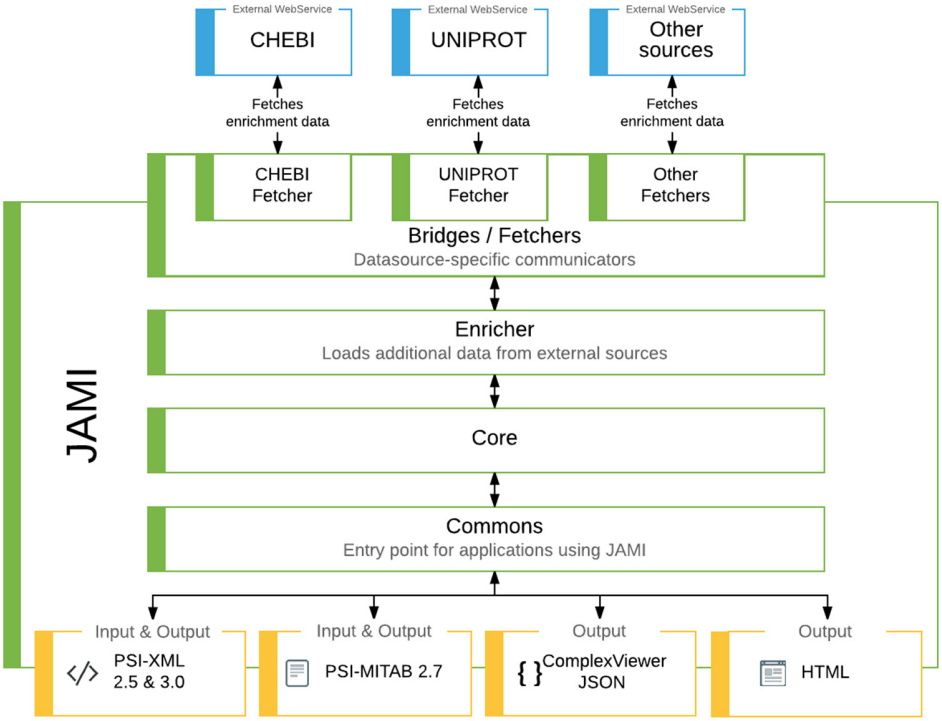


Figura 6. *Esquema de la arquitectura interna de la plataforma JAMI donde se muestra su naturaleza modular. Extraído de* **(Dumousseau et al. 2017)**

El proyecto JAMI fue impulsado por el grupo HUPO PSI-MI con el objetivo de proporcionar una plataforma de desarrollo única que permitiera el uso de los formatos y vocabularios controlados generados por el mismo grupo de trabajo, aislando así al programador de la complejidad de dichos formatos y dotándole de diferentes módulos con utilidades relacionadas con este tipo de datos.

Hoy en día dicha plataforma se mantiene y desarrolla a través de una colaboración entre el EMBL-EBI y la Universidad de Cambridge, en la que durante este trabajo de investigación se ha participado activamente gracias al hecho de ser el primer y único proyecto hasta la fecha que integra datos de tipo PSI-MI XML a través del *framework* JAMI. Los resultados de esta colaboración, que sigue vigente en la actualidad, están en proceso de publicación **(Dumousseau et al. 2017)**.

1.4 Bases de datos de referencia sobre proteínas

Cualquier interacción entre proteínas, al margen de toda la información experimental que pueda aportar, siempre deberá definirse en su forma más esencial como el conjunto (o pareja

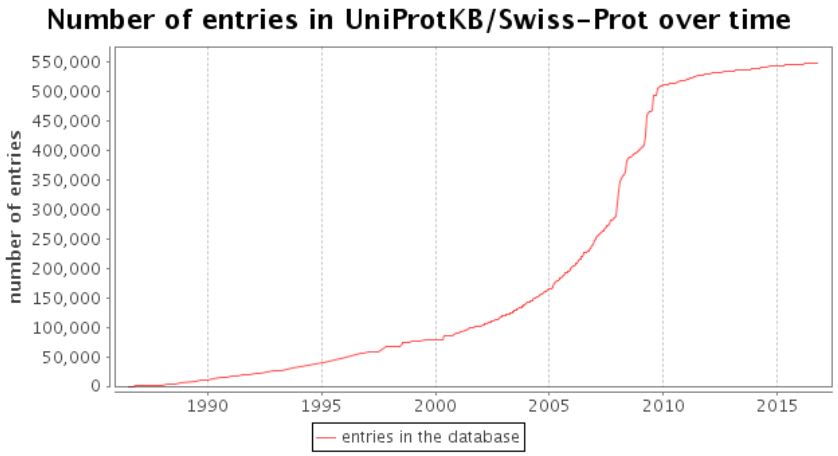
si es binaria) de proteínas que la componen. Esto es, una interacción entre dos proteínas estará identificada de forma única si, y solo si, las dos proteínas que interactúan lo están.

Esto, que a priori parece un requisito sencillo, es quizás uno de los mayores retos a la hora de integrar datos de interacciones provenientes de diferentes bases de datos: la identificación no ambigua de cada interactor.

El uso de una base de datos de referencia sobre proteínas es, por tanto, un factor clave para conseguir identificar de manera inequívoca a los participantes de una interacción. La base de datos pública sobre proteínas que se ha usado en este trabajo de investigación como referencia es UniProt **(The UniProt Consortium 2014)** (7).

UniProt (*The Universal Protein Resource)* es un repositorio que alberga diferentes bases de datos con información sobre proteínas. Recibe miles de secuencias de proteínas generadas por la comunidad científica y, a partir de estas, construye diferentes conjuntos de datos.

Las bases de datos más importantes son las contenidas en lo que se denomina UniProtKB (*Protein Knowledgebase*): **(i)** TrEMBL, una base de datos primaria procedente de la anotación automática de secuencias de proteínas y **(ii)** Swiss-Prot, una base de datos secundaria anotada por expertos. La diferencia entre ambas bases de datos radica fundamentalmente en el grado de anotación de sus registros y, por tanto, en la fiabilidad de los datos que contienen. En **Figura 7** puede observarse que, aunque el crecimiento en número de registros ha sido exponencial hasta recientemente en ambas bases de datos, existe una diferencia significativa en cuanto a su tamaño total ya que SwissProt solo contiene proteínas bien anotadas gracias a la revisión manual de sus registros por expertos. Cabe destacar también la eliminación masiva de registros de TrEMBL en marzo de 2015 debido a que se tomaron medidas para reducir la redundancia en los registros que esta contenía (8).



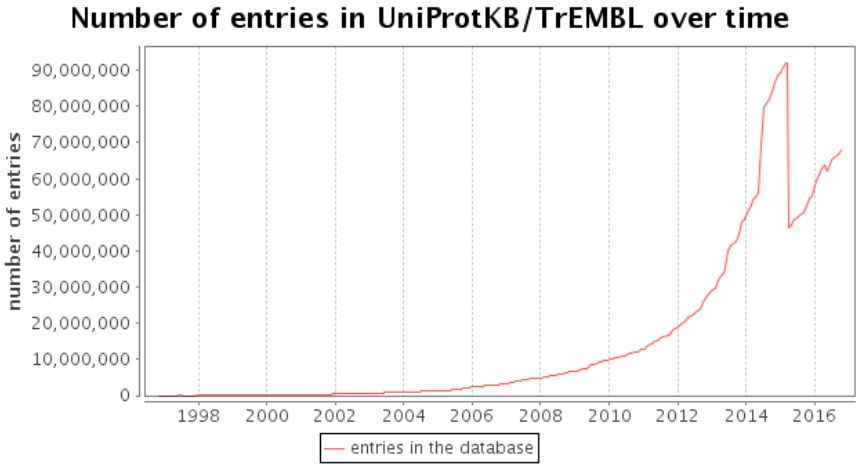


Figura 7. *Evolución del número de registros a lo largo del tiempo en las bases de datos de UniprotKB. Se puede observar el crecimiento exponencial de estas hasta su estabilización, así como la eliminación masiva de registros en TrEMBL en el año 2015 para reducir la redundancia que presentaban los registros contenidos en esta base de datos.*

En la estrategia seguida durante este trabajo de investigación a la hora de identificar y unificar interacciones, UniProtKB juega un papel esencial puesto que solo se procesan aquellas proteínas que puedan ser identificadas en alguna de sus dos bases de datos.

Otra base de datos de UniProt importante para este trabajo es UniProt Proteomes, donde se catalogan los proteomas de diferentes organismos definiendo todas las proteínas de UniProtKB que estos contienen y dándoles un identificador único. Esta base de datos se utilizaró para los cálculos de cobertura de los interactomas generados en esta Tesis Doctoral.

1.5 Bases de datos de interacciones de proteínas

La difusión de nuevo conocimiento entre la comunidad científica está basada en la publicación de los resultados experimentales en la literatura. Algunos de estos resultados, además, son almacenados y clasificados para su mejor aprovechamiento en diferentes bases de datos. Este es el caso de las interacciones de proteínas, para las que existen completos sistemas de información que proporcionan a los investigadores un entorno sobre el que trabajar y generar nuevas hipótesis.

Estos sistemas de información contienen tanto bases de datos primarias con información extraída de la literatura como meta-bases de datos que se construyen a partir estas y cuyo objetivo es el de aglutinar y, en algunos casos, reorganizar los datos disponibles en varios de estos repositorios primarios. Es en este segundo contexto donde se ubica el sistema de información diseñado en este trabajo de doctorado.

1.5.1 Bases de datos primarias

Las bases de datos primarias obtienen su información directamente desde la literatura científica gracias al trabajo de recopilación e interpretación llevado a cabo por expertos

denominados *curators.* Estos expertos siguen protocolos comunes de anotación tales como los que se establecen en **(Orchard et al. 2012)** o **(The UniProt Consortium 2014)**. Los investigadores, por su parte, son emplazados a seguir una serie de recomendaciones a la hora de describir sus resultados experimentales **(Orchard et al. 2007)**.

A continuación, se detallan brevemente las bases de datos más importantes de interacciones de proteínas descritas mediante el uso de métodos experimentales.

1. **IntAct (Kerrien et al. 2012)** (9): Base de datos creada y mantenida por el grupo de interacciones moleculares del Instituto Europeo de Bioinformática (EBI-EMBL, Cambridge). Se trata de un completo sistema de información que incluye datos detallados sobre miles de interacciones, así como un potente sistema de consultas vía web. Almacena datos de interacciones proteína-proteína, pero también de proteínas y otras pequeñas moléculas. El equipo de desarrollo forma parte del grupo PSI-MI que, como se mencionó anteriormente, se encarga de establecer estándares y fomentar el uso de estos. Sus datos sobre interacciones están integrados en otros sistemas de información biológica, como por ejemplo UniProt, y son miembros del consorcio IMEx.
2. **MINT (Licata et al. 2012)** (10): La base de datos MINT se dedica también a la extracción de datos experimentales sobre interacciones de proteínas a partir de la literatura científica. Forma parte del consorcio IMEx y, desde finales del año 2013, utiliza la infraestructura tecnológica de IntAct para poner sus datos a disposición de los investigadores **(Orchard et al. 2014)**. Así, los datos sobre interacciones proteicas *curados* por el grupo de expertos de MINT pueden consultarse a través de la página web de IntAct, donde aparecen integrados en su base de datos junto a las interacciones registradas por la propia IntAct, pero etiquetados como procedentes de MINT.
3. **HPRD (Keshava Prasad et al. 2009)** (11): La base de datos HPRD (*Human Protein Reference Database*) está especializada en proteínas humanas. Su objetivo es ofrecer una plataforma web con información general sobre dominios estructurales, modificaciones postraduccionales, interacciones y asociación con patologías para las diferentes proteínas que forman el proteoma humano. HPRD, por tanto, no es una base de datos centrada en las interacciones de las proteínas humanas, sino que contiene esta información como una anotación más entre otras muchas. Esto provoca que dichas interacciones estén pobremente anotadas, muchas veces sin identificadores UniProt para las proteínas y con una descripción del método de detección con poco detalle y sin usar ontologías. No forma parte de IMEx aunque sí ofrece sus datos en formato HUPO PSI-MI XML.
4. **DIP (Salwinski et al. 2004)** (12)**:** DIP (*Database of Interacting Proteins*) almacena también datos sobre interacciones proteína-proteína extraídos de la literatura científica. Forma parte de IMEx y, al igual que las bases de datos anteriores, ofrece la descarga de todos sus datos en ficheros con formato HUPO PSI-MI XML.

5. **BioGRID (Chatr-Aryamontri et al. 2015)** (13)**:** La plataforma web de BioGRID (*Biological General Repository for Interaction Datasets*) ofrece datos experimentales de interacciones entre proteínas pero también interacciones genéticas **(Mani et al. 2008)**. Forma parte de IMEx pero solo en calidad de *observer* tal y como se define en la web de dicho consorcio (3).
6. **BioPlex (Huttlin et al. 2015)** (14)**:** BioPlex (Biophysical Interactions of Orfeome-based Complexes) no es una base de datos como las anteriores sino un proyecto orientado a la detección de todas las interacciones existentes en el *orfeoma* (conjunto de todos los *open reading frames* o ORFs) humano **(Wiemann et al. 2016)**. Con la publicación inicial proporcionaron un conjunto de datos con aproximadamente 23.000 interacciones y, desde ese momento, en su página web ofrecen los nuevos conjuntos de datos que van obteniendo. Estas colecciones de interacciones se pueden descargar en ficheros de texto tabulados y no cumplen con ningún formato estándar, aunque son procesadas por el algoritmo desarrollado en este trabajo debido a su naturaleza experimental.

1.5.2 Meta-bases de datos

Existen también otras bases de datos que, al igual que la desarrollada durante este trabajo de doctorado, recopilan e integran la información sobre interacciones disponible en las bases de datos primarias. Estas bases de datos son conocidas como meta-bases de datos ya que agrupan y, en algunos casos, procesan e integran los datos disponibles en varias bases de datos primarias. De esta manera se convierten en algo parecido a una base de datos de bases de datos.

Este proceso de recopilación y procesado de datos puede realizarse con diferentes estrategias y objetivos. A continuación, se describen algunos de los proyectos más importantes.

1.5.2.1 Mentha

Mentha **(Calderone et al. 2013)** (15) es una meta-base de datos que, al igual que la desarrollada en esta Tesis Doctoral, contiene solo datos de interacciones descritas experimentalmente. Dispone de un sistema web que facilita la búsqueda de interacciones y el trabajo con estas gracias a un visor de redes que se ejecuta en la máquina del cliente con el uso de la plataforma JAVA.

Contiene datos de varias especies y ofrece información detallada sobra cada interacción manteniendo además el enlace a los datos originales. Establece también un *score* para cada interacción basado en el número de evidencias que la describen experimentalmente.

Mentha también almacena diversas anotaciones para las proteínas, obtenidas principalmente de las bases de datos *Gene Ontology, KEGG* y *UniProt.*

En cuanto a la integración de datos, los creadores de Mentha diseñaron un proceso automatizado para conseguir actualizar semanalmente su base de datos. Para ello utilizan la

plataforma PSICQUIC, procesando 5 de las 36 bases de datos presentes en ella: MINT, IntAct, DIP, MatrixDB **(Chautard et al. 2011)** (16) y BioGRID.

Por último, Mentha está orientado al trabajo con conjuntos de interacciones y, aunque dispondría de la información necesaria para hacerlo, no ofrece interactomas.

En resumen, Mentha es una meta-base de datos que ofrece una alta frecuencia de actualización a cambio de trabajar solo con los datos disponibles en PSICQUIC y proporciona a los investigadores un interfaz web completo con información detallada sobre cada interacción.

1.5.2.2 iRefWeb

iRefWeb **(Turner et al. 2010)** (17) es una meta-base de datos que contiene interacciones de las siguientes bases de datos primarias: BIND **(Bader et al. 2003)** (13), BioGRID, CORUM **(Ruepp et al. 2010)** (18), DIP, IntAct, HPRD, MINT, MPact **(Güldener et al. 2006)**, MPPI **(Mewes et al. 2011)** (19) y OPHID **(Brown and Jurisica 2005)** (20).

Esta meta-base de datos integra todo tipo de interacciones, experimentales y funcionales (genéticas), ofreciendo una potente interfaz de búsqueda y filtrado en su página web, tal y como se muestra en **Figura 8**. Como resultado a la búsqueda formulada por el investigador proporciona una tabla con información detallada sobre la interacción y los registros originales de las bases de datos primarias que la describen. Además, para cada interacción, presenta un *score* basado en los cálculos utilizados originalmente por la base de datos primaria MINT y descritos en **(Ceol et al. 2009)**.

Como proyecto paralelo, el equipo de trabajo de iRefWeb desarrolló un método para tratar de unificar los identificadores de todas las interacciones disponibles en las bases de datos primarias. Para ello, proponen el uso de un algoritmo que genera claves a partir de datos específicos como las secuencias de las proteínas interactuantes o sus identificadores taxonómicos **(Razick et al. 2008)**.

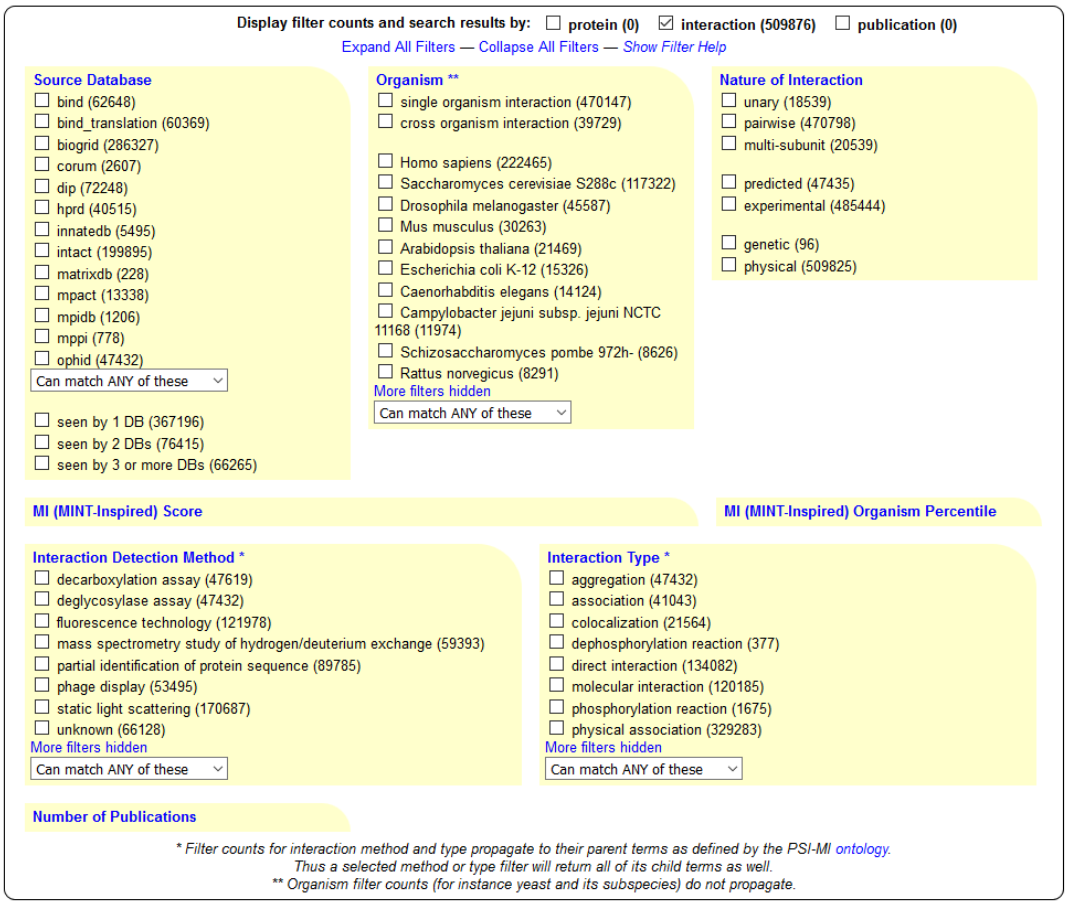


Figura 8. *Sistema de filtrado avanzado disponible en el servidor web de IrefWeb. Permite el uso de múltiples criterios de filtrado de forma combinada y actualiza los resultados numéricos correspondientes a cada categoría en tiempo real.*

1.5.2.3 Hint

HINT (High-quality INTeractomes) **(Das and Yu 2012)** (21) es una base de datos con un interfaz web sencillo que almacena interacciones de proteínas de 11 organismos modelo (inicialmente eran solo 3) y ofrece la posibilidad de consultar hasta un máximo de 10 proteínas mostrando como resultado una tabla con información resumida sobre estas.

Aunque se trata de un proyecto no comparable a otros de los que aquí se mencionan, cabe destacar que es el único que ofrece, al igual que la plataforma bioinformática propuesta en esta Tesis Doctoral, la posibilidad de obtener interactomas completos desde su aplicación web. Además, los clasifica según el tipo de interacciones que contienen, siendo estas de tipo: **(i)** binario, **(ii)** procedentes de experimentos de tipo co-complex, **(iii)** de análisis sistemáticos masivos (*high-throughput*) o **(iv)** de experimentos a pequeña escala obtenidos a partir de la literatura.

1.5.2.4 STRING

La base de datos *STRING* **(von Mering et al. 2003)** (22), de acceso público solo para uso académico, surge a partir de un consorcio formado por el *Swiss Institute of Bioinformatics* (SIB), el *Center for Protein Research* (NNF-CPR) y el *European Molecular Biology Laboratory* (EMBL). Se trata de un ambicioso proyecto cuyo objetivo principal es describir de forma exhaustiva todas las interacciones posibles entre proteínas. Para ello no solo almacenan interacciones probadas experimentalmente, sino que también predicen todo tipo de interacciones funcionales entre proteínas.

Tal y como se describe en **(Szklarczyk et al. 2016)**, estas interacciones funcionales son inferidas a partir de: **(i)** análisis sistemáticos de co-expresión de genes, **(ii)** detección de señales selectivas compartidas entre diferentes genomas, **(iii)** minería de texto automática de la literatura científica y **(iv)** extrapolación de interacciones entre organismos basada en genes ortólogos.

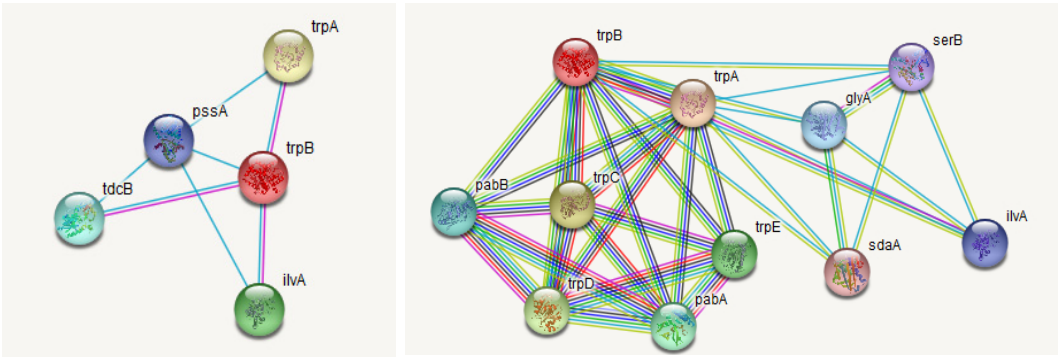


Figura 9. *Red de interacción obtenida en STRING para la proteína trpB (tryptophan synthase, beta subunit) de Escherichia coli. En el recuadro de la izquierda aparecen solo las interacciones experimentales, mientras que en el recuadro de la derecha aparecen también las interacciones funcionales inferidas por los diferentes métodos de predicción computacional que STRING usa.*

El servidor web de STRING ofrece también, en su última versión (finales de 2016), un visor de redes integrado en el navegador web y varias herramientas de análisis relacionadas con los diferentes métodos de predicción que utiliza.

1.5.2.5 GeneMania

GeneMania **(Warde-Farley et al. 2010)** (23) es una base de datos orientada a la construcción de redes de asociación de genes que, entre otros muchos datos, almacena interacciones de proteínas.

Su servidor web ofrece, en su última versión, un visualizador de redes integrado totalmente en el navegador que proporciona información sobre las asociaciones de genes/proteínas obtenidas a partir de: **(i)** interacciones físicas, **(ii)** co-expresión de genes, **(iii)** predicciones computacionales, **(iv)** rutas de señalización, **(v)** co-localización, **(vi)** interacciones genéticas y **(vi)** dominios de proteínas comunes. Estos datos están disponibles para genes y proteínas de siete organismos diferentes **(Zuberi et al. 2013)**.

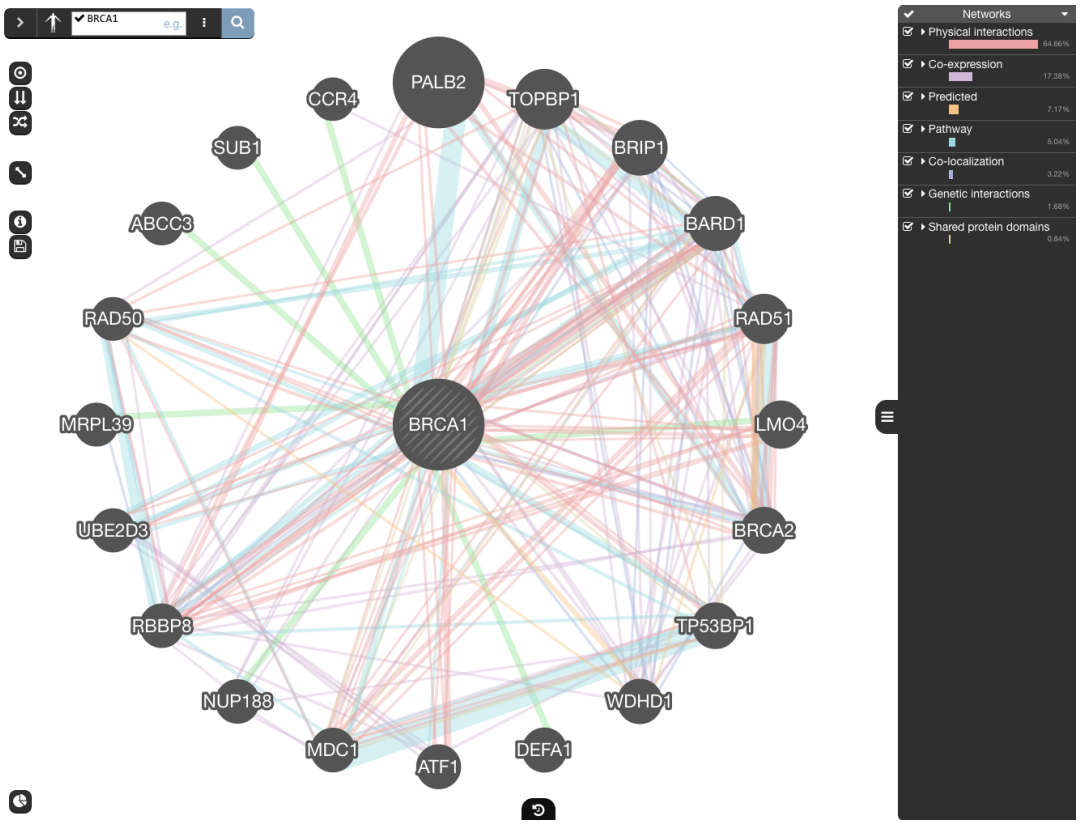


Figura 10. *Interfaz web de la base de datos de GeneMania. Toda la información disponible se muestra a través del visualizador de redes mediante el uso de cuadros de diálogo auxiliares con diferentes tipos de datos.*

1.5.2.6 Otros proyectos

Existen otras meta-bases de datos de interacciones con objetivos diferentes o que ya no se actualizan, tales como:

1. **Hippie**, *Human Integrated Protein-Protein Interaction rEference* **(Alanis-Lobato et al. 2016; Schaefer et al. 2012)** (24): Orientada a la creación e interpretación de redes de interacción entre proteínas humanas
2. **UniHI**, *Unified Human Interactome* **(Kalathur et al. 2014)** (25): Incluye información sobre drogas, genes y proteínas humanas permitiendo crear redes con las interacciones entre estas tres entidades.
3. **PINA2**, *Protein Interaction Network Analysis platform* **(Cowley et al. 2012)** (26): Incluye interacciones para 6 organismos diferentes a partir de bases de datos primarias. Última actualización en 2014.
4. **HitPredict (Patil et al. 2011; López et al. 2015)** (27): Recopila interacciones para varios organismos. Última actualización en 2015.

1.6 Bases de datos de estructuras 3D de proteínas

El estudio de la estructura tridimensional de las proteínas es clave para entender su función. Existen diversos métodos para descifrar la estructura de una proteína, desde la cristalografía de rayos X **(Omar et al. 2016)** hasta la espectroscopía de resonancia magnética nuclear **(Volkman et al. 1998)** o la microscopía electrónica **(Kostyuchenko et al. 2003)**.

Las estructuras descritas por la comunidad científica se almacenan en la base de datos *RSCB Protein Data Bank* (PDB) **(Berman et al. 2000)** (28), donde se pueden consultar todos los detalles de estas e incluso descargarse en ficheros con formato PDB.

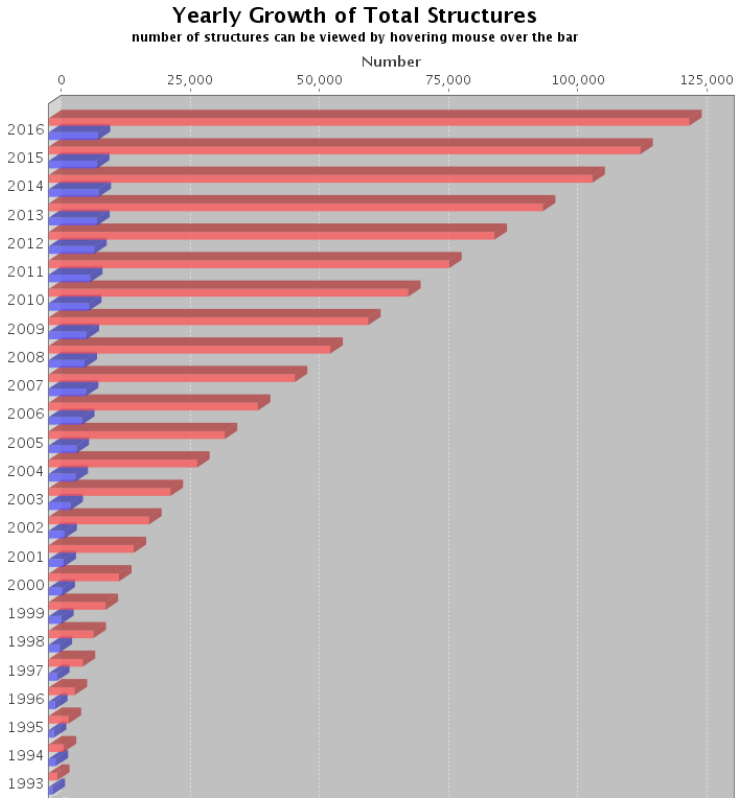


Figura 11. *Evolución temporal del número de estructuras tridimensionales almacenadas en la base de datos PDB. Las barras azules indican el incremento anual y las rojas el total. Puede observarse el carácter exponencial de dicho incremento hasta la actualidad.*

El número de estructuras tridimensionales descritas por los diferentes grupos de investigación ha aumentado de forma significativa en la última década tal y como se muestra en **Figura 11**. Esto, unido al hecho de que en gran parte de estas estructuras aparecen complejos de proteínas interaccionando, motivó la inclusión de este tipo de datos en el protocolo de integración de interacciones desarrollado en este trabajo de investigación.

Existe además otro proyecto, denominado PDBsum **(De Beer et al. 2014)** (29), que interpreta las estructuras disponibles en la base de datos PDB proporcionando a la comunidad científica información esquemática sobre las diferentes moléculas (proteínas y otras) que contiene

cada estructura y las interacciones fisicoquímicas presentes entre estas. Es esta interpretación de estructuras en concreto la que se utiliza en este trabajo para determinar si dos proteínas, cuya interacción ya está registrada, están interaccionando también en una estructura tridimensional descrita en la base de datos PDB. En **Figura 12** puede verse un ejemplo de estructura tridimensional incorporada a la base de datos generada en este trabajo como evidencia adicional para una interacción ya existente, en este caso la de las proteínas humanas HRAS y SOS1.

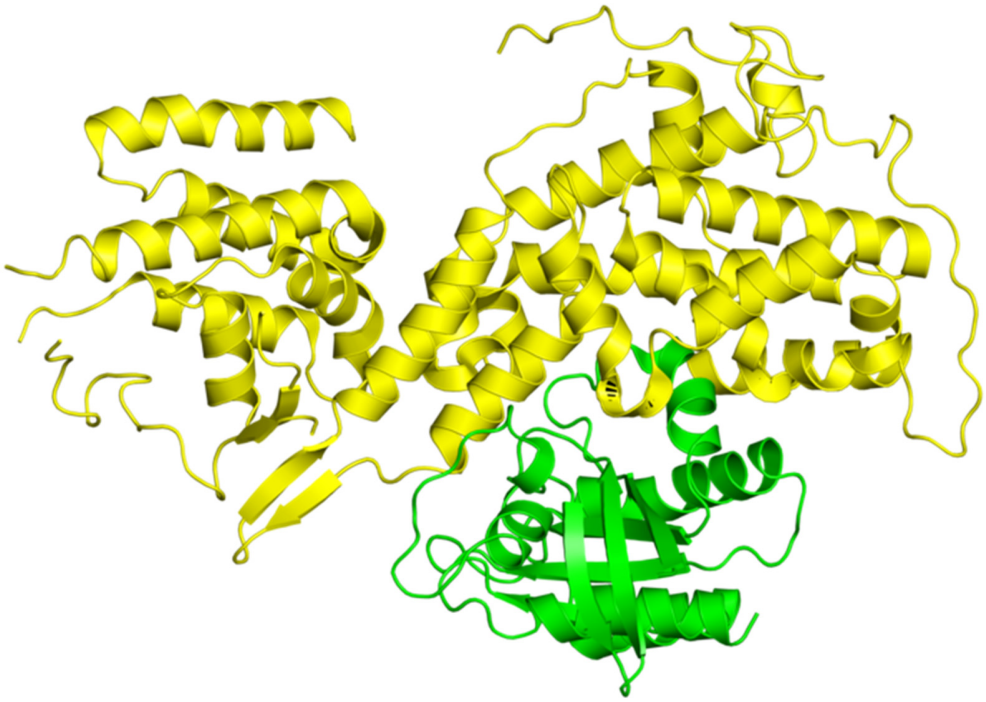


Figura 12. *Estructura tridimensional del complejo formado por las proteínas humanas HRAS (verde) y SOS1 (amarillo). Esta estructura, descrita en* **(Boriack-Sjodin et al. 1998)***, está almacenada en la base de datos PDB (con identificador 1BKD) y durante este trabajo ha sido procesada y asociada al registro de la interacción HRAS-SOS1, que aparece descrita también por otros métodos experimentales.*

1.7 Bases de datos de anotación

La caracterización funcional y estructural de las proteínas presentes en una interacción es necesaria para estudiar las consecuencias de dicha asociación. El conocimiento generado por la comunidad científica sobre miles de proteínas sujetas a estudio está contenido en la literatura, pero también estructurado formalmente a través de diferentes espacios de anotaciones.

Estas anotaciones normalmente proceden de ontologías diseñadas para capturar el conocimiento en ámbitos concretos como, por ejemplo, la localización celular o las rutas de señalización. En esta Tesis Doctoral se han utilizado varias de estas ontologías para anotar las

proteínas que han sido descritas en alguna de las interacciones que su base de datos almacena.

1.7.1 Ontologías funcionales

Las anotaciones funcionales utilizadas en este trabajo provienen del proyecto *Gene Ontology* **(Ashburner et al. 2000; Blake et al. 2015)** (30), uno de los sistemas de anotación más utilizados por la comunidad científica.

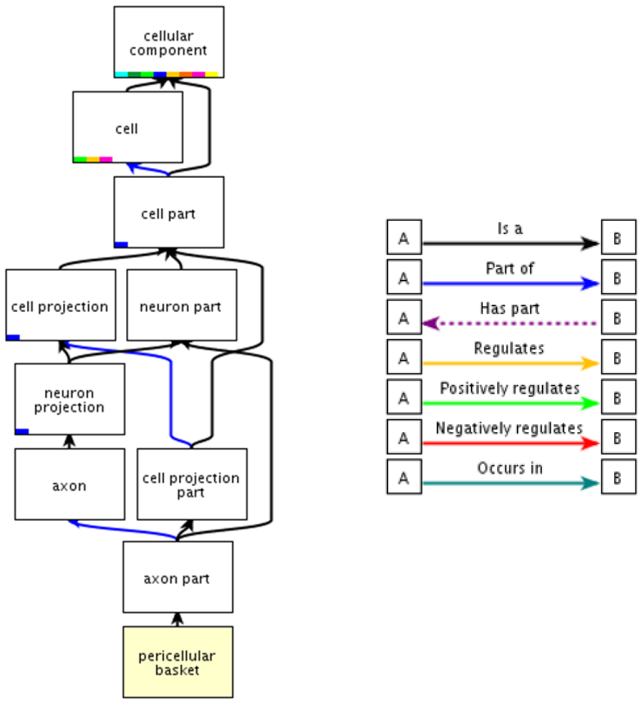


Figura 13. *Antecesores del término "pericellular basket" de la ontología cellular component de GO. Puede observarse la estructura en forma de grafo dirigido acíclico y la leyenda con el tipo de relaciones posibles entre términos. Figura generada con la herramienta AmiGO* **(Carbon et al. 2009)** *(31).*

Este proyecto, iniciado en 1988 y conocido popularmente como *GO,* proporciona un vocabulario controlado para describir una determinada entidad biológica (normalmente un gen o una proteína) anotándola con los términos adecuados en cada caso. Dicho vocabulario controlado se organiza en tres ontologías diferentes:

1. **Biological Process**: Vocabulario relacionado con la función biológica en la que una entidad está involucrada.
2. **Molecular Function**: Etiquetas referentes a la función molecular de un producto génico o proteína.
3. **Cellular Component**: Asignaciones relacionadas con la ubicación de la entidad etiquetada en un determinado componente celular.

Los términos contenidos en cada una de estas ontologías se componen de un identificador único, un nombre y una descripción. Además, estos términos están relacionados entre sí

mediante una estructura semijerárquica, en forma de grafo dirigido acíclico tal y como se muestra en **Figura 13**, y están diseñados para representar el conocimiento de forma general sin asociarlo específicamente a ningún organismo.

1.7.2 Vías metabólicas y de señalización

Otra anotación de especial interés para las proteínas es su posible implicación en determinados procesos moleculares. Estos procesos se modelan a través de rutas metabólicas y de señalización donde cada molécula participante tiene una función definida.

Existen varias bases de datos orientadas a la catalogación de estas rutas, siendo las más utilizadas KEGG (*Kyoto Encyclopedia of Genes and Genomes*) **(Ogata et al. 1999)** (32) y Reactome **(Croft et al. 2014)** (33).

Esta última es la que UniProt utiliza como anotación para sus proteínas y la que también se ha utilizado en esta Tesis Doctoral. Reactome es una base de datos primaria que extrae las rutas a partir de la literatura científica y proporciona a los investigadores diferentes herramientas para la visualización y el análisis de estas rutas o *pathways*.

1.7.3 Familias y dominios proteicos

En la secuencia de las proteínas normalmente se pueden distinguir diferentes regiones conservadas a las que se denomina dominios **(Wetlaufer 1973)**. Estos dominios suelen corresponderse con las regiones en las que se haya una mayor densidad debido a los plegamientos en su estructura y tienen un impacto directo tanto en dicha estructura como en la función de la proteína. Las diferentes combinaciones de estos dominios dan lugar a la gran diversidad de proteínas existentes en la naturaleza.

Por otro lado, las proteínas se agrupan en familias cuando descienden de un antepasado común **(Kunin et al. 2003)**. Esto da lugar a conjuntos de proteínas ortólogas que, normalmente, presentan similitud en su secuencia, estructura y función.

Existen varias bases de datos cuyo objetivo es clasificar las proteínas conocidas mediante el uso de anotaciones relacionadas con familias y dominios. En este trabajo se han procesado los datos contenidos en dos de las más utilizadas por la comunidad científica: *Pfam* e *InterPro*.

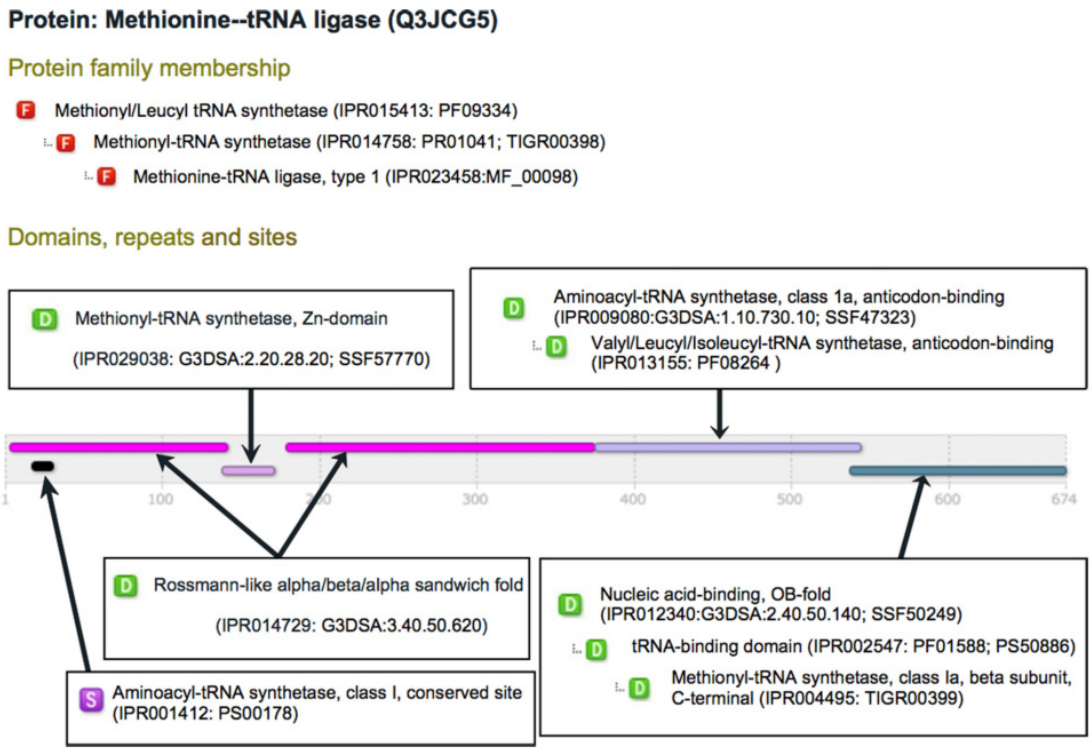


Figura 14. *Información disponible en InterPro para la proteína Metionina-ARNt Ligasa. Puede observarse la asignación de identificadores propios (con el formato "IPR...") en base a la información disponible en otras bases de datos primarias.*

Pfam **(Finn et al. 2015)** (34) es un repositorio público que mantiene una base de datos sobre familias de proteínas y dominios proteicos. Para ello utilizan alineamientos múltiples de secuencias y modelos ocultos de Markov (HMM).

InterPro **(Mitchell et al. 2015)** (35) es un meta-base de datos que integra datos sobre familias y predicciones de dominios y sitios funcionales a partir de la información disponible en diferentes bases de datos primarias.

1.8 Redes Biomoleculares

La mayoría de las funciones biológicas surgen de complejas interacciones entre diferentes componentes celulares como las proteínas, el DNA, el RNA, etc. Son, por tanto, sistemas complejos que muestran propiedades que no podemos caracterizar simplemente a través del conocimiento de sus partes.

Gracias a las nuevas tecnologías *high throughput*, cada vez se dispone de más información sobre las interacciones entre estos elementos - redes de interacción de proteínas, redes de rutas metabólicas, redes de señalización, redes reguladoras de transcripción, etc. - y se sabe que ninguna de estas redes funciona de forma independiente, sino que todas ellas forman una gran red de redes que es la verdadera responsable del comportamiento de la célula.

Uno de los grandes retos actuales de la investigación en biología molecular es, por tanto, el estudio de estas redes complejas, disciplina conocida como biología de sistemas. Y es precisamente en este contexto donde surge la conexión entre la biología molecular y las matemáticas: estos sistemas complejos se modelan sobre el papel a través de las redes y se estudian formalmente gracias a los desarrollos matemáticos provenientes de la teoría de grafos. Por otro lado, si atendemos a las diferentes definiciones disponibles en la literatura sobre visualización de información, el trabajo con grandes cantidades de datos agrupados en elementos interconectados en una red resulta un ejemplo paradigmático del tipo de problema que se pretende abordar. La visualización de información tiene que ver con el proceso de pasar de representaciones gráficas a percepciones humanas que maximicen la comprensión de la propia realidad. Es, en definitiva, una vía para adquirir nuevo conocimiento sobre dicha realidad **(Shneiderman 1996; Ware 2004)**.

En el mundo científico, la visualización se ha convertido en un método computacional más que nos ofrece la posibilidad de descubrir nuevas propiedades sobre el concepto estudiado. El tipo de visualización relacionada con el estudio de redes biológicas es la visualización de grafos. Un grafo está formado por un conjunto N de nodos o vértices y un conjunto A de aristas o arcos que relacionan pares de nodos, de forma direccional o no direccional. Esto no solo nos proporciona una correcta visualización para explorar los datos, sino que además nos ayuda a generar nuevo conocimiento a través del uso de diferentes técnicas estadísticas sobre la red representada que nos permiten estudiar su topología. Para ello se utilizan diferentes métricas pertenecientes al ámbito de la teoría de grafos, como por ejemplo el grado de un nodo, que mide el número de conexiones que este tiene con el resto de nodos.

Las redes biomoleculares, o redes biológicas, presentan una topología característica. No se organizan de forma aleatoria, sino que siguen un modelo al que se denomina *de escala libre* (*scale-free networks*), donde la mayoría de los nodos tienen pocas conexiones con el resto y un pequeño número de nodos aglutina la mayoría de las conexiones de la red **(Barabasi and Oltvai 2004)**. Este tipo de organización se descubrió inicialmente gracias al estudio de las conexiones entre páginas web a lo largo de la red *World Wide Web (WWW)* y después se comprobó que el mismo patrón aparecía en las redes biomoleculares **(Barabási and Albert 1999; Jeong et al. 2001; Jeong et al. 2000; Guelzim et al. 2002; Tong 2004)**.

En **Figura 15** se muestran las diferencias entre una red con topología aleatoria y otra con topología de escala libre. Se puede observar como la distribución del grado de los nodos de esta última no sigue un patrón de normalidad, sino que presenta una distribución de ley de potencias (*power-law*) que, al estar ambos ejes en escala logarítmica, toma la forma de una función polinómica de grado 1, es decir, una recta. Esto confirma que, mientras que un pequeño grupo de nodos tiene un alto grado, la mayoría de ellos presentan valores bajos para dicha métrica. A los nodos con alto grado se les denomina *hubs*. La ventaja de este tipo de organización podría residir en el hecho que las redes de escala libre son más tolerantes a la pérdida aleatoria de alguno de sus nodos **(Zhu et al. 2007)**.

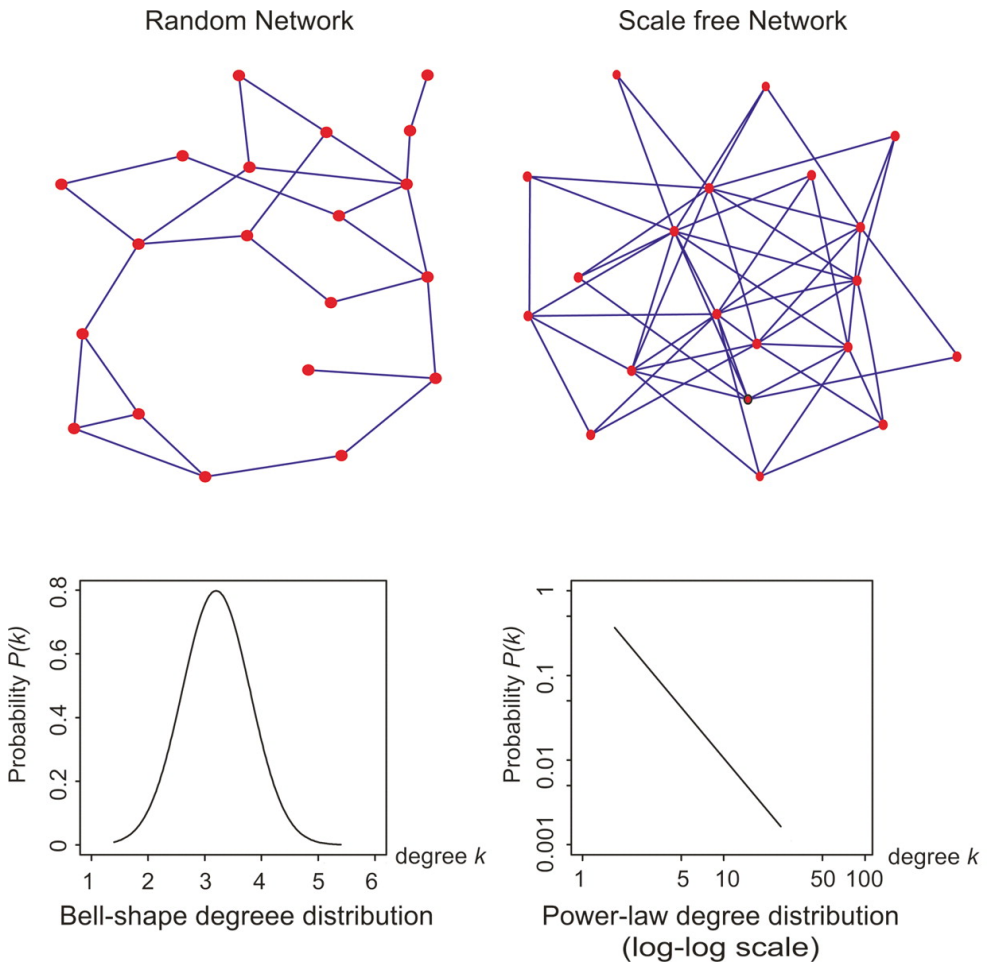


Figura 15. *Topologías de red aleatoria y de escala libre. Se muestra la representación de la red y la distribución del grado de sus nodos. Extraída de* **(Zhu et al. 2007)**.

Existen gran cantidad de herramientas software diseñadas para la visualización y el análisis de redes. Entre ellas, destaca Cytoscape **(Shannon et al. 2003)** (36), una plataforma abierta de visualización y análisis de redes desarrollada gracias a un consorcio formado por varias empresas e instituciones científicas tales como el Instituto Pasteur de Francia, el Sloan-Ketering Cancer Center de EEUU o la empresa Agilent Technologies. Como proyecto satélite de Cytoscape surge cytoscape.js **(Franz et al. 2015)**, una librería de programación que permite trasladar parte de la funcionalidad de Cytoscape al entorno de las aplicaciones web.

1.8.1 Redes de interacción de proteínas

Las redes de interacción de proteínas son un tipo concreto de red biomolecular donde lo que se representa son las interacciones binarias entre pares de proteínas para un conjunto más o menos amplio de estas. Así, se trata de grafos no direccionales donde los nodos representan proteínas y los arcos interacciones entre estas.

Estas redes son, probablemente, aquellas que mejor describen la arquitectura biológica puesto que la proteína es el componente biomolecular funcional por excelencia. Desde el

punto de vista experimental, constituye también uno 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 doble híbrido aplicada a levadura *Saccharomyces cerevisiae* **(Ito et al. 2001)**. Posteriormente se ha seguido aplicando esta técnica en otros organismos y se han desarrollado nuevos métodos de detección de interacciones de gran escala y alto rendimiento.

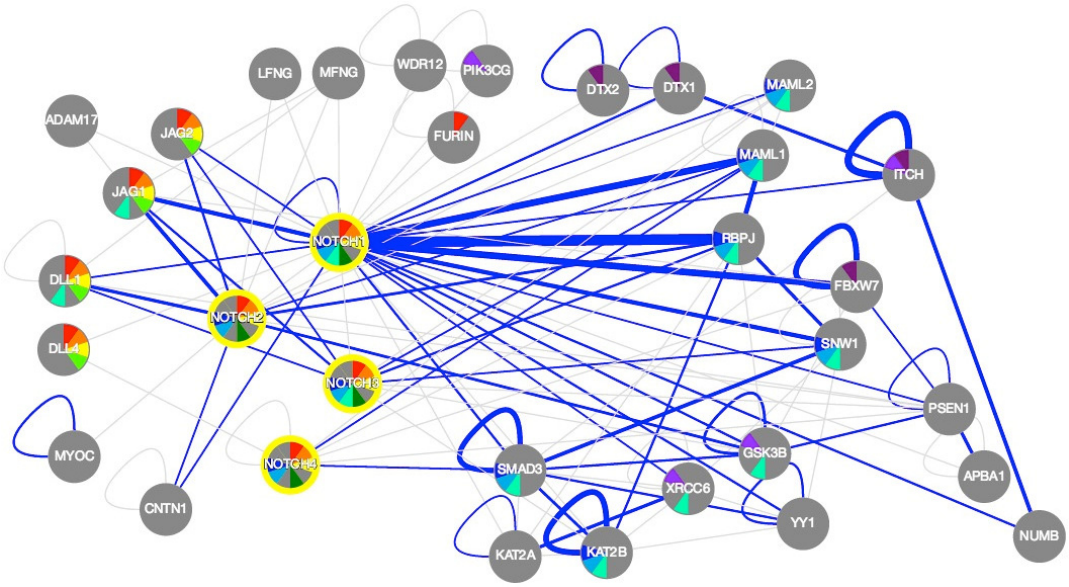


Figura 16. *Red de interacción de proteínas generada con la herramienta de visualización desarrollada como parte de este trabajo de investigación. En ella se puede observar el diferente grosor de los arcos para indicar mayor o menor evidencia experimental de cada interacción y el sistema de etiquetado por colores para las anotaciones funcionales. Ambas características se describirán con detalle a lo largo de esta Tesis Doctoral.*

2 HIPÓTESIS Y OBJETIVOS

2.1 Problema marco e hipótesis

En la biología molecular y biología celular actuales, propias de la era "ómica", es muy importante tratar de pasar de las aproximaciones enciclopédicas que hacen simples listados o elencos – más o menos globales – de los genes activos o proteínas de un organismo o de un sistema biológico, como unidades independientes, a los estudios que buscan encontrar las relaciones, conexiones o interacciones entre dichos elementos moleculares clave. Esto es especialmente importante respecto a las proteínas que construyen las máquinas moleculares dentro de todo sistema celular vivo. La generación de redes precisas de interacción entre proteínas es un paso esencial para lograr la reconstrucción de la arquitectura y dinámica molecular de las células. El mapeo preciso de todas las interacciones proteína-proteína conocidas para un determinado organismo, y derivadas de experimentos con tecnologías de pequeña o de gran escala, no es un asunto trivial y no está del todo bien resuelto actualmente en el ámbito de la bioinformática ya que no existen bases de datos completas que se centren en este objetivo específico. De hecho, un problema frecuente es la construcción y puesta a disposición de los usuarios de recursos bioinformáticos y bases de datos que mezclan muchos tipos de asociaciones entre genes y proteínas, algunos probados experimentalmente y muchos simplemente inferidos o predichos. De este modo, suele faltar un rigor procedural en el estado del arte respecto a la integración de interacciones entre proteínas y entre genes, y también respecto al valor biomolecular de las relaciones que se derivan de dichas integraciones y la confianza que se puede tener en ellas.

En este marco temático, nuestro trabajo se ha planteado abordar los problemas descritos y formular una serie de objetivos en torno a la siguiente hipótesis central.

Hipótesis central. *Es posible, utilizando y desarrollando métodos bioinformáticos y computacionales, recopilar de modo exhaustivo, para los distintos organismos, todas las interacciones proteína-proteína conocidas y demostradas experimentalmente, así como construir redes precisas de interacción entre proteínas que recapitulen la arquitectura molecular de las células y sus módulos biológicos funcionales.*

2.2 Objetivos Principales

Objetivo 1. Desarrollar un método de integración de datos sobre interacciones moleculares físicas entre proteínas (*protein-protein physical interactions*, PPIs) demostradas experimentalmente y descritas en publicaciones científicas referenciadas. Dentro de este método se diseñarán protocolos para recoger, unificar y manejar de modo eficaz y transparente la información sobre los métodos experimentales (bioquímicos, biofísicos, proteómicos, etc.) que demuestran cada interacción molecular, así como sobre las publicaciones científicas que las describen y las fuentes o bases de datos primarias de donde se obtienen los datos integrados. También se compararán los resultados de esta integración con otros recursos o bases de datos actuales que presenten aproximaciones similares.

Objetivo 2. Desarrollar una plataforma bioinformática de acceso web que permita consultar, explorar y obtener de modo eficaz las interacciones conocidas de una proteína concreta, de un listado de proteínas o de un organismo. Para ello, esta plataforma (con su estructura de datos subyacente) estará orientada hacia cada organismo tomando como referencia el proteoma de cada especie y mapeando sobre él su interactoma completo; es decir, el conjunto de interacciones proteína-proteína que se conocen actualmente para las distintas especies.

Objetivo 3. Desarrollar una herramienta para la visualización de redes de interacción que permita representar las relaciones entre proteínas de modo gráfico e interactivo. Esta herramienta estará apoyada en desarrollos actuales del consorcio internacional de *Cytoscape* y en nuevos abordajes que permitan tanto la creación de redes y subredes a medida como la inclusión gráfica de anotaciones biológicas funcionales de los nodos, así como modos prácticos de exportar los datos e información de las redes generadas a otras plataformas.

Objetivo 4. Realizar un análisis preliminar comparativo de algunos de los interactomas que se generen, abordando parámetros de topología de redes y métodos que permitan analizar sus características. Estos análisis comparativos incluirán los criterios de filtrado y de selección de redes y subredes que se quieren desarrollar en este trabajo para evaluar la validez y fiabilidad de las interacciones.

2.3 Solución Propuesta

Partiendo de la hipótesis central de este trabajo de investigación y sus objetivos principales, enunciamos aquí de modo breve la principal solución propuesta como preámbulo a la descripción detallada de todo el trabajo realizado. Se han desarrollado diferentes métodos para la integración, de forma exhaustiva y transparente, de datos experimentales de interacción de proteínas. También se ha diseñado e implementado una plataforma bioinformática para la exploración y análisis del conjunto de datos generados mediante la aplicación de dichos métodos **(Alonso-López et al. 2016)**.

Ambos desarrollos son originales y nuevos, pero parten de las ideas sobre las que se construyó, hace ya un década, una primera aplicación sobre interacción de proteínas llamada APID **(Prieto and De Las Rivas 2006)**. En este marco y por mantener el *link* o asociación con dicha primera aplicación, la nueva plataforma bioinformática construida como parte de esta Tesis Doctoral se ha denominado **APID Interactomes** (que está libremente accesible en la URL http://apid.dep.usal.es). Junto a dicha plataforma, esta Tesis Doctoral describe tanto los métodos y estrategias computacionales que se han seguido en su desarrollo, como una serie de análisis de los datos biomoleculares obtenidos. También describe la estructura de la base de datos construida y las herramientas de visualización y análisis que incorpora la propia plataforma bioinformática.

3 MATERIALES Y MÉTODOS

3.1 Obtención de interacciones registradas en las bases de datos primarias

Los datos originales sobre interacciones de proteínas se obtuvieron mediante la descarga de los ficheros disponibles en cada una de las bases de datos primarias que se iban a incluir en APID Interactomes.

Entre las diferentes bases de datos primarias existe cierta heterogeneidad en cuanto a los procesos de revisión de literatura y registro de interacciones que cada una de ellas lleva a cabo. Además, el nivel de adaptación a las recomendaciones del grupo PSI-MI es variable, aunque todas, excepto BioPlex, ofrecen su información en el formato PSI-MI XML 2.5. En el caso de esta, la información se proporciona en una página web sencilla de apoyo al proyecto, que dispone de un apartado con un enlace a un fichero de texto tabulado con toda la información sobre las interacciones.

A continuación, se detallan los ficheros utilizados para la extracción de los datos originales de cada una de ellas.

3.1.1 IntAct

La base de datos IntAct **(Kerrien et al. 2012)**, además de poderse explorar a través de un servidor web, dispone de un servidor FTP (37) donde se proporciona acceso libre, tanto a nivel académico como comercial y con licencia *Apache License 2.0* (38), a todos los datos que contiene.

Como se puede ver en **Figura 17**, los datos sobre interacciones están disponibles en los formatos PSI-MI TAB y PSI-MI XML 2.5. También se ofrecen algunos conjuntos de datos auxiliares como vocabularios controlados o subconjuntos de datos específicos de determinadas patologías o áreas de conocimiento.

En cuanto a los datos de interacciones en formato PSI-MI XML 2.5, se ofrecen agrupados por especie o por publicación. Para la integración de datos en APID Interactomes se optó por utilizar los ficheros organizados por identificador de taxonomía (especie).

- Contents of the IntAct FTP directory:
  - all.zip: entire contents of this directory, compressed.
  - cv: controlled vocabularies.
  - psi25: IntAct data in PSI-MI 2.5 format. See the PSI website for a detailed format description.
  - psimitab: IntAct data in PSI-MI TAB format. See the README for a detailed format description.
    intact.zip: entire contents of the database in a single file.
    pmidMITAB.zip: entire contents of the database, each file represents a publication.
  - various: Data subsets for exchange with collaborating databases.
  - Sentences for text-mining - textual evidence of manually curated interactions.
- Contents of the IntAct Biological Complexes FTP directory:
  - psi25: IntAct complexes in PSI-MI 2.5 format. See the PSI website for a detailed format description.

Figura 17. *Contenido del servidor FTP de IntAct (37).*

Para algunos organismos también se ofrecen pequeños conjuntos de interacciones para las que se ha descrito su no ocurrencia en determinadas condiciones. Estas interacciones se registran en ficheros con el sufijo "_negative", que son filtrados para no incluirse en el protocolo de recolección inicial de datos.

Para la versión inicial de APID Interactomes se procesaron un total de 1282 ficheros de datos en formato PSI-MI XML 2.5 correspondientes a 821 especies diferentes.

3.1.2 MINT

Desde Septiembre de 2013, justo al inicio de este trabajo, las bases de datos de MINT **(Licata et al. 2012)** e IntAct **(Kerrien et al. 2012)** llegaron a un acuerdo para que el equipo de expertos de MINT pasase a utilizar el protocolo de registro de interacciones desarrollado por IntAct **(Orchard et al. 2014)**, por lo que todas las interacciones registradas en MINT están contenidas ahora en los ficheros que se obtienen del servidor FTP de IntAct.

> **[NEW] Starting September 2013, MINT uses the IntAct database infrastructure to limit the duplication of efforts and to optimise future software development. Data manually curated by the MINT curators can now be accessed from the IntAct homepage at the EBI. Data maintenance and release, MINT PSICQUIC and IMEx services are under the responsibility of the IntAct team, while curation effort will be carried by both groups.**

Figura 18. *Mensaje informativo extraído de la página web de MINT (10) donde se anuncia el traslado de sus datos de interacciones a la base de datos de IntAct.*

Para la integración de estas interacciones se decidió seguir también el protocolo de adquisición de IntAct y asignarle dicha base de datos como origen con el identificador único que esta asigna a cada interacción que reporta.

3.1.3 HPRD

La información almacenada en la base de datos HPRD **(Keshava Prasad et al. 2009)** puede descargarse desde (39), previo registro de usuario. El acceso es libre para uso académico.

Se ofrece la información en ficheros de texto tabulados, en formato XML simple y en formato PSI-MI 2.5. Para este último existe la posibilidad de obtener todas las interacciones concatenadas en un único fichero XML. Dicho fichero fue el utilizado para procesar en APID Interactomes los datos originales de HPRD.

FILES AVAILABLE FOR DOWNLOAD

File name	File description	File type	Version number	Date created
HPRD_Release9_041310.tar.gz	This file contains human binary protein-protein interactions in tab delimited format.	Tab delimited	Release 9	04-13-10 Apr 13, 2010
HPRD_FLAT_FILES_041310.tar.gz	This file contains all entries contained in HPRD including features of proteins such as post-translational modifications, tissue expression, subcellular localization and protein-protein interactions in tab delimited file format as per the users request.	Tab delimited	Release 9	04-13-10 Apr 13, 2010
HPRD_XML_041310.tar.gz	This file contains all entries contained in HPRD including features of proteins such as post-translational modifications, tissue expression, subcellular localization and protein-protein interactions etc.	XML	Release 9	04-13-10 Apr 13, 2010
HPRD_PSIMI_041310.tar.gz	This file contains only protein-protein interactions of all entries contained in HPRD. The data is in PSI-MI format version 2.5	XML	Release 9	04-13-10 Apr 13, 2010
HPRD_SINGLE_PSIMI_041310.xml.tar.gz	This file contains only protein-protein interactions but all interaction records are concatenated into a single XML file. The data is in PSI-MI format version 2.5	XML	Release 9	04-13-10 Apr 13, 2010

Figura 19. *Ficheros disponibles para su descarga en la base de datos HPRD.*

3.1.4 BioGRID

Los datos almacenados en BioGRID **(Chatr-Aryamontri et al. 2015)** están disponibles de forma abierta tanto para uso académico como para uso comercial. Pueden descargarse desde (40) en diferentes formatos, incluyendo PSI-MI TAB y PSI-MI XML 2.5. Las interacciones están agrupadas por organismos, pero al igual que en HPRD, existe la posibilidad de descargarse todas en un único fichero.

El uso de este fichero único estuvo supeditado a la capacidad computacional disponible en cada momento puesto que el número de interacciones almacenadas en BioGRID es muy superior al de las almacenadas en HPRD, en cuyo caso siempre se pudo optar por el fichero único.

Para integrar BioGRID en la versión inicial de APID Interactomes se procesó un fichero de 3.7 Gb que debía ser almacenado completamente en memoria para poder evaluar de forma global el esquema XML y hacer el mapeado de objetos correspondiente. Para ello se utilizó el *framework* JAMI como se explicará más adelante cuando se hable del registro de interacciones.

3.1.5 BioPlex

Como ya se ha mencionado, en el caso de BioPlex **(Huttlin et al. 2015)** no hay una base de datos, pero sí una página web desde donde se pueden descargar tanto las interacciones contenidas en la publicación inicial como las disponibles en la última actualización.

Accediendo a (41) se puede obtener un fichero tabulado con la información básica de cada interacción. Para la integración de datos se utilizó el fichero correspondiente a las interacciones descritas a finales del año 2015. Este fichero contiene aproximadamente unas 56.000 interacciones mientras que el fichero que acompañaba a su publicación al inicio de 2015 contenía unas 24.000 interacciones.

3.2 Obtención de datos de referencia sobre proteínas

Una interacción se define esencialmente como el conjunto de participantes que la forman. Para definir una interacción de proteínas es clave la correcta identificación de estas, ya que cualquier otra información relativa a dicha interacción - como por ejemplo el método de detección - será siempre una anotación secundaria a la propia definición de sus participantes.

Si, además, como es el caso de este trabajo de investigación, se busca recopilar e integrar datos sobre interacciones provenientes de distintos repositorios y distintos métodos de registro, la identificación inequívoca de cada participante es uno de los puntos más críticos del proceso. Existe la posibilidad, como se verá más adelante, de que un mismo participante no sea registrado de la misma manera en todos los repositorios. Es decir, para la correcta integración de datos se deberá contar con un protocolo de correspondencia o *mapeo* de los códigos de identificación de los participantes de cada interacción.

Así, lo que se hizo fue construir un catálogo de proteínas a partir de todos aquellos participantes en interacciones que pudieron ser identificados como proteínas de forma inequívoca.

La construcción de este catálogo de referencia sobre proteínas se basó en la información disponible en UniProtKB **(The UniProt Consortium 2014)**, tanto en su versión revisada (SwissProt) como en su versión no revisada (TrEMBL). La información extraída de dicha base de datos también se utilizó para el sistema de validación y actualización de identificadores.

Los datos están contenidos en ficheros con formato UniProtKB XML, cuyo esquema puede consultarse en (42). Son ficheros de gran tamaño que solo pueden leerse y analizarse de forma secuencial usando eventos, es decir, sin cargarse en memoria. Esto se consigue mediante el uso de la librería *Simple API for XML (SAX)* (43).

De estos ficheros se extraen también las anotaciones funcionales para cada proteína, como se verá más adelante.

3.3 Obtención de datos de estructura 3D de proteínas

Para obtener los datos de las estructuras tridimensionales se descargaron de la base de datos PDB **(Rose et al. 2015)** (28) todas aquellas estructuras que contuvieran proteínas completas y fueran heterodímeras.

Para procesar la información con la interpretación de cada estructura en PDBsum **(De Beer et al. 2014)** (29) se implementó un pequeño script en Ruby que lee el código HTML de la página web correspondiente a cada estructura y genera una tabla con la información de los diferentes interfaces obtenidos.

Los datos resultantes se almacenaron en dos tablas de la base datos a partir de las cuales, mediante un script SQL, se asignaron identificadores PDB a las interacciones correspondientes, tal y como se explica en la sección **Asignación de estructuras 3D**.

3.4 Obtención de datos de ontologías

Para tener la posibilidad de referirse a determinadas entidades o conceptos de manera formal durante el proceso de integración se necesitan clasificaciones y vocabularios controlados. En APID Interactomes se ha recurrido a la clasificación estándar de especies y a la ontología de interacciones moleculares definida por HUPO PSI-MI.

3.4.1 Taxonomía de especies

Lo datos sobre la taxonomía de especies se obtuvieron de los servidores FTP del NCBI (44). Desde estos puede descargarse el archivo *taxdmp.zip* que contiene varios ficheros con el volcado de la base de datos de taxonomías de GenBank **(Benson et al. 2013)**. Entre ellos está el archivo *names.dmp* que almacena los nombres e identificadores de cada especie.

3.4.2 Métodos de detección de interacciones de proteínas

Uno de los principales objetivos de la organización HUPO PSI-MI es la estandarización de formatos de intercambio de datos de interacciones moleculares. Este proceso de estandarización incluye la creación y mantenimiento de una ontología pública con términos relativos a todo tipo de información relacionada con las interacciones moleculares.

Esta ontología puede explorarse a través del servidor web *Ontology Lookup Service* **(Côté et al. 2010)** y descargarse en formato OBO **(Osumi-Sutherland 2010)** (45).

De este fichero se pudo extraer toda la jerarquía de términos relacionados con los métodos de detección de interacciones de proteínas.

3.5 Obtención de proteomas

APID Interactomes, para poder calcular los tamaños de los interactomas, se basa en la cobertura que estos ofrecen sobre el proteoma del organismo correspondiente: se contabilizan todas las proteínas que participan en alguna interacción y se contrastan con el total de proteínas descritas para ese organismo.

La caracterización del proteoma de cada organismo es, por tanto, fundamental para cuantificar los interactomas que APID Interactomes ofrece. Por ello, la definición de proteoma para cada organismo se extrajo del proyecto Proteomes de UniProt (46), siempre que dicho organismo existiera en la base de datos. En caso de que no fuese así, se contabilizó el número total de proteínas asignadas a cada organismo en SwissProt y en TrEMBL. Para la versión inicial de APID Interactomes, con interactomas de 453 especies, se obtuvo el proteoma de un total de 263 organismos (58%) en UniProt Proteomes y 190 (42%) contabilizando el número de proteínas asignadas en SwissProt y TrEMBL.

En el caso de los proteomas disponibles en UniProt Proteomes, se incorporaron a la base de datos a partir de un fichero de texto tabulado que puede descargarse de su página web. Una vez incorporados estos datos en bruto, se cruzaron con los datos de taxonomías de organismos asignando a las 263 especies disponibles el número total de proteínas descritas para su proteoma.

Para obtener el proteoma de los otros 190 organismos se escribió un programa en Java que, utilizando los servicios REST de UniProt (47), calcula el número de proteínas asignadas a un organismo dado, tanto en SwissProt como en TrEMBL, y almacena el resultado en la tabla correspondiente de la base de datos.

3.6 Algoritmo de integración de interacciones

Todo el algoritmo de integración de interacciones que se detalla en la sección correspondiente del capítulo de resultados está implementado en el lenguaje de programación orientado a objetos JAVA.

Dicha implementación hace uso de algunas librerías como log4j (48) para el sistema de registro de eventos o los drivers JDBC (49) para el acceso a la base de datos.

Para la lectura y el procesamiento de los ficheros con formato estándar HUPO PSI-MI XML 2.5 se utilizó la plataforma JAMI (50). Para ello, al inicio del proyecto, se hizo una breve estancia en el equipo de desarrollo de JAMI, en el Instituto Europeo de Bioinformática (EBI, Cambridge), que después se convertiría en una colaboración a largo plazo, que hoy sigue vigente, para efectuar diversas pruebas y mejorar la implementación de JAMI. Gracias a esto, APID Interactomes es la primera herramienta pública que utiliza la plataforma JAMI para procesar las interacciones en el formato PSI-MI XML.

3.7 Arquitectura de la plataforma web APID Interactomes

La plataforma web desarrollada durante esta Tesis Doctoral está construida sobre el servidor web Tomcat y la plataforma STRUTS 2, proyectos ambos de *The Apache Software Foundation*. Todo el desarrollo implementado en el servidor está escrito en Java y su función principal es obtener datos de la base de datos de APID Interactomes para dar respuesta a las peticiones recibidas desde la parte cliente. En esta última, además de utilizar conjuntos de etiquetas STRUTS, se han empleado varias librerías adicionales para llevar a cabo diferentes tareas tal y como se explicará en la sección **Implementación en el cliente**.

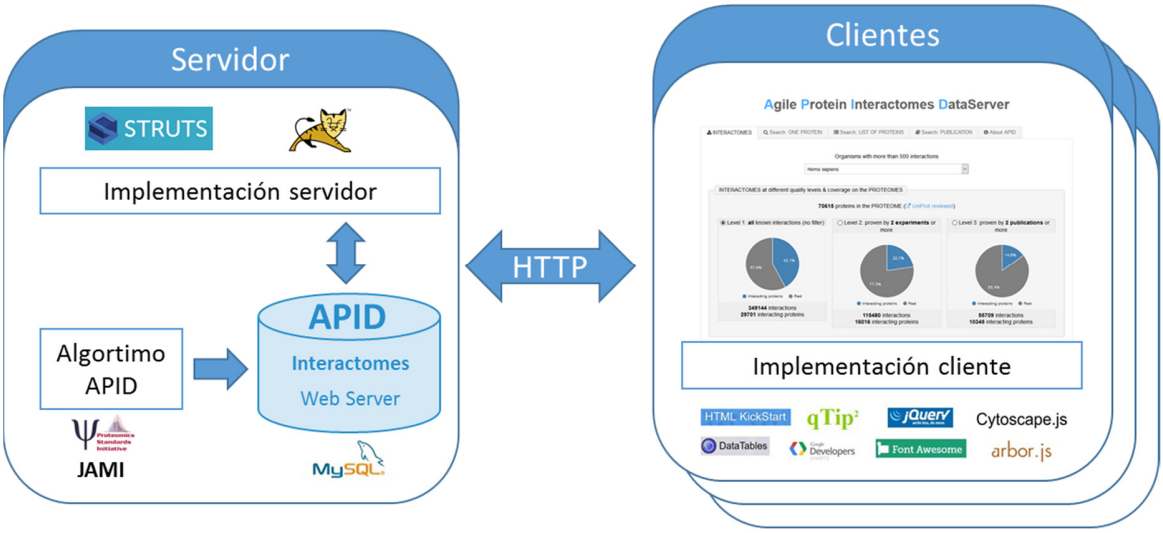


Figura 20. *Diagrama esquemático que representa la arquitectura de la plataforma web de APID Interactomes y las diferentes tecnologías utilizadas tanto a nivel de servidor como a nivel de cliente.*

En **Figura 20** se representa esquemáticamente la arquitectura de la plataforma web de APID Interactomes con algunas de las tecnologías utilizadas en cada contexto.

3.8 Implementación en el servidor

La parte servidor de la plataforma diseñada consiste en un desarrollo JAVA construido sobre la plataforma STRUTS en su versión 2 y cuyo objetivo principal es procesar las peticiones que recibe del lado cliente de la aplicación.

Para ello se define un catálogo de servicios web compuesto por diferentes métodos, o *acciones* en terminología STRUTS, que representan todas aquellas operaciones que el servidor es capaz de proporcionar al cliente. Así, para cada una de estas *acciones*, el servidor debe procesar los argumentos que el cliente aporta, obtener la información correspondiente de la base de datos, y enviar al cliente los resultados en el formato adecuado.

Un ejemplo simplificado de este sistema de peticiones sería el siguiente:

1. El servidor publica en su catálogo de servicios web la operación *buscaProteina* con el argumento *codigoUniProt*.
2. El cliente efectúa una petición del tipo *buscaProteina(Q03330)*.
3. El servidor recibe la petición, procesa el argumento *Q03330* y busca en la base de datos si existe una proteína con dicho identificador UniProt.
4. Si existe, obtiene los datos disponibles y construye un texto con el nombre de la proteína y el organismo.
5. El cliente recibe el texto *Histone acetyltransferase GCN5 (Saccharomyces cerevisiae)*.

De esta manera el servidor puede implementar cualquier operación, más o menos compleja, que tenga como dato de entrada una determinada información literal proveniente del cliente y como resultado cualquier procesamiento de esta que, normalmente, requerirá el uso de información adicional contenida en la base de datos APID Interactomes.

El servidor, por tanto, ha de implementar un sistema de acceso eficaz y seguro a la base de datos. Al tratarse de una base de datos relacional SQL, la comunicación directa entre el código Java y el sistema gestor de base de datos se construyó sobre el controlador MySQL de las librerías JDBC. Además, para convertir los tipos de datos de este sistema relacional en los utilizados en un lenguaje orientado a objetos como Java, se optó por seguir la técnica de programación denominada mapeo objeto-relacional o ORM (acrónimo de su nombre en lengua inglesa) de la que se hablará a continuación.

3.8.1 Mapeo objeto-relacional (ORM)

El mapeo objeto-relacional es una técnica de programación que trata de enlazar los tipos de datos disponibles en una base de datos relacional (como MySQL) con los presentes en un lenguaje de programación orientada a objetos (como Java). Para ello se construye una capa intermedia que proporciona cierto nivel de abstracción sobre las características propias del sistema gestor de bases de datos y permite al programador trabajar con clases y objetos puros. De esta manera, resulta posible implementar la lógica del programa siguiendo el paradigma de la programación orientada a objetos sin preocuparse de la persistencia de dichos objetos.

Existen varios desarrollos software, tanto libres como comerciales, destinados a ocupar esta capa intermedia y servir de *motor de persistencia*. Algunos de los más utilizados son Hibernate (51) para desarrollos en Java o Doctrine (52) para desarrollos en PHP. También, en muchos proyectos software, el motor de persistencia se implementa desde cero para adaptarse a sus necesidades de una forma más personalizada.

En el desarrollo de la plataforma web de APID Interactomes se optó por esta última estrategia: programar desde cero la capa ORM para así tener más control sobre las conversiones de tipo y las clases resultantes. De esta manera, se generó todo un paquete de clases Java dedicadas a establecer las correspondencias necesarias entre las tablas de la base de datos relacional y la estructura de clases de la aplicación.

Al igual que ocurre en los desarrollos de terceros, estas clases también implementan los métodos de acceso, control de tipos y transformación de datos necesarios para el funcionamiento de la aplicación. Además, se construyeron clases con información combinada de varias tablas y pre-cálculos de determinados parámetros para dotar de mayor agilidad a la aplicación web.

3.8.2 Sistema gestor de bases de datos

Como sistema gestor de bases de datos se utilizó MySQL (53). Se valoró la posibilidad de utilizar bases de datos orientadas a grafos como Neo4j (54) pero conceptualmente no era la mejor solución: APID Interactomes construye redes a partir de interacciones binarias, pero no almacena redes como grafos ni hace consultas sobre estas. Sin embargo, sí resultó de enorme utilidad realizar ciertas optimizaciones adaptadas al uso que la aplicación web hace de los datos. Estas optimizaciones se describen en la sección **Optimización de la base de datos** del capítulo de resultados.

3.9 Implementación en el cliente

Toda la interfaz de la parte cliente de la aplicación web está basada en los estándares HTML y CSS (55). El desarrollo de interfaces web está supeditado siempre a dichos estándares ya que los navegadores solo pueden construir en pantalla los elementos definidos en estos. Con el uso del lenguaje de programación Javascript, y las librerías construidas sobre este, se pueden conseguir combinaciones complejas de elementos HTML formateados convenientemente con CSS y con un comportamiento concreto definido en dicho lenguaje. Es así como todo el interfaz de APID Interactomes fue desarrollado.

Como punto de partida, hoja de estilos general, y base para la maquetación y distribución de espacios en la aplicación se utilizó la plataforma HTML Kickstart (56).

A lo largo de toda la implementación se usó de forma intensiva la librería jQuery (57), que permite un manejo ágil del árbol de objetos de una página web a través del API DOM (*Documento Object Model*), y además es la base sobre la que se han diseñado gran parte de las librerías que se usaron en este desarrollo.

Para ofrecer pequeñas ayudas al usuario, en cada pantalla existe un enlace denominado “Show help” que, al pulsarlo, muestra algunas indicaciones contextuales sobre diferentes elementos de la interfaz activa. Para conseguir esto se utilizó la librería qTip2 (58).

Font Awesome (59) se utilizó para el uso de caracteres gráficos como iconos a lo largo de toda la aplicación.

Para mostrar de forma gráfica la cobertura de los diferentes interactomas sobre su proteoma correspondiente se utilizaron las librerías Charts de Google Developers (60).

De especial utilidad resultó la librería DataTables (61) para implementar las tablas de datos, así como varias de las funcionalidades asociadas a ellas como, por ejemplo, los sistemas de filtrado.

3.9.1 Visualizador de redes

Para la representación interactiva de la red se utilizó la librería cytoscape.js **(Franz et al. 2015)** (62). Se trata de una librería JavaScript con licencia MIT (63) que trata de llevar al entorno de los navegadores web algunas de las funcionalidades del software Cytoscape.

Esta librería está basada en Javascript y, por tanto, es compatible con cualquier navegador actual sin ningún tipo de complemento de terceros. Además, proporciona una visualización interactiva básica a partir de una estructura de datos JSON con la información sobre la red y un API con el que se pueden implementar funcionalidades adicionales.

Para la implementación del *layout* arbor se utilizó la librería Arbor.js (64) y la extensión cytoscape-arbor (65).

3.10 Análisis de interactomas

Por último, en la parte final donde se llevan a cabo diferentes análisis sobre algunos de los interactomas generados por el algoritmo de APID Interactomes, se utilizaron varias herramientas de forma combinada tal y como se describe a continuación.

El primer paso fue generar las redes correspondientes a los interactomas de cada organismo. Para poder visualizar y analizar dichas redes se utilizó la aplicación Cytoscape **(Shannon et al. 2003)**. Desde dicha aplicación se importó el fichero de datos generado por APID Interactomes incluyendo como atributos de los arcos las métricas calculadas para cada interacción. A continuación, con Cytoscape, se diseñaron diferentes filtros basados en dichas métricas y se creó un estilo visual adecuado para grandes redes. Por último, se crearon subredes para los diferentes análisis usando estos filtros para establecer el número mínimo de experimentos por interacción.

Para analizar topológicamente las redes se utilizó la aplicación NetworkAnalizer **(Doncheva et al. 2012)**, que funciona como complemento de Cytoscape y permite calcular multitud de parámetros en un tiempo razonable a pesar de tratarse de redes de gran tamaño.

Para explorar las redes en busca de agrupamientos de nodos altamente conectados se utilizó la aplicación MCODE **(Bader and Hogue 2003)**, que funciona también como complemento de Cytoscape.

Por último, para tratar de caracterizar funcionalmente los genes asociados a estos agrupamientos, se utilizaron varios algoritmos de enriquecimiento: *GeneTerm Linker* **(Fontanillo et al. 2011)** (66), *GeneCodis* **(Carmona-Saez et al. 2007; Nogales-Cadenas et al. 2009; Tabas-Madrid et al. 2012)** (67) y *Enrichr* **(Kuleshov et al. 2016)** (68).

Para la identificación de algunos de estos agrupamientos se utilizó también la base de datos C*omplex Portal* del EMBL-EBI **(Meldal et al. 2015)** (69).

4 RESULTADOS

4.1 Protocolos para la integración de datos sobre interacciones moleculares físicas entre proteínas

Como se ha mencionado anteriormente, uno de los objetivos principales de este trabajo es la construcción de un conjunto de datos de interacciones de proteínas que integre y unifique la información disponible en diferentes repositorios abiertos a la comunidad científica. Dicha integración de datos presenta diversos retos, tanto a nivel conceptual como a nivel técnico y computacional. A lo largo de este capítulo se describen las diferentes propuestas y decisiones de diseño tomadas para lograr una integración de datos exhaustiva y con un nivel de calidad controlado.

La base de datos de APID Interactomes integra solo interacciones físicas entre proteínas que han sido descritas en publicaciones científicas mediante el uso de métodos experimentales. Existen también en la literatura descripciones de interacciones predichas computacionalmente, así como asociaciones funcionales, ambas sin soporte experimental que pruebe el contacto físico entre los interactores. Este tipo de interacciones funcionales, denominadas en algunos casos *interacciones genéticas* **(Mani et al. 2008)**, son recogidas en algunas de las bases de datos ya mencionadas anteriormente **(Kerrien et al. 2012; Zuberi et al. 2013; Szklarczyk et al. 2015)**

Para conseguir esto, el algoritmo diseñado recopila solo datos experimentales y además lo hace usando siempre como referencia la base de datos UniProt **(The UniProt Consortium 2014)**, tanto para las proteínas como para los proteomas. De esta manera, y como se explicará más adelante, el algoritmo descartará cualquier interacción cuyos interactores no puedan ser relacionados de forma inequívoca con un código UniProt. También, para poder evaluar la cobertura de los interactomas obtenidos, recurrirá a los datos disponibles en el proyecto UniProt Proteomes siempre que sea posible.

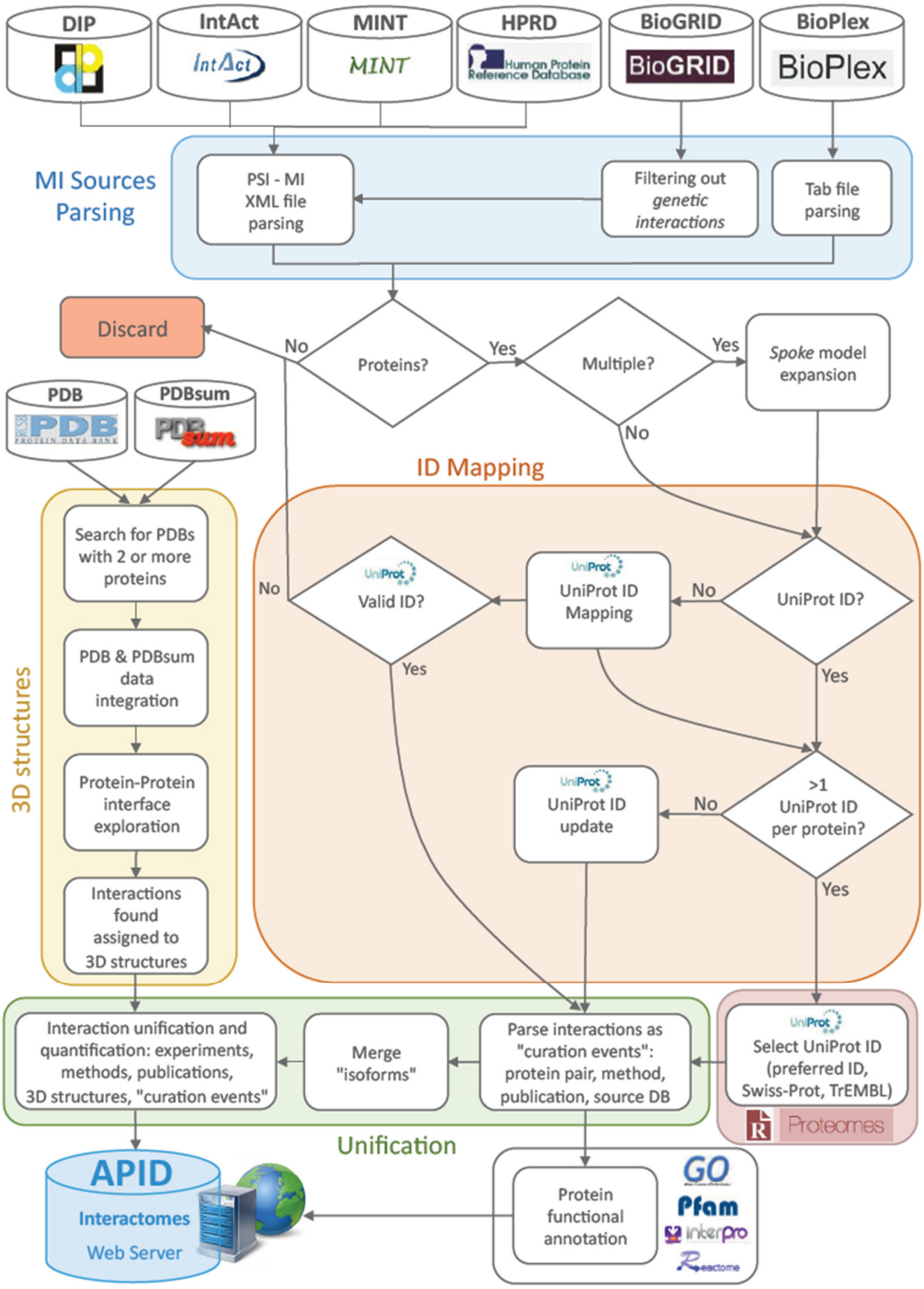


Figura 21. *Diagrama de flujo que representa el protocolo de adquisición e integración de datos de interacciones de proteínas implementado en APID Interactomes. Extraído de* **(Alonso-López et al. 2016)**.

A partir de los datos de interacciones experimentales cuyos interactores hayan sido identificados como proteínas, se llevará a cabo un proceso de unificación que permitirá registrar una interacción

única para cada par de proteínas a la que, además, se asociará una serie de métricas calculadas a partir de todos los registros procesados que describen dicha interacción.

En cuanto a los datos estructurales, el algoritmo usará como referencia el catálogo de estructuras disponible en la base de datos PDB **(Rose et al. 2015)** y analizará la interpretación de dichas estructuras que el proyecto PDBsum **(De Beer et al. 2014)** proporciona para intentar detectar en ellas la presencia de alguna de las interacciones que previamente han sido unificadas.

Finalmente, se llevará a cabo un proceso de anotación sobre todas las proteínas que participen en alguna de las interacciones procesadas. En **Figura 21** puede verse un diseño esquemático de las diferentes fases del algoritmo, que se explican de forma más detallada a continuación.

4.1.1 Procesamiento de datos primarios

Como estrategia general, a la hora de recoger los datos registrados en las diferentes bases de datos y repositorios, se tuvieron en cuenta los siguientes criterios:

1. Obtener la información menos procesada posible: Puesto que el objetivo es integrar los datos según nuestro criterio, siempre se ha intentado partir de los datos más crudos y detallados que cada base de datos o repositorio ofrezca.
2. Priorizar los formatos definidos por HUPO PSI-MI: el cumplimiento de estos estándares debería garantizar en cierto modo la calidad en el proceso de curación de datos.
3. Asegurar dentro de lo posible el uso de ontologías como complemento indispensable al punto anterior.
4. Control de versiones: Cada versión de la base de datos de APID Interactomes estará asociada a las versiones concretas de cada base de datos origen, tanto de interacciones como de referencia de proteínas y estructuras 3D.

En base a estos criterios, y de modo general, se recopilaron datos sobre:

1. **Interacciones**: Como ya se ha indicado anteriormente, la base de datos de APID Interactomes integra datos experimentales de interacciones físicas entre proteínas. En **Tabla 1** se detalla en número de interacciones obtenidas de cada base de datos primaria.
2. **Proteínas**: Cada interacción está definida por, al menos, dos participantes. Es crucial poder identificar de forma inequívoca a dichos participantes puesto que el éxito de dicha identificación será el criterio para aceptar o no cada interacción procesada. Siendo el dato principal de APID Interactomes las interacciones proteína-proteína, disponer de un catálogo de estas es requisito indispensable.
3. **Datos de estructura 3D de proteínas**: Como se explicará más adelante, APID Interactomes utiliza datos sobre estructuras tridimensionales, descritas en la literatura científica, para identificar contactos físicos entre proteínas y asignarlos a la interacción correspondiente como soporte experimental adicional.
4. **Datos de ontologías**: En el protocolo de adquisición de datos e integración de interacciones se utilizan diversas ontologías que proporcionan vocabularios controlados para diferentes dominios de conocimiento.

5. **Proteomas**: Se recopilan también los datos disponibles sobre proteomas para diversos organismos con el objeto de mejorar el cálculo de cobertura del interactoma correspondiente.

Tabla 1. *Tabla comparativa que incluye las diferentes fuentes de datos de interacciones usadas por APID Interactomes. En ella se detallan el tipo de interacciones recogidas en cada caso, así como el número de proteínas e interacciones registradas según cada página web.*

Base de datos	Tipo de interactores	Tipo de interacciones	Organismos	Número de proteínas	Número de interacciones
IntAct	Proteínas y otras moléculas	Experimentales	Todos	86812	349690
MINT	Proteínas	Experimentales	Todos	30089	83744
HPRD	Proteínas	Experimentales	Humano	30047	41327
DIP	Proteínas	Experimentales	Todos	27701	79339
BioGRID	Proteínas y otras moléculas	Experimentales y genéticas	Todos	56907	557106
BioPlex	Proteínas	Experimentales	Humano	7668	23744

4.1.2 Registro de ontologías

Una vez obtenidos todos los datos, el primer paso fue generar las tablas auxiliares correspondientes a las dos ontologías principales, la de organismos y la de métodos de detección de interacciones.

id	psimi_id	description
1	MI:0001	interaction detection method
2	MI:0002	participant identification method
3	MI:0003	feature detection method
4	MI:0004	affinity chromatography technology
5	MI:0005	alanine scanning
6	MI:0006	anti bait coimmunoprecipitation
7	MI:0007	anti tag coimmunoprecipitation
8	MI:0008	array technology
9	MI:0009	bacterial display
10	MI:0010	beta galactosidase complementation

Figura 22*. Extracto de la tabla "method" de la base de datos de APID Interactomes, que almacena la ontología correspondiente a los términos que describen los diferentes métodos de detección de interacciones.*

En este último caso, se trata de una tabla sencilla que simplemente relaciona un número entero, que se usa como clave primaria, con el término de la ontología PSI-MI que coincide numéricamente con él. En **Figura 22** puede verse un extracto de dicha tabla.

En el caso de los organismos, aparte del identificador numérico y nombre del término de la taxonomía, se almacenan otros campos necesarios para el cálculo de la cobertura de los interactomas de cada organismo con tres niveles de calidad diferentes. Aunque de esto se hablará más adelante, cabe adelantar que estos campos contienen datos de los proteomas obtenidos anteriormente y de los resultados de los cálculos de tamaño de cada interactoma presente en APID Interactomes en tres niveles distintos de calidad. En **Figura 23** puede verse un extracto de la tabla *taxon*, que es la que almacena todos estos datos.

name	proteome	reviewed	interactions1	interactingproteins1
Homo sapiens	69986	UP000005640	32092	8706
Saccharomyces cerevisiae (strain ATCC 204508 / S288c)	6721	UP000002311	19325	3878
Escherichia coli (strain K12)	4252	UP000000318	4020	1389
Mus musculus	50189	UP000000589	2489	1950
Arabidopsis thaliana	31477	UP000006548	1960	1309
Drosophila melanogaster	22005	UP000000803	1336	1258
Caenorhabditis elegans	26596	UP000001940	1327	1138
Schizosaccharomyces pombe (strain 972 / ATCC 24843)	5121	UP000002485	916	689
Campylobacter jejuni subsp. jejuni serotype O:2 (strain NCTC 11168)	1623	UP000000799	471	373
Rattus norvegicus	29885	UP000002494	326	366
Treponema pallidum (strain Nichols)	1028	UP000000811	172	110
Bos taurus	24112	UP000009136	93	136
Xenopus laevis	16615	NA	67	78
Synechocystis sp. (strain PCC 6803 / Kazusa)	3507	UP000001425	61	63
Helicobacter pylori (strain ATCC 700392 / 26695)	1553	UP000000429	59	92

Figura 23. *Extracto de la tabla "taxon" de la base de datos de APID Interactomes, que almacena datos sobre todos los organismos para los que se han registrado interacciones entre sus proteínas.*

Estas dos ontologías son las únicas que se almacenan localmente puesto que forman parte intrínsecamente del protocolo de obtención e integración de interacciones. No obstante, en la fase de anotación de proteínas de dicho proceso, se trabaja también con diversas ontologías funcionales que, en este caso, no necesitan ser almacenadas en la propia base de datos de APID Interactomes.

4.1.3 Sistema para la validación de participantes

Como ya se ha explicado anteriormente, la toma de decisiones en APID Interactomes sobre la aceptación o no de un participante en una interacción está basada en la posible catalogación de este en UniProtKB. Todo participante que, a pesar de estar declarado como proteína y contener su identificador correspondiente, no puede catalogarse en UniProtKB, se descarta. Y, como consecuencia, se descarta también la interacción en la que participa independientemente de que el otro participante sea válido o no.

Teóricamente, si los identificadores fueran únicos y estables en el tiempo, y si los protocolos de registro de interacciones de las diferentes bases de datos primarias tuvieran idénticos criterios a la hora de asignar identificadores a los participantes, bastaría con hacer una simple comprobación de la existencia de cada participante en UniProtKB para decidir si la interacción en la que aparece debe registrarse en APID Interactomes o no.

Sin embargo, la realidad demuestra que esto no es así ya que existen diferentes situaciones problemáticas relacionadas con la identificación de los participantes. A saber:

1. Hay participantes que no son proteínas: Este es el caso más evidente y menos problemático puesto que el tipo de participante está controlado por las ontologías del PSI-MI y, si el protocolo de registro de interacciones de la base de datos primaria correspondiente es correcto, se usará el término adecuado del vocabulario controlado.
2. Existen participantes que se representan como proteínas, pero no lo son: Durante la incorporación de datos se han encontrado participantes declarados como proteínas que finalmente resultan ser otro tipo de molécula. Esto puede deberse a una mala interpretación por parte de los expertos o a errores en la publicación o en el protocolo de registro de la base de datos primaria.
3. Los identificadores pueden no ser únicos: Cada participante puede tener varios identificadores, en forma de referencias cruzadas a diferentes bases de datos, que representen diferentes registros.
4. Los identificadores pueden referirse a distintos tipos de entidades biológicas: Un mismo participante puede tener varios identificadores correspondientes a diferentes conceptos biológicos, por ejemplo, dos identificadores para la proteína, uno para el gen asociado y dos más para dos secuencias posibles de dicho gen.
5. Un mismo participante puede tener varios identificadores del mismo tipo: En el caso de las bases de datos HPRD y BioGRID, existen interacciones donde un participante aparece con varios identificadores de proteína sobre la misma base de datos, normalmente UniProtKB.

La estrategia más directa ante este tipo de problemas es la de descartar todas las interacciones en las que aparezcan participantes cuya identificación no sea inequívoca, pero esto provocaría una pérdida de cobertura importante. Es por ello que, a la hora de diseñar el protocolo de integración de interacciones, se ha intentado refinar la identificación de participantes mediante el uso de diferentes estrategias que se detallarán en los siguientes apartados.

4.1.3.1 Participantes que no son proteínas

El primer filtrado que se establece en el protocolo de incorporación de interacciones es el de que la interacción sea proteína-proteína. Para identificar el tipo de participantes en cada interacción se utiliza el término *interactor type* (MI:0313) de la ontología PSI-MI. En **Figura 24** puede verse parte del árbol de términos definidos para dicha identificación.

Así, toda interacción en la que intervenga algún participante no declarado como proteína con el término *protein* (MI:0326) se descarta. Existe también la posibilidad de que un participante declarado como proteína resulte no serlo, en cuyo caso el protocolo de asignación de identificador no será capaz de catalogarlo en UniProtKB y la interacción en la que participa será desechada igualmente.

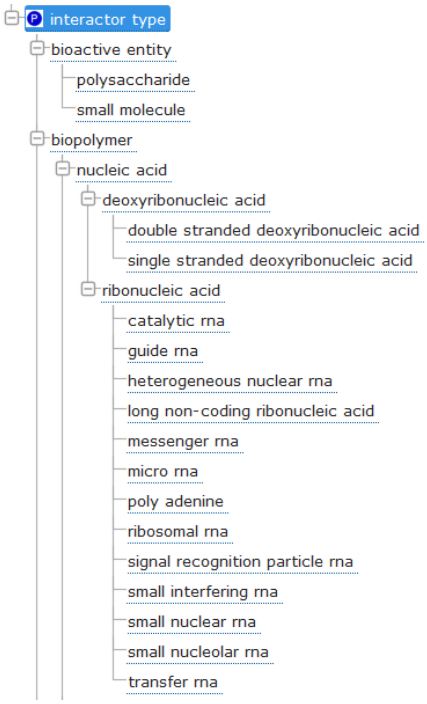


Figura 24. *Extracto de la jerarquía de términos definida por el PSI-MI para la identificación del participante o interactor.*

Como ejemplo de este primer proceso de filtrado, para la versión inicial de APID Interactomes y durante la exploración de las interacciones registradas en IntAct para el organismo Mus Musculus, el algoritmo procesó 184575 participantes entre los que encontró 1907 (aproximadamente un 1%) no declarados como proteínas. En Tabla 2 puede verse el desglose de estos participantes por tipos.

Tabla 2. *Número de ocasiones en las que el algoritmo detectó un participante no declarado como proteína para la exploración de los datos de interacciones del organismo Mus Musculus en la base de datos IntAct.*

Término ontología PSI-MI	Identificador	Número de apariciones
gene	MI:0250	869
small molecule	MI:0328	593
peptide	MI:0327	145
dna	MI:0319	137
ds dna	MI:0681	50
rna	MI:0320	41
ss dna	MI:0680	24
poly a	MI:0679	15
snrna	MI:0607	13
complex	MI:0314	12
mrna	MI:0324	6
ds dna	MI:0681	1
nucleic acid	MI:0318	1

4.1.3.2 Identificación de participantes declarados como proteínas

Una vez filtrados todos aquellos participantes que no son proteínas, el algoritmo intenta asignar un identificador UniProtKB válido a cada participante que evalúa. Este proceso de asignación contempla diferentes escenarios tal y como se detalla a continuación. Su funcionamiento esquemático puede verse en la **Figura 25**.

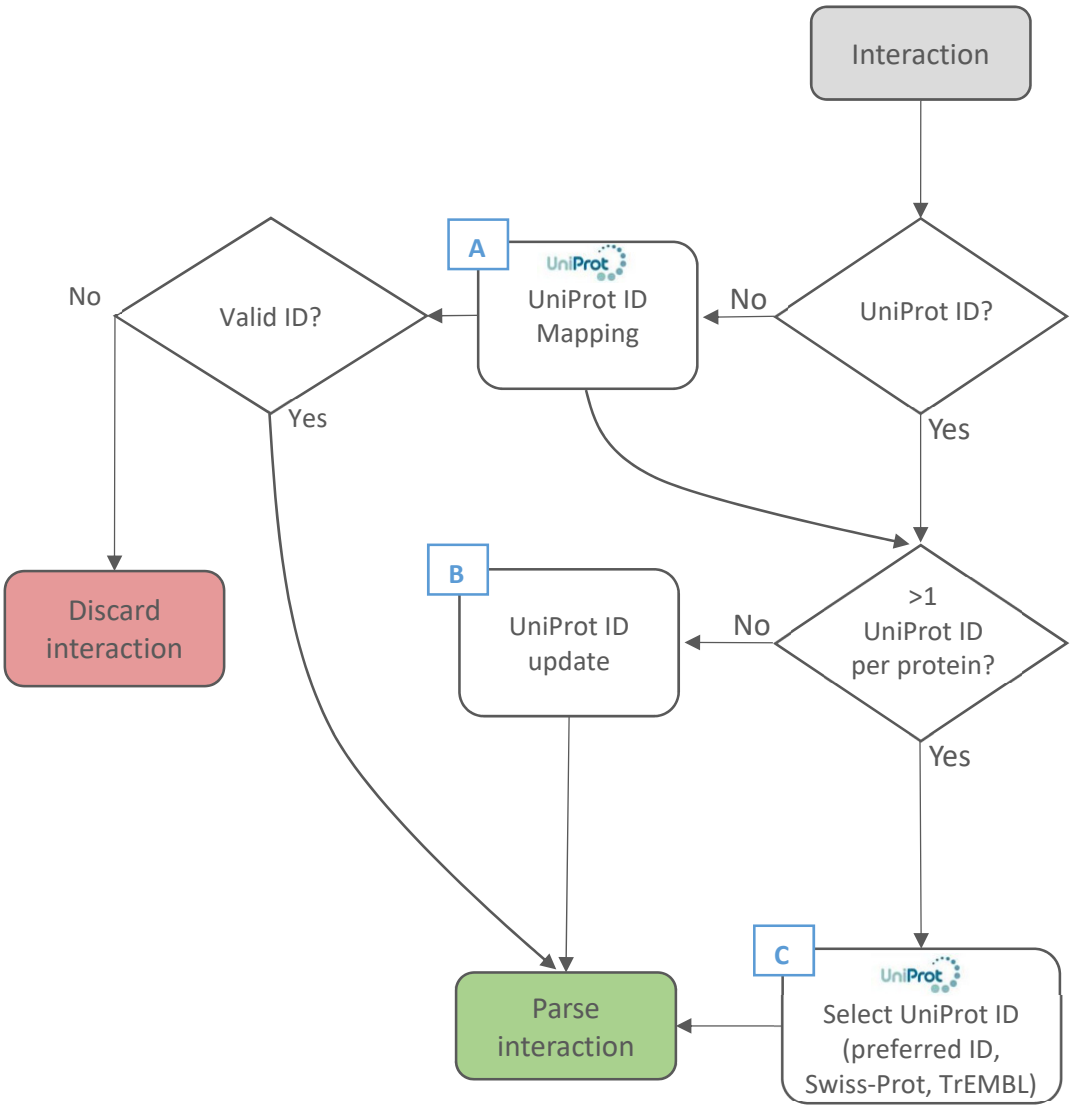


Figura 25. *Diagrama de flujo que representa el proceso de asignación de identificador a cada participante que ha sido declarado como proteína en una interacción.*

Lo primero que el algoritmo comprueba es la existencia de al menos un identificador en la base de datos UniProtKB para cada uno de los participantes de la interacción. Si este es único, se procede a comprobar si está actualizado y, de ser así, se procesa la interacción. En caso contrario, se actualiza previamente. Si hay múltiples identificadores UniProt se procede a seleccionar el más adecuado (por orden de prioridad, como se explicará después) y después se registra la interacción usando dicho identificador. Por último, en el caso de que no haya ningún identificador UniProt, si existen otros

identificadores de diferentes bases de datos para los que pueda encontrarse una equivalencia con un identificador de UniProt, se almacena la interacción. En caso contrario, se descarta.

El algoritmo contiene tres procedimientos principales, etiquetados en **Figura 25** como A, B y C. En dos de ellos (A y C, marcados con el logo de UniProt) se utilizan los servicios REST del servidor web de UniProt (70), en concreto cuando se necesita convertir un identificador de otra base de datos a UniProtKB a través del servicio *UniProt ID Mapping* (71) (A) y cuando se tiene más de un identificador de UniProtKB para la misma proteína (C), mientras que en el caso del procedimiento de actualización de identificadores UniProtKB (B), este se hace de forma local utilizando la base de datos de APID Interactomes. A continuación, se detallan cada uno de estos tres procedimientos.

4.1.3.2.1 Conversión de identificadores

Cuando el algoritmo encuentra un participante sin identificador UniProt, procede a explorar el fichero XML en busca de otro tipo de identificadores puesto que, según la especificación HUPO PSI-MI, todo participante debe contener al menos uno.

Cada identificador encontrado se envía a la herramienta de conversión de identificadores de UniProt utilizando la interfaz REST desde el código de Java. Si para alguno de los identificadores encontrados se obtiene un equivalente en UniProt se procede al etiquetado del participante con el identificador obtenido y se continúa con los siguientes pasos del algoritmo.

Existe la posibilidad de que, como resultado de este proceso iterativo de conversión de identificadores, se obtenga más de un identificador UniProt. En ese caso, la lista de identificadores UniProt se envía al procedimiento de selección de identificador, de la misma manera que se haría si inicialmente, en la información procedente de la base de datos primaria, el participante estuviera registrado con más de un identificador UniProt.

4.1.3.2.2 Actualización de identificadores

Como se mencionaba anteriormente, este procedimiento se lleva a cabo sin necesidad de utilizar los servicios REST de UniProt. Esto es posible porque, durante la adquisición de datos de referencia de proteínas, se construye también una tabla de equivalencias entre identificadores antiguos y actuales.

Para ello, se examinan las cabeceras *<entry>* de cada elemento XML que almacena una proteína y se registra la historia de identificadores que esta ha tenido desde que existe en la base de datos. En **Figura 26** puede verse un extracto de la cabecera del fichero XML correspondiente al registro de UniProt para el receptor de insulina en *Homo Sapiens* (INSR_HUMAN). En dicho extracto puede apreciarse que esta proteína ha tenido, a lo largo del tiempo, un total de seis identificadores diferentes.

```
− <uniprot xsi:schemaLocation="http://uniprot.org/uniprot http://www.uniprot.org/support/docs/uniprot.xsd">
  − <entry dataset="Swiss-Prot" created="1988-01-01" modified="2016-05-11" version="225">
      <accession>P06213</accession>
      <accession>Q17RW0</accession>
      <accession>Q59H98</accession>
      <accession>Q9UCB7</accession>
      <accession>Q9UCB8</accession>
      <accession>Q9UCB9</accession>
      <name>INSR_HUMAN</name>
```

***Figura 26**. Extracto del fichero XML de UniProt correspondiente al receptor de insulina en Homo Sapiens.*

Así, a partir de estos ficheros, el protocolo diseñado almacena todos los posibles identificadores que cada proteína ha tenido a lo largo del tiempo y los asocia con el actual de tal manera que siempre puedan detectarse participantes con referencias obsoletas. Esto es importante ya que permite evitar ambigüedades y duplicidades a la hora de registrar las interacciones.

Cuando el servidor web de UniProt recibe un identificador obsoleto, automáticamente proporciona el registro de la proteína con su identificador actual lo que provoca que, siguiendo con el ejemplo del receptor de la insulina, una comprobación de existencia de proteína para el identificador "Q9UCB7" se ejecute con éxito. Pero también lo haría otra para el identificador "Q17RW0" y, por supuesto, para el "P06213" que, a fecha de hoy, es el correcto. Este comportamiento de UniProt, unido al hecho de que las anotaciones de participantes en las bases de datos primarias no siempre son actuales, provoca que una misma interacción pueda ser registrada tantas veces como identificadores obsoletos diferentes tengan sus participantes.

De esta manera, la actualización de identificadores implementada en el protocolo de integración de interacciones de APID Interactomes evita que aparezcan problemas de este tipo, evitando así la redundancia en la denominación de las proteínas participantes y el aumento, de forma artificial, del número total de interacciones.

4.1.3.2.3 Selección de identificador UniProt

La selección de identificador UniProt es necesaria siempre que un participante tenga asignado más de un identificador UniProt. Esto puede ser debido a que inicialmente aparezca así anotado o a que se le asigne más de un identificador UniProt durante el proceso de conversión de identificadores descrito anteriormente.

En **Figura 27** puede verse un fragmento de fichero XML, de la base de datos HPRD, que contiene los identificadores disponibles - P28482, Q499G7 y Q1HBJ4 en UniProt - para la proteína MAPK1/ERK2. Para que la identificación del participante sea inequívoca y la interacción pueda registrarse en base a sus dos componentes, este debe tener un solo identificador y, por tanto, el algoritmo debe seleccionar uno entre los disponibles.

```
− <interactor id="01496">
  + <names></names>
  − <xref>
      <primaryRef db="HPRD" dbAc="MI:0468" id="01496" refType="identity" refTypeAc="MI:0356"/>
      <secondaryRef db="uniprot" dbAc="MI:0486" id="P28482,Q499G7,Q1HBJ4"/>
      <secondaryRef db="entrezgene" dbAc="MI:0477" id="5594"/>
      <secondaryRef db="omim" dbAc="MI:0480" id="176948"/>
      <secondaryRef db="pdb" dbAc="MI:0460" id="1PME"/>
    </xref>
```

Figura 27. *Extracto del fichero XML obtenido de la base de datos HPRD donde puede verse un participante con más de un identificador UniProt.*

La selección de identificador implementada se basa en la calidad de la anotación a la que representa dicho identificador, es decir, busca el registro que tiene la información más revisada. Para el ejemplo que aquí se expone, al consultar UniProt a través de sus servicios REST se obtienen los resultados que se detallan en **Tabla 3**. En este caso, el algoritmo seleccionaría el identificador P28482 que es el que pertenece a la base de datos SwissProt y en consecuencia está marcado como *reviewed*.

Tabla 3. *Resultados para la consulta de los tres posibles identificadores de la proteína MAPK1/ERK2 en UniProt.*

Identificador UniProt	Estado	Nivel de anotación
P28482	Reviewed (SwissProt)	5/5
Q499G7	Unreviewed (TrEMBL)	2/5
Q1HBJ4	Unreviewed (TrEMBL)	5/5

Cuando el algoritmo encuentra un conjunto de identificadores con igual nivel de anotación se queda con el primero de ellos.

En las especificaciones más recientes del formato XML de HUPO PSI-MI existe la posibilidad de especificar el identificador más adecuado usando la entidad *preferredId*. El algoritmo está preparado para, en caso de encontrar esta entidad, respetar la decisión del equipo de anotación de la base de datos primaria. Sin embargo, a día de hoy, la mayoría de las bases de datos que presentan identificadores múltiples no hacen uso de esta posibilidad.

4.1.4 Registro de interacciones

Cuando los participantes de una interacción han sido validados y catalogados con su referencia a UniProt, el algoritmo pasa a examinar y registrar en la base de datos la interacción entre ellos. El protocolo siempre evalúa interacciones binarias con solo dos participantes porque, cuando se detecta una interacción múltiple en la que están implicadas más de dos proteínas, esta es convertida a varias interacciones binarias, tal y como se explica más adelante en el apartado **Interacciones múltiples**.

Dado que una misma interacción puede estar registrada en diferentes bases de datos primarias, y para poder establecer diferentes niveles de fiabilidad para dicha interacción, APID Interactomes mantiene dos entidades de datos diferentes:

1. **Interacción** (tabla *interaction*): Almacena cada interacción única registrada en la base de datos. Dicha interacción se representa con un identificador interno y con el par de proteínas que participan en ella. También contiene los parámetros que se calculan a partir de los diferentes registros que dicha interacción tenga en las bases de datos primarias. Los campos de la tabla de la base de datos de APID Interactomes son:
    1. **Id**: Identificador local único de cada interacción.
    2. **Participant1**: Identificador UniProt de una de las proteínas participantes en la interacción.
    3. **Participant2**: Identificador UniProt de la otra proteína participante en la interacción.
    4. **Methods**: Número de métodos de detección de interacciones distintos que aparecen en los registros de las bases de datos primarias referentes a esta interacción. Parámetro calculado en la unificación de interacciones.
    5. **Papers**: Número de publicaciones distintas que aparecen en los registros de las bases de datos primarias referentes a esta interacción. Parámetro calculado en la unificación de interacciones.
    6. **Experiments**: Número de combinaciones método de detección-publicación distintas que aparecen en los registros de las bases de datos primarias referentes a esta interacción. Parámetro calculado en la unificación de interacciones.
    7. **Curationevents**: Número total de registros en bases de datos primarias referentes a esta interacción. Parámetro calculado en la unificación de interacciones.
    8. **Pdb**: Número de estructuras 3D de la base de datos PDB que el algoritmo ha asignado a esta interacción.
2. **Registros en bases de datos primarias** (tabla *intmetpaper*): Almacena cada registro de base de datos primaria que se refiere a esa interacción. Es lo que en APID Interactomes se ha denominado *curation event*. Estos registros almacenan información sobre la publicación donde se describe dicha interacción y sobre el método de detección utilizado. También se almacena la información necesaria para que no se pierda la trazabilidad con el registro original de la base de datos primaria. Los campos de la tabla de la base de datos de APID Interactomes son:
    1. **Id**: Identificador local único del registro.
    2. **Interaction**: Clave foránea para referirse al identificador único de la interacción en la tabla *interaction* de la base de datos de APID Interactomes.
    3. **Method**: Método de detección de la interacción. Contiene el identificador del término de la ontología HUPO PSI-MI que representa el método utilizado para detectar la interacción.
    4. **Paper**: Contiene el identificador PubMed de la publicación donde se describe el experimento a partir del cual se ha demostrado la interacción.
    5. **Multiple**: Marcador utilizado para indicar si la interacción binaria descrita procede de una interacción múltiple que ha sido expandida.
    6. **Source**: Contiene un identificador que indica de qué base de datos primaria procede el registro.

7. **SourceId**: Contiene el identificador original utilizado por la base de datos primaria para acceder al registro original.

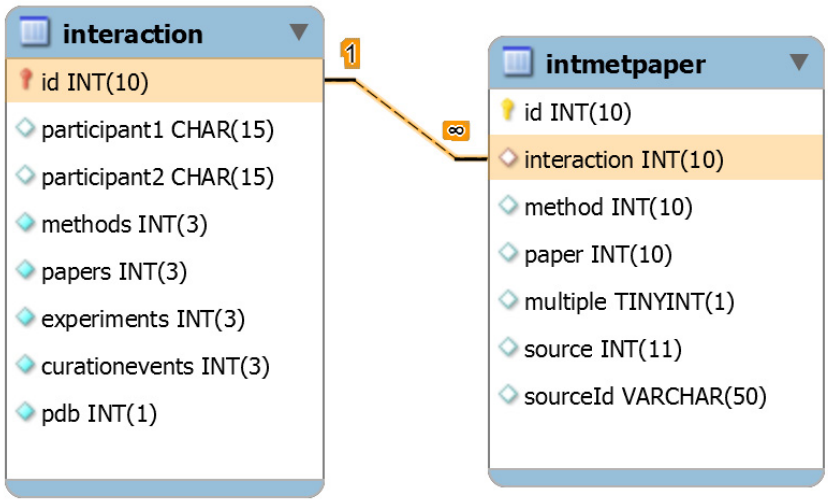


Figura 28. *Extracto del diagrama entidad-relación de la base de datos de APID Interactomes donde pueden verse las tablas que almacenan la información sobre las interacciones únicas y los registros de cada una de ellas en las diferentes bases de datos primarias.*

Esta organización de la información permite establecer un esquema de vista general y detalle con una cardinalidad 1:N entre cada interacción y sus posibles registros. Dicho contexto, mostrado en **Figura 28**, permite también unificar los registros dotando a cada interacción única de una serie de parámetros que permitan evaluar su fiabilidad tal y como se explicará más adelante.

4.1.4.1 Interacciones múltiples

Como ya se mencionó en la sección **Introducción y estado del arte**, algunos de los métodos experimentales utilizados para detectar interacciones entre proteínas generan resultados en los que aparecen varias proteínas interactuando en forma de complejo.

Estos resultados, aunque describen una interacción múltiple, no especifican las interacciones binarias a partir de las cuales se forma dicho complejo. Es por ello que, para poder almacenar las interacciones en formato binario, debe aplicarse un algoritmo de expansión que convierta la interacción múltiple en varias interacciones simples.

La expansión de interacciones múltiples es un proceso necesario pero artificial que, de forma inevitable, genera errores en el registro de interacciones. Como puede verse en **Figura 29**, un hipotético complejo, representado en la zona inferior derecha, podría ser detectado mediante un experimento de tipo TAP-MS como se representa en la parte izquierda de la imagen, con la proteína de color rojo actuando como cebo (*bait*). Los datos que generaría dicho experimento podrían ser expandidos mediante dos métodos diferentes denominados *matrix* y *spoke*. En el primero de ellos, se asume que todas las proteínas interactúan entre ellas y, en consecuencia, se genera un alto

número de falsos positivos en forma de interacciones que se registran, pero no reflejan la realidad. APID Interactomes utiliza el segundo método, *spoke* o *spoke model*, donde se asume que es la proteína cebo la que interacciona con todas las demás. Esta aproximación es la más extendida entre la comunidad científica porque produce menos falsos positivos **(Hakes et al. 2007)** aunque, como en casi cualquier otro caso de detección de patrones, la reducción de falsos positivos puede tener como efecto secundario el incremento de falsos negativos. Así, se admite la posibilidad de perder información fidedigna con el objetivo de reducir la aportación de información errónea.

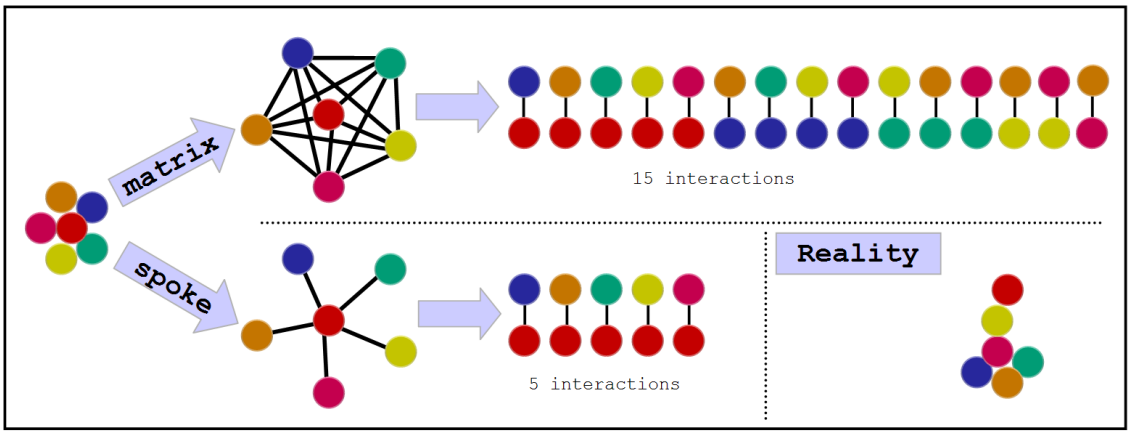


Figura 29. *Esquema extraído de la página web del EMBL-EBI que representa la expansión de interacciones múltiples mediante los algoritmos matrix y spoke, contrastando el resultado que ambos ofrecen con las interacciones que realmente existen en el complejo de proteínas.*

De esta manera, cuando el algoritmo de registro de interacciones detecta que una interacción es múltiple, la expande mediante *spoke model* y envía cada una de las interacciones binarias al procedimiento de registro mencionado anteriormente. Además, marca el campo *multiple* de la tabla *interaction* para que esas interacciones binarias puedan ser identificadas como procedentes de la expansión de una interacción múltiple.

4.1.5 Unificación de interacciones

Después de examinar y almacenar todos los registros válidos de interacciones procedentes de las bases de datos primarias, el algoritmo de APID Interactomes comienza la unificación de dichas interacciones. Esta unificación tiene como objetivo integrar toda la información disponible para cada interacción única descrita en, al menos, una de las bases de datos primarias. Para lograr esto se llevan a cabo, básicamente, tres procedimientos diferentes:

1. **Asignar todos los registros de las bases de datos primarias (*curation events*) a la interacción correspondiente basándose en la combinación única de sus dos participantes**: El primer paso a la hora de unificar los registros obtenidos de las bases de datos primarias es asignar de forma correcta cada uno de ellos a la interacción única correspondiente. Para conseguir esto es clave la actualización de identificadores UniProt que se mencionó en el apartado **Actualización de identificadores** ya que permite asignar, a una misma interacción única en APID Interactomes, registros que en las bases de datos primarias aparecen como interacciones diferentes debido a la ambigüedad a la hora de identificar a sus participantes.

2. **Unificar isoformas**: En algunos registros primarios de interacciones los participantes se identifican como una isoforma concreta de la proteína correspondiente. Esto provoca que dos isoformas de la misma proteína sean tratadas como diferentes participantes y representen, por tanto, dos interacciones diferentes. Incluso, puede darse el caso de que en una interacción se especifique la isoforma concreta y en otra no, contabilizándose de nuevo como dos interacciones diferentes. Para solucionar esto, el algoritmo de APID Interactomes convierte todos los identificadores de isoformas a los identificadores canónicos de la proteína correspondiente. Con ello se evita sobreestimar el número de registros disponibles para una interacción dada sin perder información puesto que siempre se conserva el identificador original de la interacción en la base de datos primaria donde se especifica la isoforma.
3. **Calcular los parámetros de fiabilidad de cada interacción**: Para cada interacción única, APID Interactomes calcula una serie de parámetros que caracterizan dicha interacción y permiten asignarle mayor o menor fiabilidad experimental. Estos parámetros, enumerados ya en la sección **Registro de interacciones** y que se muestran en la **Figura 30**, son los siguientes:
    a. **Número de experimentos**: Se calcula como el número de ocurrencias únicas del par de datos <*identificador de publicación (PMID), identificador del método de detección de interacción (ontología HUPO PSI-MI)*>. Se trata de la métrica principal que APID Interactomes ofrece para filtrar y priorizar interacciones como se verá más adelante. Lo que se intenta con esta medida es representar el *quién* y el *cómo* de cada prueba experimental que aparece en la literatura científica describiendo una interacción particular. De esta manera, la interacción única correspondiente en APID Interactomes acumulará fiabilidad a medida que sea demostrada por diferentes investigadores y/o diferentes métodos.
    b. **Número de métodos**: Se trata del número de métodos distintos, representados por el identificador del término de la ontología HUPO PSI-MI correspondiente, a través de los cuales ha sido probada una interacción.
    c. **Número de publicaciones**: En este caso se representa el número de artículos científicos, mediante su identificador de PubMed, que describen la interacción.
    d. **Número de registros en bases de datos primarias**: Contabiliza el número de veces que una interacción ha sido registrada en cualquiera de las bases de datos primarias procesadas por el algoritmo de APID Interactomes. Estos registros se almacenan como *curation events* y representan el hecho de que un equipo especializado de *curators* de una base de datos de interacciones de proteínas haya decidido registrar una nueva interacción aparecida en la literatura científica.
    e. **Número de estructuras 3D**: Especifica el número de estructuras 3D procedentes de PDB/PDBsum que han sido asignadas a la interacción.

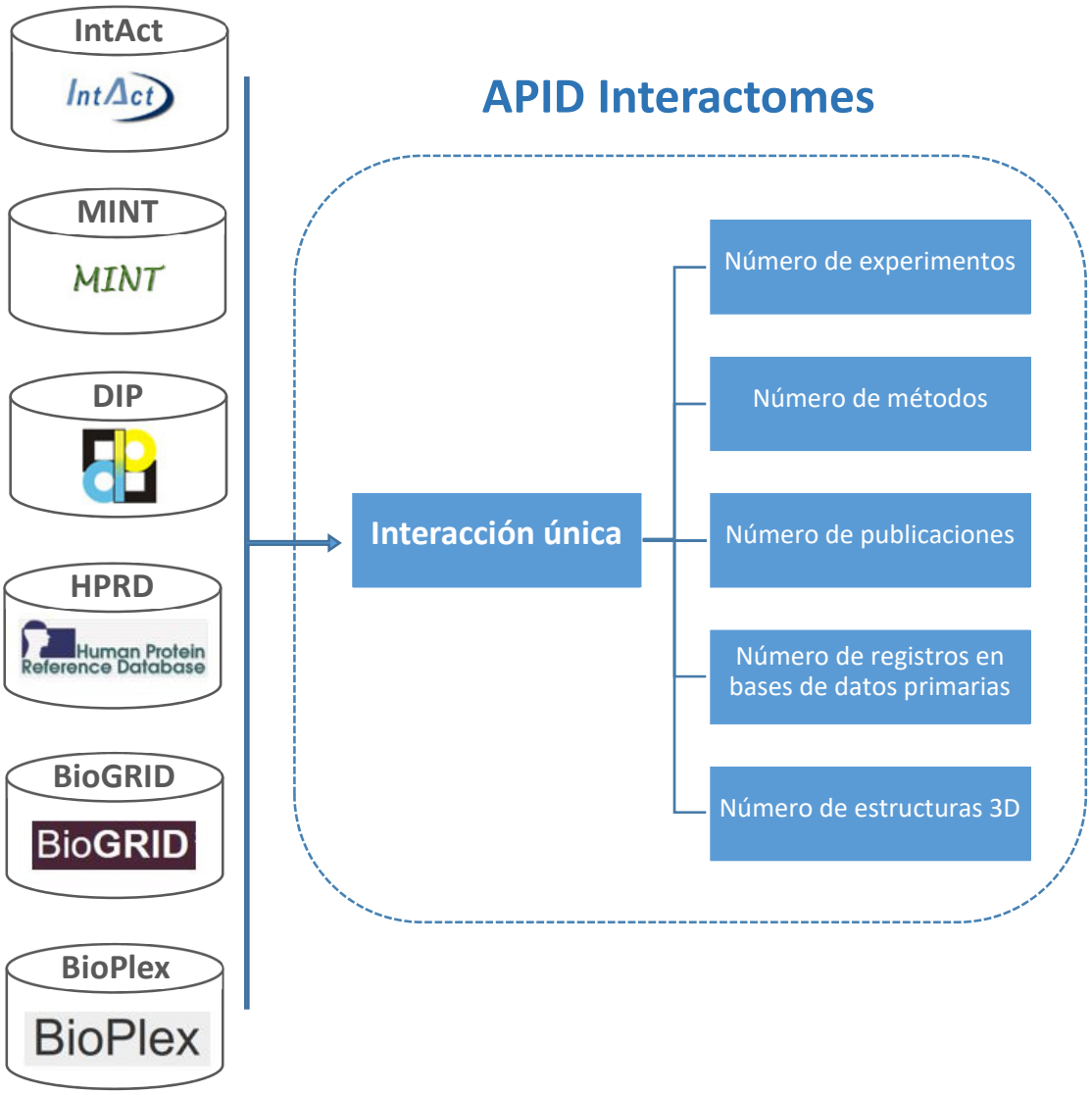


Figura 30. *Diseño esquemático del proceso de unificación de interacciones. A partir de los diferentes registros disponibles en las bases de datos primarias, APID Interactomes genera una interacción única con diversos parámetros de fiabilidad.*

4.1.6 Asignación de estructuras 3D

Una vez unificadas las interacciones disponibles en las bases de datos primarias, el algoritmo de APID Interactomes pasa a examinar las estructuras 3D disponibles y a asignar estas a las interacciones correspondientes. Para ello se utilizan una serie de scripts escritos en SQL que, partiendo de los datos extraídos de los ficheros de la base de datos PDB, asignan identificadores de dicha base datos a las interacciones unificadas que aparezcan en cada uno de ellos.

Los datos iniciales extraídos de los ficheros con las estructuras 3D registradas en PDB se almacenan en dos tablas diferentes, una con las características de cada interfaz y otra con las equivalencias entre cadenas peptídicas e identificadores de la base de datos UniProt.

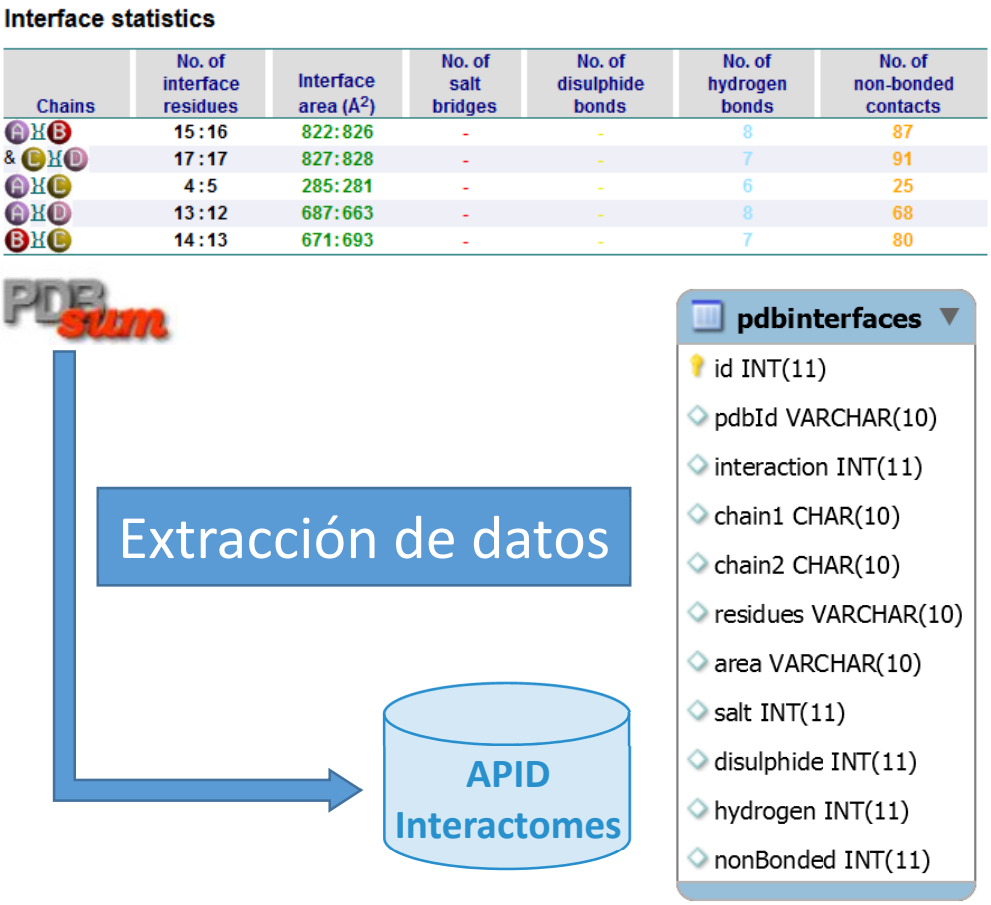


Figura 31. *Esquema de la extracción de datos de la web de PDBsum. APID Interactomes almacena en una tabla todos las interfaces que PDBsum interpreta para cada fichero de la base de datos PDB.*

En **Figura 31** se presenta la información que el algoritmo extrae de la página web de PDBSum **(De Beer et al. 2014)** (29) relativa a los interfaces presentes en cada estructura 3D examinada. Estas interfaces pueden representar enlaces químicos de tres tipos: puentes salinos, enlaces disulfuros o puentes de hidrógeno. Cada uno de estos posibles enlaces se da entre dos cadenas peptídicas que se etiquetan con una letra - A, B, C, etc. – y que posteriormente deben ser identificados.

Para poder asignar las diferentes cadenas presentes en los interfaces a una proteína concreta, el algoritmo extrae de la página web de PDBsum dicha información tal y como se representa en **Figura 32**.

A partir de estas dos tablas, un script SQL enlaza identificadores de proteínas (UniProt), estructuras tridimensionales (PDB) e interacciones (APID Interactomes) para establecer las relaciones correspondientes y poder calcular el valor de la métrica referente a las estructuras para cada interacción única.

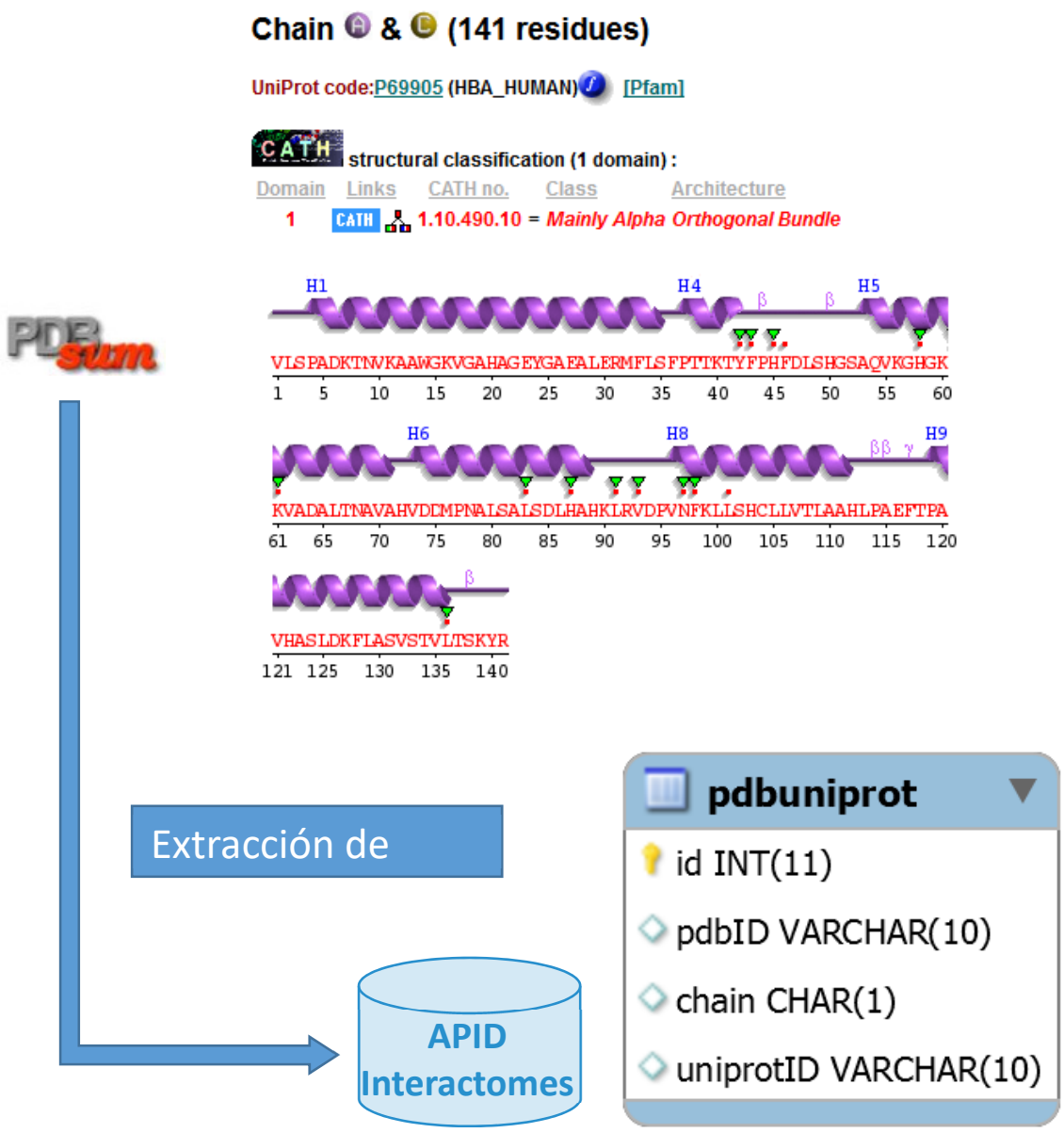


Figura 32. *Esquema de la extracción de datos de la web de PDBSum. APID Interactomes almacena en una tabla todas las identificaciones con código UniProt de cadenas peptídicas que PDBsum almacena para cada estructura 3D.*

4.1.7 Construcción de interactomas

Uno de los objetivos principales de APID Interactomes es el de ofrecer a la comunidad científica los interactomas de aquellas especies para las que han sido descritas interacciones entre sus proteínas. Además, estos interactomas deben tener asociado un nivel de calidad o fiabilidad en las interacciones que contienen.

A continuación, se describen algunos de los aspectos más importantes relacionados con la construcción de los interactomas.

4.1.7.1 Parámetros de calidad

Gracias al registro de interacciones de diferentes bases de datos primarias y al proceso posterior de unificación, es posible establecer unos niveles de calidad controlados basándose en los parámetros derivados de esta última. Así, APID Interactomes establece tres niveles globales de calidad sobre los que generar los interactomas. Para ello, define los requisitos que deben cumplir las interacciones que se incluyan en cada uno de dichos interactomas.

Los requisitos para cada uno de los tres niveles de calidad son los siguientes:

- **Nivel 1**: Incluye todas las interacciones conocidas. No hay ningún tipo de filtrado.
- **Nivel 2**: Agrupa todas las interacciones que presenten 2 o más *experimentos*. Esto es, que en su proceso de unificación se hayan encontrado al menos dos pares distintos <*publicación, método de detección de la interacción*>. Esto incluirá las interacciones descritas en diferentes artículos y con diferentes métodos, pero también aquellas descritas en diferentes artículos con igual método o en el mismo artículo con diferente método. Es decir, debe existir un mínimo de heterogeneidad al menos en uno de los parámetros.
- **Nivel 3**: Contiene todas las interacciones que hayan sido descritas en al menos 2 publicaciones. Este requisito es el más riguroso de los tres puesto que solo considerará las interacciones que hayan sido descritas al menos dos veces en diferente publicación, independientemente del método de detección utilizado. Es decir, una interacción descrita en una publicación que haya sido probada por un único grupo con varios métodos diferentes no entraría en un interactoma de nivel 3.

Esta elección de requisitos está basada en la interpretación de los diferentes parámetros que se calculan en la unificación de interacciones y tiene como estrategia principal la valoración de *cuantos grupos diferentes* y *con cuantos métodos diferentes* en lo que a la demostración experimental de cada interacción se refiere. Así, para cada nivel de calidad, se obtiene un tamaño de interactoma concreto y su correspondiente cobertura, tal y como se detalla a continuación.

4.1.7.2 Cobertura sobre los proteomas

En **Figura 33** puede verse la cobertura de diferentes interactomas, para seis organismos modelo, en los tres niveles de calidad mencionados anteriormente. Esta cobertura está calculada sobre el proteoma de referencia, es decir, se extrae el número de proteínas diferentes que participan en las interacciones contenidas para cada nivel de calidad de un interactoma dado y se representa la proporción de estas frente a las descritas en el proteoma del organismo correspondiente.

Los proteomas utilizados, como se describió en la sección **Obtención de proteomas**, provienen del proyecto UniProt Proteomes cuando el organismo está presente en dicha base de datos o de un cálculo propio basado en el número de proteínas asignadas a un organismo dado, tanto en SwissProt como en TrEMBL.

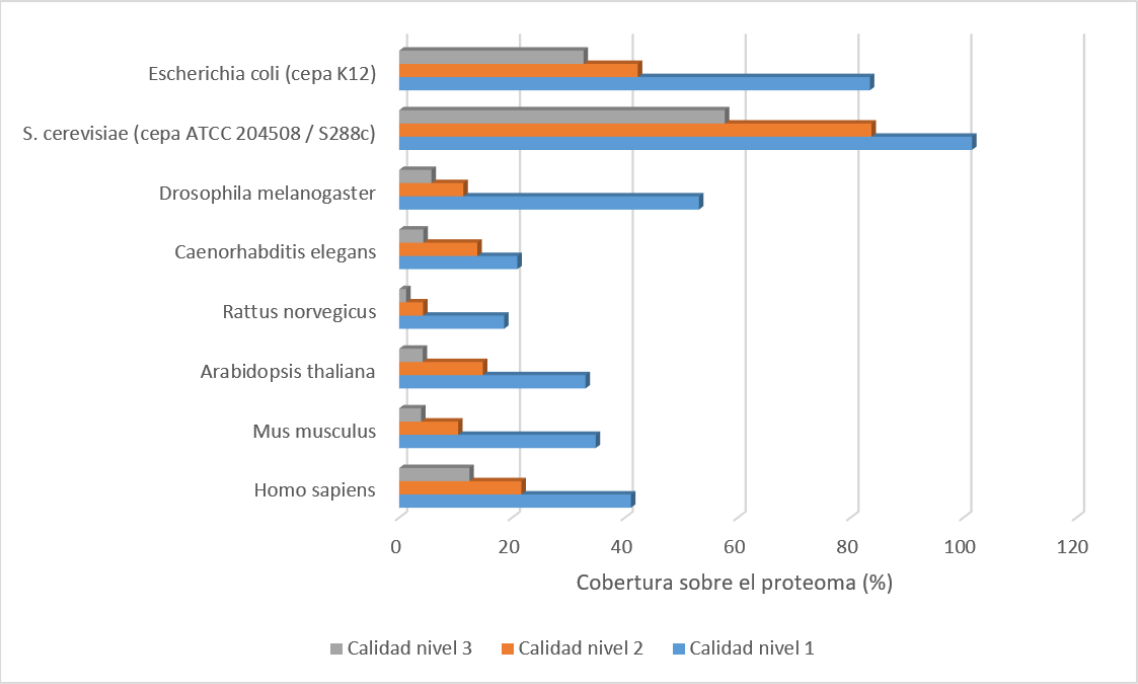


Figura 33. *Cobertura sobre el proteoma de los interactomas de algunos organismos modelo para los tres niveles de calidad disponibles.*

En el caso del organismo *S. cerevisiae (cepa ATCC 204508 / S288c)*, la cobertura es ligeramente superior al 100% (101.52%) debido a que el proteoma de referencia, en este caso disponible en UniProt Proteomes, contiene un número inferior de proteínas si se comparan con las que el algoritmo de APID Interactomes ha detectado entre todas las interacciones descritas para dicho organismo. Esto es debido a que existen interacciones cuyos participantes son proteínas que están registradas en TrEMBL y asignadas al organismo *S. cerevisiae (cepa ATCC 204508 / S288c)*, pero no contenidas en el proteoma registrado en UniProt Proteomes para el mismo organismo.

En **Tabla 4** pueden verse los datos y coberturas de los interactomas de seis organismos modelo para las tres calidades disponibles en APID Interactomes. Además, se representa la cobertura media que pasa de aproximadamente un 50% en el nivel sin filtrado a un 15% cuando se aplica para estos organismos el filtro más exigente.

Tabla 4. *Cobertura de los interactomas de seis organismos modelo para cada uno de los tres niveles de calidad disponibles en APID Interactomes.*

		Nivel 1 (todas las interacciones)			Nivel 2 (al menos 2 experimentos)			Nivel 3 (al menos dos publicaciones)		
	Proteínas en el PROTEOMA	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el proteoma (%)	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el proteoma (%)	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el proteoma (%)
Homo sapiens	69 986	**28 749**	**334 817**	41.08	**15 160**	**88 864**	21.66	**8 706**	**32 092**	12.44
Mus musculus	50 189	**17 478**	**53 676**	34.82	**5 243**	**7 897**	10.45	**1 950**	**2 489**	3.89
Arabidopsis thaliana	31 477	**10 394**	**40 018**	33.02	**4 668**	**9 716**	14.83	**1 309**	**1 960**	4.16
Rattus norvegicus	29 885	**5 556**	**8 797**	18.59	**1 265**	**1 244**	4.23	**366**	**326**	1.22
Caenorhabditis elegans	26 596	**5 569**	**14 762**	20.94	**3 683**	**5 849**	13.85	**1 138**	**1 327**	4.28
Drosophila melanogaster	22 005	**11 692**	**59 977**	53.13	**2 505**	**3 590**	11.38	**1 258**	**1 336**	5.72
S. cerevisiae (cepa ATCC 204508 / S288c)	6 721	**6 823**	**128 740**	101.52	**5 626**	**43 039**	83.71	**3 878**	**19 325**	57.70
Escherichia coli (strain K12)	4 252	**3 547**	**25 335**	83.42	**1 798**	**6 967**	42.29	**1 389**	**4 020**	32.67
		Cobertura media sobre el PROTEOMA (%)		**48.32**			**25.30**			**15.26**

4.1.7.3 Generación de ficheros

La construcción de un interactoma a partir de las interacciones almacenadas en la base de datos es un proceso computacionalmente costoso que hace inviable su ejecución bajo demanda a través de la aplicación web. Por ello, una de las fases finales del protocolo que se describe en esta sección es la construcción y el almacenamiento en ficheros de todos los interactomas disponibles en la base de datos de APID Interactomes.

Para llevar a cabo este proceso se escribió un programa en Java que, para cada organismo con interacciones en la base de datos, genera tres interactomas según los niveles de calidad descritos anteriormente y los almacena en tres ficheros de texto tabulado diferentes. De esta manera, la aplicación web solo necesita referenciar el archivo adecuado para cada petición del usuario.

La nomenclatura seguida para nombrar estos ficheros contiene el organismo, mediante su identificador numérico de la ontología de taxonomías, y el nivel de calidad del interactoma, mediante el uso de las etiquetas Q1, Q2 o Q3. Así, para los interactomas de *Homo Sapiens* se generan tres ficheros con nombres: 9606_Q3.txt, 9606_Q2.txt y 9606_Q1.txt.

```
1  InteractionID   UniprotID_A UniprotName_A   GeneName_A  UniprotID_B UniprotName_B   GeneName_B
   Experiments Methods Publications    3DStructures    CurationEvents
2  2627577 A0A087WX54  A0A087WX54_HUMAN    ZNF385C Q9NVV9  THAP1_HUMAN THAP1   2   1   2   0   2
3  2627578 A0A087WX54  A0A087WX54_HUMAN    ZNF385C P25788  PSA3_HUMAN  PSMA3   2   1   2   0   2
4  2899154 A0A087X2D5  A0A087X2D5_HUMAN    MRPL45  Q9HD33  RM47_HUMAN  MRPL47  2   2   2   0   2
5  2973930 A0AVI4  TM129_HUMAN TMEM129 Q9BUN8  DERL1_HUMAN DERL1   2   1   2   0   3
6  2973936 A0AVI4  TM129_HUMAN TMEM129 P04439  1A03_HUMAN  HLA-A   2   1   2   0   2
7  2820171 A0AVT1  UBA6_HUMAN  UBA6    P62837  UB2D2_HUMAN UBE2D2  3   2   3   0   3
8  2880257 A0AVT1  UBA6_HUMAN  UBA6    P0CG48  UBC_HUMAN   UBC 4   4   3   0   4
9  2880258 A0AVT1  UBA6_HUMAN  UBA6    O15205  UBD_HUMAN   UBD 3   3   2   0   3
10 2880720 A0AVT1  UBA6_HUMAN  UBA6    Q9H832  UBE2Z_HUMAN UBE2Z   6   3   5   0   7
11 2900246 A0AVT1  UBA6_HUMAN  UBA6    P15374  UCHL3_HUMAN UCHL3   2   1   2   0   2
12 2900585 A0AVT1  UBA6_HUMAN  UBA6    Q00341  VIGLN_HUMAN HDLBP   2   1   2   0   2
13 2905317 A0AVT1  UBA6_HUMAN  UBA6    Q86UV5  UBP48_HUMAN USP48   2   1   2   0   2
14 2937584 A0AVT1  UBA6_HUMAN  UBA6    P17812  PYRG1_HUMAN CTPS1   2   1   2   0   2
15 2947929 A0AVT1  UBA6_HUMAN  UBA6    P68036  UB2L3_HUMAN UBE2L3  2   2   2   0   2
```

***Figura 34**. Captura de pantalla de las quince primeras líneas del fichero de texto tabulado con el interactoma completo de Homo Sapiens en el nivel de calidad 1 (9606_Q1.txt).*

Cada fichero contiene el conjunto de interacciones que forman el interactoma del organismo para el nivel de calidad correspondiente. La información que se escribe en el fichero sobre cada una de las interacciones consiste en una línea de texto con todos los parámetros calculados en el proceso de unificación. Las columnas del fichero son las siguientes:

1. **InteractionID**: Identificador numérico interno de la base de datos de APID Interactomes para la interacción.
2. **UniprotID_A**: Identificador UniProt de la primera proteína participante en la interacción.
3. **UniprotName_A**: Nombre UniProt de la primera proteína participante en la interacción.
4. **GeneName_A**: Nombre del gen asociado a la primera proteína participante en la interacción.
5. **UniprotID_B**: Identificador UniProt de la segunda proteína participante en la interacción.
6. **UniprotName_B**: Nombre UniProt de la segunda proteína participante en la interacción.

7. **GeneName_B**: Nombre del gen asociado a la segunda proteína participante en la interacción.
8. **Experiments**: Número de experimentos calculados en el proceso de unificación para la interacción.
9. **Methods**: Número de métodos de detección de interacción diferentes calculados en el proceso de unificación para la interacción.
10. **Publications**: Número de publicaciones diferentes calculados en el proceso de unificación para la interacción.
11. **3DStructures**: Número de estructuras 3D asignadas a la interacción.
12. **CurationEvents**: Número de veces que la interacción ha sido registrada en una base de datos primaria.

Estos ficheros están especialmente pensados para ser compatibles con la mayoría del software de análisis de redes disponible. Son ficheros de texto tabulados que pueden visualizarse con cualquier editor de texto e importarse sin problema en software de análisis de redes como por ejemplo Cytoscape **(Shannon et al. 2003)**.

4.1.7.4 Filtrado de interacciones inter-especies

Para la mayoría de los organismos más estudiados, aparecen descritas en la literatura científica interacciones entre alguna de sus proteínas y otras proteínas de diferentes organismos.

Se trata de un hecho conocido y asumido hasta el punto de que, cuando se habla del interactoma de un organismo concreto, normalmente se incluyen también estas interacciones inter-especie. Es por ello que en los interactomas que APID Interactomes construye también han sido incluidas.

Sin embargo, y de manera adicional, se proporcionan esos mismos interactomas sin interacciones inter-especie. El usuario puede indicar esto desde la aplicación web y, en ese caso, descargará un fichero diferente que, al igual que los anteriores, ha sido previamente generado.

Para disponer de esta posibilidad, el algoritmo de APID Interactomes encargado de construir los interactomas genera tres ficheros adicionales para cada organismo a cuyo nombre añade la etiqueta "noISI". De esta manera, para el caso del organismo *Homo Sapiens*, tendríamos finalmente los siguientes seis ficheros: 9606_Q3.txt, 9606_noISI_Q3.txt, 9606_Q2.txt, 9606_noISI_Q2.txt, 9606_Q1.txt y 9606_noISI_Q1.txt.

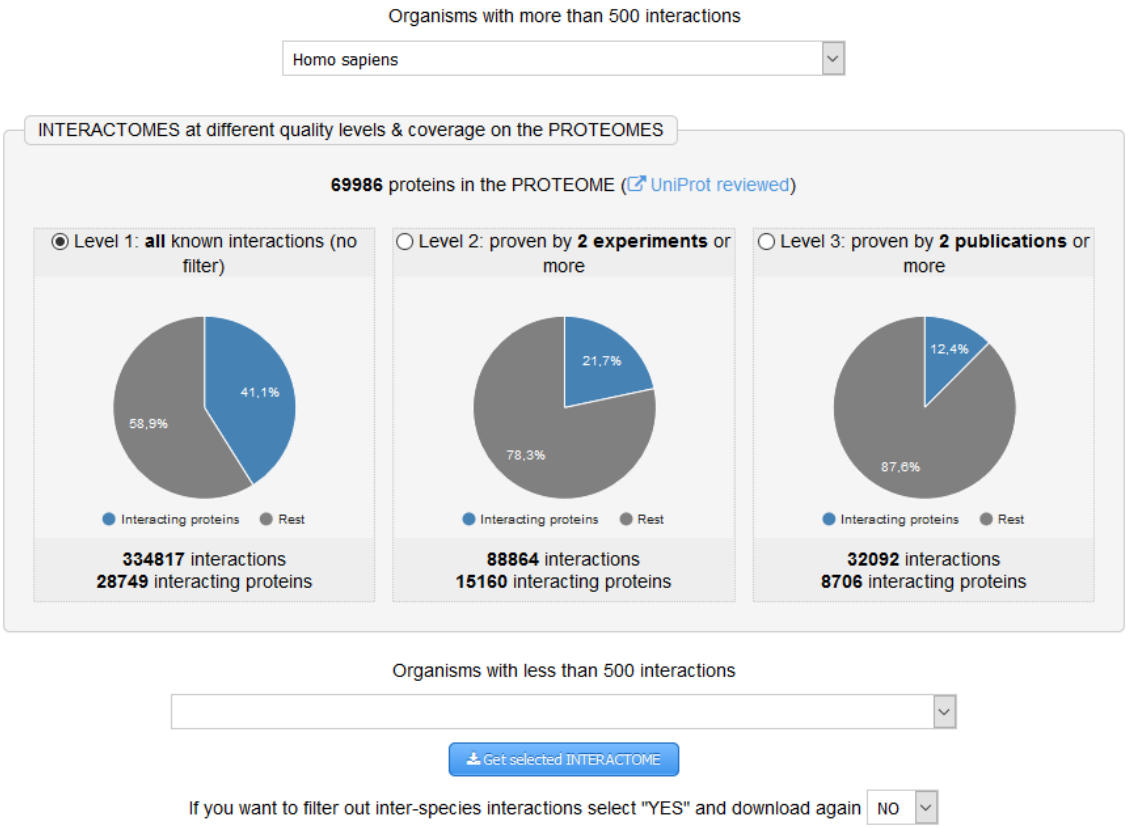


Figura 35. *Captura de la aplicación web de APID Interactomes. En ella puede observarse la cobertura de los tres interactomas de Homo Sapiens sobre su proteoma. También pueden verse los sistemas de selección de organismo, nivel de calidad del interactoma y filtrado de interacciones inter-especie.*

4.1.8 Anotación de proteínas

La última fase del protocolo para la integración de datos de interacciones de proteínas está dedicada a la recopilación de información sobre estas. Una vez se han capturado y unificado todas las interacciones y se han construido los interactomas correspondientes, el siguiente objetivo es contextualizar dichas interacciones a nivel funcional. Para ello, se recurre a la asignación de etiquetas a cada proteína procedentes de diferentes espacios de anotación, cada uno de ellos con su ontología y vocabulario controlado correspondiente.

Los espacios de anotación que se utilizan corresponden a bases de datos que son referencia en diferentes ámbitos:

1. **Anotación funcional**: Gene Ontology **(Ashburner et al. 2000; Blake et al. 2015)** (30).
2. **Estructura de proteínas**: InterPro **(Mitchell et al. 2015)** (35) y Pfam **(Finn et al. 2015)** (34).
3. **Rutas y señalización**: Reactome **(Croft et al. 2014)** (33).

La asignación de anotaciones a cada proteína se hace a partir de los datos disponibles en UniProt. El algoritmo utiliza el mismo procedimiento que lleva a cabo para la extracción de los

identificadores, es decir, analiza de forma secuencial, y orientada a eventos, los ficheros XML procedentes tanto de SwissProt como de TrEMBL.

La diferencia es que, en esta ocasión, lo que se procesa son las entidades que contienen anotaciones sobre la proteína correspondiente. Estas entidades XML son de tipo *referencia cruzada* y apuntan a los espacios de anotación representados por las bases de datos antes mencionadas.

El objetivo de este proceso de anotación es, como se verá más adelante, proporcionar al usuario un entorno funcional asociado a cada red que este construya. De esta manera, la función aislada de una proteína pueda interpretarse mejor contextualizándola en base a las anotaciones de las proteínas que interactúan con ella y, por tanto, pertenecen a dicha red.

A continuación, se detallan las entidades recogidas para cada espacio de anotación.

4.1.8.1 Gene Ontology

Para la anotación funcional de las proteínas se recurrió a la base de datos Gene Ontology **(Ashburner et al. 2000; Blake et al. 2015)** (30). Como ya se introdujo al principio de este documento, Gene Ontology está compuesta por tres espacios diferentes de anotación que pueden ser combinados usándose sus etiquetas de la misma manera.

```
− <dbReference type="GO" id="GO:0005886">
    <property type="term" value="C:plasma membrane"/>
    <property type="evidence" value="ECO:0000314"/>
    <property type="project" value="UniProtKB"/>
  </dbReference>
− <dbReference type="GO" id="GO:0005525">
    <property type="term" value="F:GTP binding"/>
    <property type="evidence" value="ECO:0000314"/>
    <property type="project" value="UniProtKB"/>
  </dbReference>
```

Figura 36. *Anotaciones funcionales con etiquetas de la base de datos Gene Ontology. Extraídas del fichero P01112.xml de UniProt, correspondiente a la proteína RASH_HUMAN.*

En **Figura 36** puede verse un extracto del fichero XML correspondiente a la proteína RASH_HUMAN que contiene dos referencias a términos de la base de datos de Gene Ontology. Como puede observarse, estas entidades contienen información sobre:

1. El identificador único de la etiqueta o término en base de datos: *GO:005886* o *GO:0005525*
2. La definición corta de dicha etiqueta: *C:Plasma membrane* o *F:GTP binding*. Aquí, el prefijo *C* o *F* indica a que espacio de anotación de los tres disponibles en Gene Ontology pertenece dicho término (*Cellular Component* y *Molecular Function* respectivamente).
3. El tipo de evidencia: En este caso, las evidencias son de tipo *ECO:0000314*, ya que es el dedicado a aquellas evidencias procedentes de experimentos, pero asignadas manualmente.

4. La información sobre quién asigna la etiqueta: Aparece en la última línea, tratándose en ambos casos de la propia base de datos UniProtKB.

El algoritmo de APID Interactomes, a través del uso de este tipo de entidades, construye una tabla en la base de datos que relaciona cada identificador UniProt que representa una proteína con los diferentes identificadores de tipo GO extraídos de los ficheros XML.

4.1.8.2 InterPro

En cuanto a las anotaciones relacionadas con la estructura de las proteínas, se utilizaron dos bases de datos: InterPro **(Mitchell et al. 2015)** (35) y Pfam **(Finn et al. 2015)** (34).

InterPro es una base datos de familias, dominios y sitios funcionales que extrae e integra información de otras muchas bases de datos, incluída Pfam.

```
−<dbReference type="InterPro" id="IPR027417">
    <property type="entry name" value="P-loop_NTPase"/>
  </dbReference>
−<dbReference type="InterPro" id="IPR005225">
    <property type="entry name" value="Small_GTP-bd_dom"/>
  </dbReference>
−<dbReference type="InterPro" id="IPR001806">
    <property type="entry name" value="Small_GTPase"/>
  </dbReference>
```

Figura 37. *Anotaciones estructurales con etiquetas de la base de datos InterPro. Extraídas del fichero P01112.xml de UniProt correspondiente a la proteína RASH_HUMAN.*

En **Figura 37** pueden verse el tipo de entidades referentes a InterPro que están contenidas en los ficheros XML de UniProt. A partir de estas entidades el algoritmo de APID Interactomes extrae el identificador propio de la base de datos InterPro – por ejemplo, *IPR027417* en el primer caso – y la descripción corta de dicha etiqueta.

4.1.8.3 Pfam

Como espacio de anotación adicional en el ámbito estructural se utilizó la base de datos Pfam **(Finn et al. 2015)**.

```
−<dbReference type="Pfam" id="PF00071">
    <property type="entry name" value="Ras"/>
    <property type="match status" value="1"/>
  </dbReference>
```

Figura 38. *Anotaciones estructurales con etiquetas de la base de datos Pfam. Extraídas del fichero P01112.xml de UniProt correspondiente a la proteína RASH_HUMAN.*

En **Figura 38** puede verse una entidad XML que contiene una etiqueta de anotación perteneciente a la base de datos Pfam. En ella aparece el identificador de la etiqueta correspondiente, así como

su descripción corta. También contiene una línea adicional que indica el número de *hits* para el dominio representado por la etiqueta correspondiente.

4.1.8.4 Reactome

Por último, en cuanto a rutas de señalización y *pathways* se utilizó la base de datos Reactome **(Croft et al. 2014)** (33). En **Figura 39** pueden observarse dos entidades con etiquetas procedentes de dicha base de datos. De nuevo, la información contenida en estas etiquetas incluye el identificador único de la etiqueta en la base de datos y el texto corto de descripción de dicha etiqueta.

```
− <dbReference type="Reactome" id="R-HSA-112412">
    <property type="pathway name" value="SOS-mediated signalling"/>
  </dbReference>
− <dbReference type="Reactome" id="R-HSA-1169092">
    <property type="pathway name" value="Activation of RAS in B cells"/>
  </dbReference>
```

Figura 39. *Anotaciones de rutas de señalización con etiquetas de la base de datos Reactome. Extraídas del fichero P01112.xml de UniProt, correspondiente a la proteína RASH_HUMAN.*

4.2 Plataforma bioinformática de acceso *web* para la exploración y generación de conjuntos de interacciones de proteínas

Una vez generada la base de datos de interacciones proteína-proteína, se procedió a diseñar e implementar una plataforma bioinformática de acceso web que proporcionara a la comunidad científica la posibilidad de explorar los datos contenidos en ella de una forma intuitiva y eficaz. Puede accederse a dicha plataforma en la URL http://apid.dep.usal.es.

Para conseguir esto, se estudiaron los casos de uso más habituales para este tipo de datos, desde la obtención directa del interactoma de un organismo concreto, con una calidad determinada, hasta la consulta de todas las interacciones registradas para una única proteína de interés, y se diseñó una aplicación que se adaptase de forma satisfactoria a dichos escenarios.

4.2.1 Arquitectura del sistema

Uno de los objetivos más importantes de este trabajo de investigación ha sido el de tratar de poner a disposición de la comunidad científica los datos generados de la forma más eficaz posible. Es por esto que, desde un principio, se optó por desarrollar una aplicación web en vez de una aplicación de escritorio. Aunque esta última opción ofrecía mucha más potencia y flexibilidad en lo que a interfaz de usuario se refiere, la inmediatez del acceso a la aplicación web hizo que se seleccionara dicha alternativa.

Así, la aplicación desarrollada se basa en el modelo cliente-servidor, donde las tareas se reparten entre los proveedores de recursos o servicios, llamados servidores, y los demandantes, llamados clientes. Un caso concreto de este tipo de arquitectura son los servicios web, donde históricamente la parte cliente consistía en un visualizador especializado de determinados documentos (estáticos o dinámicos) que la parte servidor proveía, asumiendo esta última casi toda la carga computacional.

En los últimos años se ha producido una pequeña revolución en la parte cliente con el desarrollo de múltiples tecnologías orientadas a enriquecer las capacidades de estos visualizadores especializados conocidos como navegadores web. Las nuevas especificaciones HTML 5 y CSS 3, unidas a la proliferación de librerías multipropósito basadas en JavaScript y la mayor capacidad computacional disponible en el lado del cliente, hacen posible el desarrollo de aplicaciones web complejas que ya nada tienen que ver con la mera representación de los documentos generados por un servidor. Hoy en día, muchas aplicaciones web tienen una implementación en el lado del cliente con un grado de complejidad muy similar a la del servidor.

La aplicación desarrollada en esta Tesis Doctoral es un claro ejemplo de esta nueva tendencia ya que, como se describirá a lo largo de este capítulo y el siguiente, la lógica implementada en el lado del cliente gestiona en muchos casos tanta información como la presente en el lado del servidor.

La parte cliente de la aplicación APID Interactomes es lo que el usuario realmente percibe como aplicación ya que es la única parte visible para este. Así, el objetivo que se planteó al inicio del desarrollo fue el de diseñar una interfaz ágil y sencilla que proporcionara al investigador un acceso rápido a los datos generados por el algoritmo de integración de interacciones proteína-proteína. Además, estos datos debían proporcionarse como información ordenada, con la posibilidad de ser filtrada y, en ciertos casos, agrupada.

Para conseguir esto se recurrió a la plataforma STRUTS, pero también a diversas librerías orientadas al tratamiento de la información y al diseño de interfaces complejas en el lado cliente, similares estos últimos a los que se pueden implementar en las aplicaciones de escritorio.

Con el objetivo de establecer diferentes *caminos* para llegar a la información contenida en la base de datos, se estudiaron los diferentes contextos en los que un investigador puede necesitar datos de interacción de proteínas. Así, se definieron cuatro puntos de partida posibles utilizando como criterio la información de la que dispone el investigador inicialmente y lo que APID Interactomes podía aportar a partir de esta. Son los siguientes:

1. **Un organismo**: el investigador quiere estudiar un organismo específico a nivel global y para ello necesita acceso a su interactoma completo.
2. **Una proteína**: el investigador trabaja sobre una única proteína de interés y quiere acceder a todas las interacciones de esta.
3. **Una lista de proteínas**: el investigador dispone de una lista de proteínas de interés y necesita ver cómo se relacionan entre ellas.
4. **Un artículo científico**: en este caso el contexto de trabajo es el conjunto de interacciones que se han descrito en un artículo específico. Aquí APID Interactomes proporcionará dichas interacciones, pero con toda la información adicional que sobre ellas contiene procedente de su algoritmo de integración.

De esta manera, se diseñaron interfaces específicas para cada uno de estos escenarios. A continuación, se describe cada una de ellas con más detalle.

4.2.2 Interactomas de un organismo específico

El primero de los escenarios es aquel en el que el investigador necesita toda la información disponible en APID Interactomes sobre las interacciones entre las proteínas de un organismo concreto, esto es, su interactoma.

El interactoma de un organismo es el conjunto de datos más importante que APID Interactomes proporciona ya que su generación es fruto de la integración de todos los datos disponibles en las bases de datos primarias que procesa y, además, para cada interacción única, lleva asociadas las diferentes métricas de calidad que APID Interactomes calcula.

En **Figura 40** puede verse la interfaz diseñada para que el investigador pueda seleccionar el interactoma que quiere descargar. Como ya se ha explicado anteriormente, APID Interactomes ofrece los interactomas bajo tres niveles de calidad que se sustentan en diferentes criterios, más o menos exigentes, de filtrado de interacciones. Esto influye de forma significativa en el número de interacciones que contiene cada interactoma y, por tanto, en la cobertura que este ofrece sobre el proteoma del organismo. Lo que se ha querido conseguir con esta interfaz es que el investigador pueda identificar visualmente, y de forma sencilla, la cobertura de los diferentes interactomas en cada nivel de calidad.

La parte servidor de APID Interactomes envía las estadísticas pre-calculadas de los interactomas de cada especie en sus tres niveles de calidad y estas se visualizan a través de las gráficas correspondientes tal y como se puede observar en **Figura 40**.

Para seleccionar el organismo de interés se crearon dos listas de selección, una primera con los organismos más estudiados – esto es, con más interacciones descritas en la literatura científica - y otra segunda con el resto. El umbral para separar estos dos grupos de organismos se estableció en 500 interacciones. Lógicamente, el concepto de interactoma como conjunto masivo de todas las interacciones de proteínas descritas para un organismo es demasiado grande para aquellos organismos donde se ha descrito un número pequeño de interacciones, pero por simplicidad y homogeneidad, se decidió mantener dicha referencia.

Además, tal y cómo se explicó anteriormente, los interactomas se construyeron con y sin interacciones inter-especie. Es por esto que la interfaz incluye también un control para que el investigador pueda decidir qué tipo de interactoma quiere descargarse. No obstante, los cálculos de cobertura sobre el proteoma están realizados con la versión sin filtrar, es decir, la que incluye todas las interacciones descritas, intra e inter-especie.

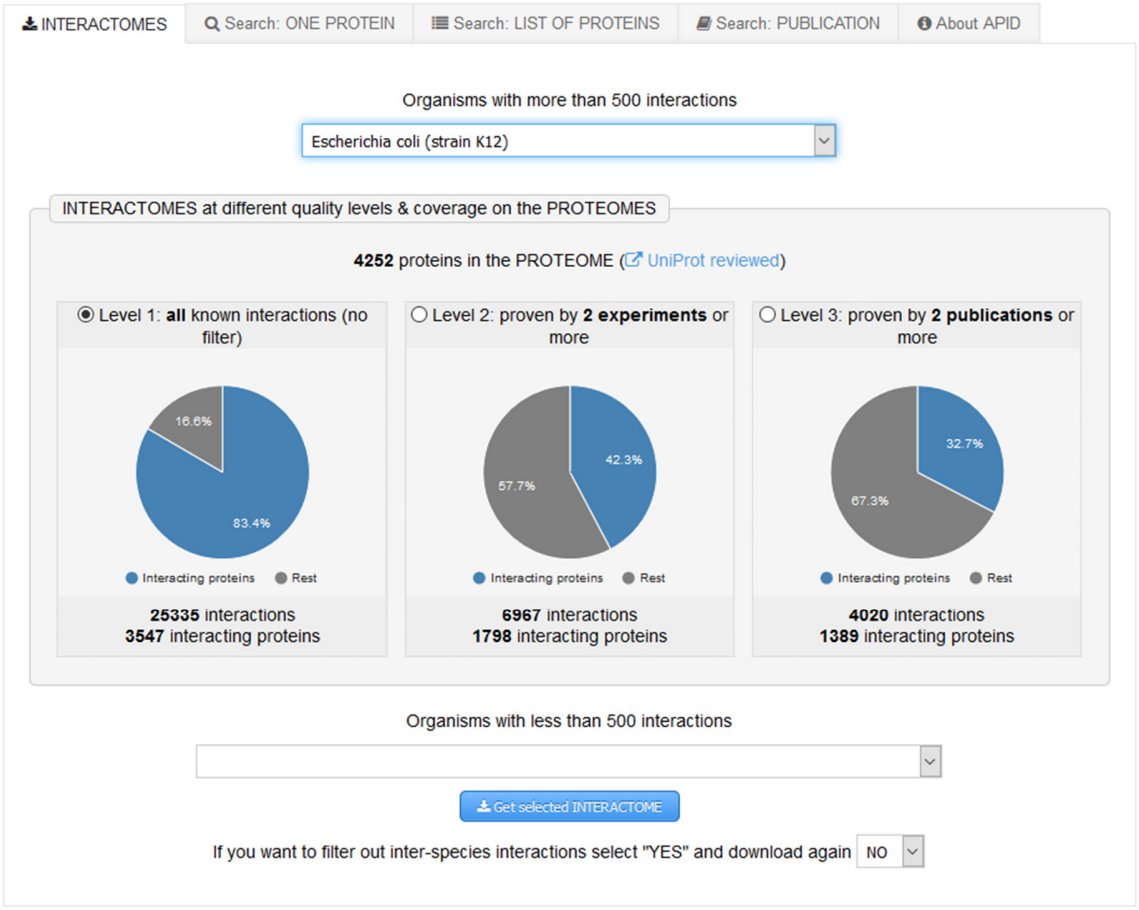


Figura 40. *Interfaz diseñada para el acceso a los interactomas de un organismo específico con diferentes niveles de calidad.*

4.2.3 Interacciones de una proteína específica

En este contexto, el investigador parte de una única proteína de interés y la aplicación le ofrece toda la información disponible sobre sus diferentes interacciones.

Para ello, como puede verse en **Figura 41**, se diseñó un cuadro de búsqueda sencillo donde el usuario puede introducir de forma completa o parcial el nombre UniProt de la proteína, su identificador UniProt AC o el nombre del gen correspondiente.

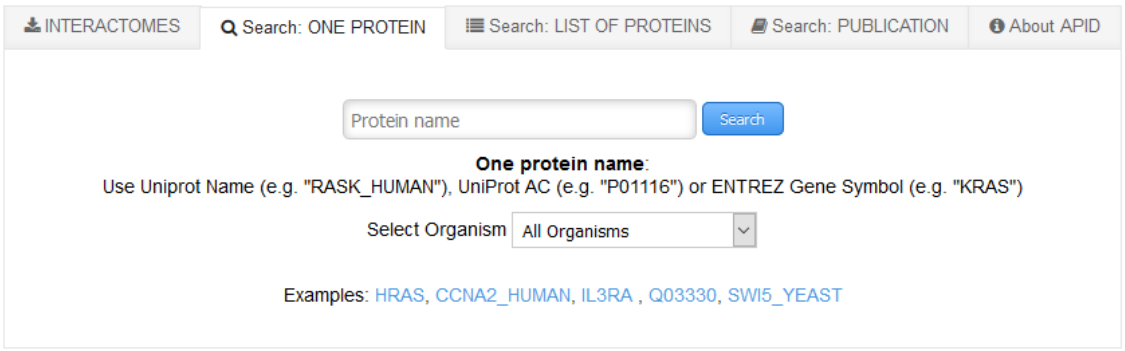


Figura 41. *Cuadro de búsqueda de proteínas.*

Partiendo de esta información, la aplicación presenta una segunda pantalla donde se muestran las proteínas que encajan con la búsqueda planteada por el investigador. En **Figura 42** puede verse la pantalla resultante para la búsqueda textual "HRAS".

6 proteins found for "HRAS"

Showing 1 to 6 of 6 entries — Search:

Uniprot ID	Uniprot Name	Gene Name	Organism	Description	Interactions	[Display RESULTS]
P01112	RASH_HUMAN	HRAS	Homo sapiens	GTPase HRas	165	
Q61411	RASH_MOUSE	Hras	Mus musculus	GTPase HRas	21	
P20171	RASH_RAT	Hras	Rattus norvegicus	GTPase HRas	6	
Q96KN8	HRSL5_HUMAN	HRASLS5	Homo sapiens	Ca(2+)-independent N-acyltransferase	6	
Q86WS9	Q86WS9_HUMAN	HRASLS	Homo sapiens		1	
Q8UUZ4	Q8UUZ4_XENLA	hras	Xenopus laevis		1	

Show 25 entries — Previous 1 Next

Figura 42. *Resultado mostrado a partir de una búsqueda textual para "HRAS".*

Para cada una de las proteínas contenidas en la tabla se ofrecen tres operaciones diferentes: consultar las interacciones únicas ( ), explorar todos los *curation events* registrados en las bases de datos primarias ( ) y generar una red con las interacciones disponibles ( ).

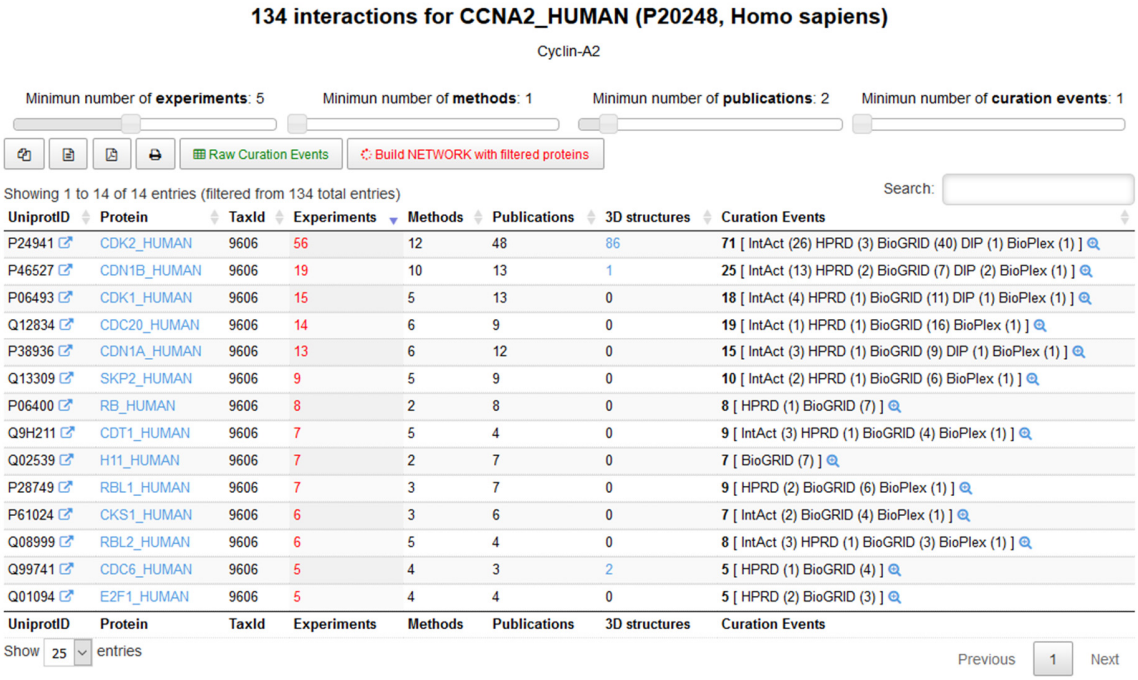
La primera de las opciones disponibles ofrece una nueva tabla con todas las interacciones únicas de dicha proteína y los parámetros calculados por el algoritmo de integración. Así, para cada interacción única en la que participa la proteína de interés, la aplicación muestra el número de experimentos, métodos, publicaciones y estructuras tridimensionales que validan su detección. En **Figura 43** puede verse un ejemplo de esto para la proteína CCNA2_HUMAN.

Además de mostrar dicha información, la interfaz ofrece la posibilidad de filtrar el conjunto de interacciones en tiempo real en base a algunas de estas métricas. Así, la tabla mostrará solo aquellas

interacciones que cumplan con los criterios establecidos pudiéndose utilizar varios de ellos de forma combinada.

La aplicación también permite ordenar las interacciones de forma ascendente o descendente usando cualquiera de las columnas y buscar sobre el contenido de la tabla. Tanto el ordenado como la búsqueda se efectúan sobre los datos que la tabla contiene en cada momento en función de los diferentes filtros que el investigador establezca. Es decir, todos los controles implementados funcionan en tiempo real y de forma simultánea.

De esta manera, se proporciona al investigador una herramienta potente con la que hacer una exploración de datos rápida y eficaz sobre las interacciones disponibles para una proteína específica.



134 interactions for CCNA2_HUMAN (P20248, Homo sapiens)

Cyclin-A2

Minimun number of **experiments**: 5 | Minimun number of **methods**: 1 | Minimun number of **publications**: 2 | Minimun number of **curation events**: 1

Raw Curation Events | Build NETWORK with filtered proteins

Showing 1 to 14 of 14 entries (filtered from 134 total entries) — Search:

UniprotID	Protein	TaxId	Experiments	Methods	Publications	3D structures	Curation Events
P24941	CDK2_HUMAN	9606	56	12	48	86	**71** [IntAct (26) HPRD (3) BioGRID (40) DIP (1) BioPlex (1)]
P46527	CDN1B_HUMAN	9606	19	10	13	1	**25** [IntAct (13) HPRD (2) BioGRID (7) DIP (2) BioPlex (1)]
P06493	CDK1_HUMAN	9606	15	5	13	0	**18** [IntAct (4) HPRD (1) BioGRID (11) DIP (1) BioPlex (1)]
Q12834	CDC20_HUMAN	9606	14	6	9	0	**19** [IntAct (1) HPRD (1) BioGRID (16) BioPlex (1)]
P38936	CDN1A_HUMAN	9606	13	6	12	0	**15** [IntAct (3) HPRD (1) BioGRID (9) DIP (1) BioPlex (1)]
Q13309	SKP2_HUMAN	9606	9	5	9	0	**10** [IntAct (2) HPRD (1) BioGRID (6) BioPlex (1)]
P06400	RB_HUMAN	9606	8	2	8	0	**8** [HPRD (1) BioGRID (7)]
Q9H211	CDT1_HUMAN	9606	7	5	4	0	**9** [IntAct (3) HPRD (1) BioGRID (4) BioPlex (1)]
Q02539	H11_HUMAN	9606	7	2	7	0	**7** [BioGRID (7)]
P28749	RBL1_HUMAN	9606	7	3	7	0	**9** [HPRD (2) BioGRID (6) BioPlex (1)]
P61024	CKS1_HUMAN	9606	6	3	6	0	**7** [IntAct (2) BioGRID (4) BioPlex (1)]
Q08999	RBL2_HUMAN	9606	6	5	4	0	**8** [IntAct (3) HPRD (1) BioGRID (3) BioPlex (1)]
Q99741	CDC6_HUMAN	9606	5	4	3	2	**5** [HPRD (1) BioGRID (4)]
Q01094	E2F1_HUMAN	9606	5	4	4	0	**5** [HPRD (2) BioGRID (3)]

Show 25 entries — Previous 1 Next

Figura 43. *Interfaz principal para la exploración de interacciones únicas de la proteína CCNA2_HUMAN. Se muestran las métricas calculadas para cada interacción, así como el número de estructuras tridimensionales asignadas a estas y la procedencia de los curation events correspondientes. También puede observarse el sistema de filtrado en tiempo real, en este caso presentando las interacciones con 5 o más experimentos y un mínimo de dos publicaciones, que son 14 de un total de 134 interacciones registradas para la proteína CCNA2_HUMAN.*

En la última de las columnas se ofrece información resumida sobre los registros en bases de datos primarias para cada interacción única. Esta información puede consultarse en detalle pulsando sobre el icono con forma de lupa que aparece al final de la línea. En **Figura 44** puede verse la tabla que se obtiene al pulsar sobre dicho icono en el caso de la interacción entre las proteínas CCNA2_HUMAN y RBL2_HUMAN.

8 Curation Events for CCNA2_HUMAN - RBL2_HUMAN interaction

Interactions of CCNA2_HUMAN | Export raw curation events in MITAB format | Show 25 entries | Search:

ProteinA	ProteinB	Method	Publication	Source
RBL2_HUMAN	CCNA2_HUMAN	anti bait coimmunoprecipitation (MI:0006)	Litovchick, L. et al., 2007 (PMID:17531812)	IntAct (Acc: EBI-1389907)
RBL2_HUMAN	CCNA2_HUMAN	anti bait coimmunoprecipitation (MI:0006)	Litovchick, L. et al., 2007 (PMID:17531812)	IntAct (Acc: EBI-1390043)
RBL2_HUMAN	CCNA2_HUMAN	anti tag coimmunoprecipitation (MI:0007)	Litovchick, L. et al., 2007 (PMID:17531812)	IntAct (Acc: EBI-1389763)
RBL2_HUMAN	CCNA2_HUMAN	affinity chromatography technology (MI:0004)	Li, Y. et al., 1993 (PMID:8253383)	BioGRID (Acc: 287790)
RBL2_HUMAN	CCNA2_HUMAN	in vivo (MI:0493)	Li, Y. et al., 1993 (PMID:8253383)	HPRD (Acc: P20248)
RBL2_HUMAN	CCNA2_HUMAN	affinity chromatography technology (MI:0004)	Huttlin, EL. et al., 2015 (PMID:26186194)	BioGRID (Acc: 1198049)
RBL2_HUMAN	CCNA2_HUMAN	affinity chromatography technology (MI:0004)	Huttlin, EL. et al., 2015 (PMID:26186194)	BioPlex
RBL2_HUMAN	CCNA2_HUMAN	pull down (MI:0096)	Farkas, T. et al., 2002 (PMID:12006580)	BioGRID (Acc: 796995)
ProteinA	ProteinB	Method	Publication	Source

Showing 1 to 8 of 8 entries | Previous | 1 | Next

Figura 44. *Vista en detalle de los diferentes registros contenidos en las bases de datos primarias (curation events) para una interacción única específica. Cada elemento contiene un enlace a su registro original en la base de datos externa correspondiente. En el caso de las publicaciones se proporciona, además, la posibilidad de consultar el conjunto de interacciones descritas en cada artículo científico dentro de la propia plataforma de APID Interactomes.*

Esta misma información puede obtenerse, para el total de las interacciones únicas, con la segunda de las opciones disponibles en la pantalla inicial ( ). Así, en **Figura 45** pueden verse todos los *curation events* o registros originales contenidos en las bases de datos primarias en los que participa la proteína RASH_HUMAN.

Esta tabla contiene la información detallada sobre cada registro con el método de detección de la interacción (representado por el término correspondiente de la ontología PSI-MI), el artículo científico donde se describe y la base de datos primaria donde está almacenado dicho registro, así como su identificador único original.

Para todos los identificadores únicos de entidades pertenecientes a bases de datos externas se incluye un enlace al registro original ( ). En el caso de la columna donde se detalla el artículo científico, además del enlace correspondiente a la entrada en PUBMED, se proporciona la posibilidad de interrogar a APID Interactomes sobre todas las interacciones descritas en dicho artículo ( ).

Por último, la tercera y última opción que la aplicación ofrece para las proteínas encontradas es la de construir una red con las interacciones de dichas proteínas. De esta última parte se habla más adelante, en la sección **Herramienta *web* para generación, visualización y análisis de redes de interacción de proteínas**.

371 Curation Events for RASH_HUMAN (P01112, Homo sapiens)

GTPase HRas

Interactions of RASH_HUMAN | Export raw curation events in MITAB format | Show 25 entries | Search:

ProteinA	ProteinB	Method	Publication	Source
KPCZ_HUMAN	RASH_HUMAN	in vivo (MI:0493)	Zhao, Y. et al., 2005 (PMID:15721486)	HPRD (Acc: P01112)
RASH_HUMAN	RAF1_HUMAN	two hybrid (MI:0018)	Zhang, XF. et al., 1993 (PMID:8332187)	IntAct (Acc: EBI-8653209)
RASH_HUMAN	RAF1_HUMAN	pull down (MI:0096)	Zhang, XF. et al., 1993 (PMID:8332187)	IntAct (Acc: EBI-8653240)
RASH_HUMAN	RAP1B_HUMAN	two hybrid (MI:0018)	Zhang, XF. et al., 1993 (PMID:8332187)	HPRD (Acc: P01112)
RASH_HUMAN	ARAF_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	IntAct (Acc: EBI-537048)
RASH_HUMAN	ARAF_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	HPRD (Acc: P01112)
RASH_HUMAN	RAF1_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	IntAct (Acc: EBI-537265)
RASH_HUMAN	RAF1_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	IntAct (Acc: EBI-537342)
RASH_HUMAN	RAF1_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	IntAct (Acc: EBI-537202)
RASH_HUMAN	RAF1_HUMAN	two hybrid (MI:0018)	Yuryev, A. et al., 2003 (PMID:12620389)	BioGRID (Acc: 260018)
RASH_HUMAN	AGTR1_HUMAN	in vivo (MI:0493)	Yang, H. et al., 1996 (PMID:8753867)	HPRD (Acc: P01112)
RASH_HUMAN	AFAD_HUMAN	pull down (MI:0096)	Yamamoto, T. et al., 1999 (PMID:10334923)	BioGRID (Acc: 300899)
RASH_HUMAN	AFAD_HUMAN	affinity chromatography technology (MI:0004)	Yamamoto, T. et al., 1999 (PMID:10334923)	BioGRID (Acc: 300901)
RABX5_HUMAN	RASH_HUMAN	enzymatic study (MI:0415)	Xu, L. et al., 2010 (PMID:20655225)	BioGRID (Acc: 674560)
RASH_HUMAN	TLR9_HUMAN	in vivo (MI:0493)	Xu, H. et al., 2003 (PMID:12867418)	HPRD (Acc: P01112)
RASH_HUMAN	KPCI_HUMAN	in vivo (MI:0493)	Wooten, MW. et al., 2000 (PMID:10848576)	HPRD (Acc: P01112)
ICMT_HUMAN	RASH_HUMAN	in vivo (MI:0493)	Winter-Vann, AM. et al., 2003 (PMID:12750467)	HPRD (Acc: P01112)
RGS12_HUMAN	RASH_HUMAN	in vivo (MI:0493)	Willard, MD. et al., 2007 (PMID:17380122)	HPRD (Acc: P01112)
RASH_HUMAN	RAF1_HUMAN	pull down (MI:0096)	Warne, PH. et al., 1993 (PMID:8332195)	IntAct (Acc: EBI-8653777)
RASH_HUMAN	ZHX2_HUMAN	affinity chromatography technology (MI:0004)	Wang, JB. et al., 2005 (PMID:15824104)	BioGRID (Acc: 727509)
RASH_HUMAN	K1C18_HUMAN	two hybrid (MI:0018)	Wang, J. et al., 2011 (PMID:21988832)	BioGRID (Acc: 893135)
RASF2_HUMAN	RASH_HUMAN	in vivo (MI:0493)	Vos, MD. et al., 2003 (PMID:12732644)	HPRD (Acc: P01112)
RASH_HUMAN	RASF1_HUMAN	in vivo (MI:0493)	Vos, MD. et al., 2000 (PMID:10998413)	HPRD (Acc: P01112)
RASH_HUMAN	RASF1_HUMAN	pull down (MI:0096)	Vos, MD. et al., 2000 (PMID:10998413)	BioGRID (Acc: 275128)
RASH_HUMAN	RASF1_HUMAN	affinity chromatography technology (MI:0004)	Vos, MD. et al., 2000 (PMID:10998413)	BioGRID (Acc: 275129)
ProteinA	ProteinB	Method	Publication	Source

Showing 1 to 25 of 371 entries | Previous 1 2 3 4 5 … 15 Next

Figura 45. *Vista de todos los registros disponibles en las bases de datos primarias que describen interacciones para la proteína RASH_HUMAN.*

4.2.4 Interacciones entre las proteínas de una lista de interés

Otro contexto experimental muy frecuente es aquel en el que el investigador obtiene como resultado de sus experimentos una lista concreta de genes o proteínas. Para este caso, la aplicación web de APID Interactomes proporciona la posibilidad de especificar listas proteínas (o sus genes codificantes) y así obtener todas las posibles interacciones entre estas.

En **Figura 46** puede verse la interfaz diseñada para enviar listas de proteínas a la aplicación. De nuevo, estas proteínas pueden especificarse con su nombre UniProt, su identificador UniProt AC o el nombre del gen correspondiente. También puede determinarse el organismo con el que se está trabajando para acotar la búsqueda.

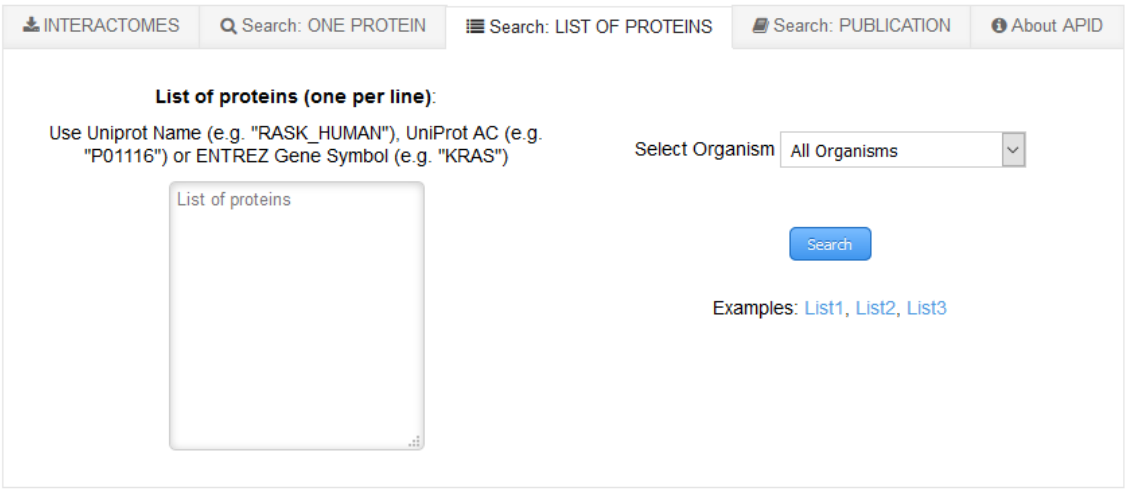


Figura 46. *Cuadro de búsqueda para una lista de proteínas de interés.*

Partiendo de esta lista de proteínas, APID Interactomes hace una búsqueda para cada una de ellas y ofrece los resultados en una nueva pantalla, como se ve en **Figura 47**, de la misma manera que lo hacía en la búsqueda de una sola proteína. En este caso, la diferencia reside en que APID Interactomes permite seleccionar entre todos los resultados aquellas proteínas que son de interés y crear una red personalizada con las interacciones que existan entre ellas. Al crear esta red, APID Interactomes enviará todos los datos disponibles sobre las interacciones de esta lista de proteínas al visualizador de redes del que se habla en la sección **Herramienta *web* para generación, visualización y análisis de redes de interacción de proteínas**.

List of proteins

Select: All, none | Build NETWORK with selected proteins

Showing 1 to 33 of 33 entries — Search:

Uniprot ID	Uniprot Name	Gene Name	Organism	Description	Interactions	[Display RESULTS]
P78536	ADA17_HUMAN	ADAM17	Homo sapiens	Disintegrin and metalloproteinase domain-containing protein 17	32	
Q02410	APBA1_HUMAN	APBA1	Homo sapiens	Amyloid beta A4 precursor protein-binding family A member 1	31	
Q12860	CNTN1_HUMAN	CNTN1	Homo sapiens	Contactin-1	40	
O00548	DLL1_HUMAN	DLL1	Homo sapiens	Delta-like protein 1	14	
Q9NR61	DLL4_HUMAN	DLL4	Homo sapiens	Delta-like protein 4	2	

Figura 47. *Extracto de los resultados que la aplicación muestra cuando el usuario efectúa una búsqueda de proteínas a partir de un listado concreto, en este caso uno de los disponibles como ejemplo en la plataforma web (denominado List1).*

Con esta red, el investigador puede estudiar un subconjunto del interactoma de un organismo a partir de un conjunto de proteínas que ha sido seleccionado experimentalmente, lo cual resulta extremadamente útil en muchos contextos. Valga como ejemplo una lista de genes que se expresan diferencialmente entre dos grupos de muestras: estudiar las relaciones entre sus proteínas añade una nueva capa de información funcional muy valiosa a la hora de establecer conclusiones a nivel biológico.

Además, para cada una de las proteínas encontradas, APID Interactomes proporciona las mismas opciones que cuando se hace una búsqueda para una única proteína: consultar las interacciones únicas ( ), explorar todas los *curation events* registrados en las bases de datos primarias ( ) y generar una red con las interacciones disponibles ( ).

4.2.5 Interacciones descritas en un determinado artículo científico

Por último, APID Interactomes ofrece la posibilidad de consultar todas las interacciones descritas en una publicación científica. Para ello, el usuario debe especificar el identificador PUBMED de la publicación.

Figura 48. *Cuadro de búsqueda para las interacciones descritas en un artículo científico determinado.*

En **Figura 48** puede verse la interfaz disponible para efectuar dicha operación. A partir del identificador de la publicación, APID Interactomes agrupa todas las interacciones descritas en dicho documento y añade además toda la información disponible, en su propia base de datos, sobre dichas interacciones.

Como puede verse en **Figura 49**, APID Interactomes añade a las interacciones descritas en la publicación especificada por el investigador todas las métricas que han sido calculadas por el algoritmo de integración de interacciones. Esto es importante porque el investigador podrá comprobar si una o varias de esas interacciones aparecen descritas también en otras publicaciones, resultando de especial interés aquellas que hayan sido detectadas mediante el uso de diferentes métodos experimentales.

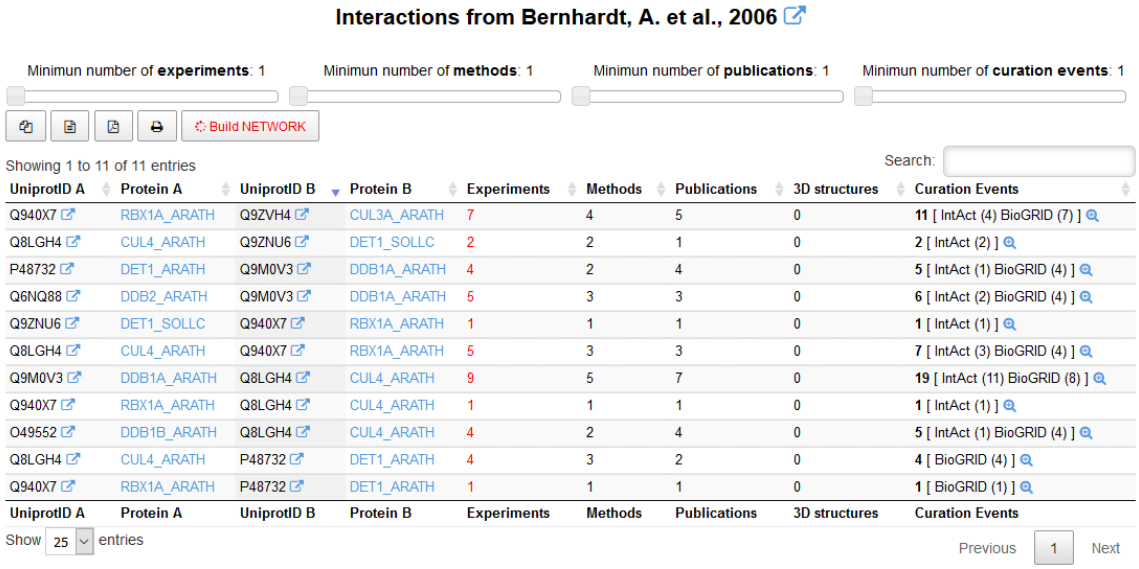


Figura 49. *Pantalla que muestras las interacciones únicas descritas en* **(Bernhardt et al. 2006)**.

También se ofrece la posibilidad de construir una red con las interacciones descritas en la publicación añadiendo también aquellas que APID Interactomes tiene registradas en su base de datos de tal manera que la red que inicialmente se podría construir con los datos descritos en la publicación se ve enriquecida con los datos aportados por otras publicaciones.

La búsqueda de información en APID Interactomes sobre las interacciones descritas en una publicación científica determinada es una operación computacionalmente costosa. Para garantizar una buena experiencia de uso se decidió establecer un control para aquellos artículos que describieran más de 200 interacciones. Dicho control consiste en una advertencia al usuario (**Figura 50**) de que el cálculo puede tardar en completarse ofreciéndole la posibilidad de descargar un fichero con el resultado pre-calculado y almacenado en el mismo formato en el que se ofrecen los interactomas. En cualquier caso, se ofrece también la posibilidad de llevar a cabo dicho cálculo si el usuario lo cree conveniente.

Esta decisión se tomó debido a que existen artículos científicos cuyo propósito es describir el interactoma completo de un organismo y, por tanto, incluyen decenas de miles de interacciones, como, por ejemplo, **(Rolland et al. 2014)** o **(Huttlin et al. 2015)**. En estos casos, aunque el servidor es capaz de procesar dicha cantidad de información, lo más probable es que el navegador del cliente no pueda manejar tablas de ese tamaño. Incluso en el caso de que fuera capaz de construir la interfaz para mostrar dicho contenido, su usabilidad sería escasa. Sin embargo, la descarga de un fichero de texto con toda la información disponible puede resultar de gran utilidad si se utiliza en una aplicación de escritorio para el análisis de redes como por ejemplo Cytoscape **(Shannon et al. 2003)** (36).

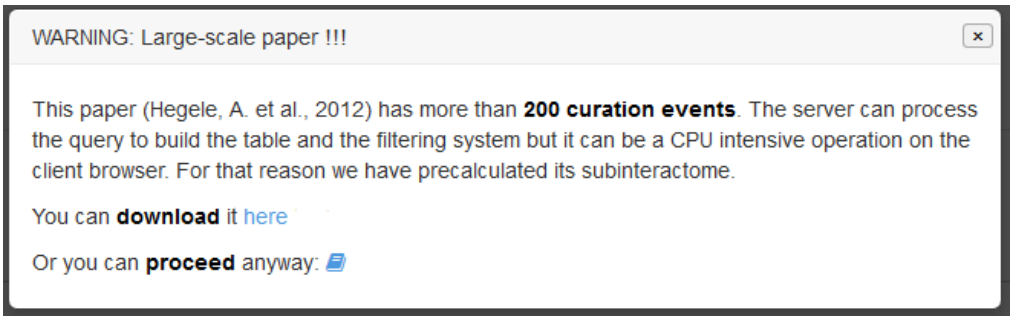


Figura 50. *Advertencia al usuario sobre el coste computacional de la operación que pretende realizar.*

4.2.6 Optimización de la base de datos

Algunas de las operaciones que lleva a cabo la aplicación de APID Interactomes requieren el uso de grandes cantidades de información. La base de datos de APID Interactomes contiene más de 1.200.000 registros de interacciones, así como más de 90.000 proteínas con cientos de miles de anotaciones funcionales y estructurales.

El uso de servidores de alto rendimiento permite la ejecución de consultas sencillas sobre una única tabla en unos segundos, pero cuando se trata de consultas complejas que involucran varias tablas o que deben descomponerse en miles de consultas sencillas, el aumento del tiempo necesario para ejecutarse es exponencial. Esto hace que la comunicación cliente-servidor se vea afectada y puede llegar a provocar errores si se sobrepasan ciertos tiempos de espera máximos o *timeouts.* Por supuesto, también afecta negativamente a la experiencia de uso de la aplicación.

Para conseguir mejorar el rendimiento de este tipo de consultas se trabajó en tres niveles diferentes: pre-calcular determinadas métricas, optimizar el SQL de las consultas textuales y desnormalizar la base de datos.

Como resultado de estas optimizaciones, los tiempos de acceso de la aplicación se redujeron de forma significativa, mejorando la experiencia del usuario e incluso haciendo posible la implementación de determinadas operaciones complejas como la búsqueda de interacciones por publicación científica o la construcción de redes a partir de una lista de proteínas.

A continuación, se describen con más detalle cada una de las optimizaciones que se llevaron a cabo.

4.2.6.1 Pre-cálculo de métricas de interacción

La primera optimización consistió en almacenar los resultados de los cálculos de métricas para todas las interacciones. Partiendo de la base de datos generada por el algoritmo de unificación de interacciones, se calcularon para cada interacción sus métricas correspondientes y se almacenaron en la tabla de interacciones únicas. Además, se añadieron campos para almacenar el número de registros contenidos en cada base de datos primaria para cada interacción única.

Así, la tabla original de interacciones únicas paso a contener un total de diez nuevos campos numéricos asociados a los siguientes conceptos:

1. Número de métodos de detección de interacción diferentes.
2. Número de publicaciones científicas que describen la interacción.
3. Número de experimentos.
4. Número total de registros en bases de datos primarias (*curation events)*.
5. Número de registros en la base de datos IntAct.
6. Número de registros en la base de datos HPRD.
7. Número de registros en la base de datos DIP.
8. Número de registros en la base de datos BioGrid.
9. Número de registros en la base de datos BioPlex.
10. Número de estructuras tridimensionales en la base de datos PDB.

De esta manera, con una consulta sencilla a una única tabla, la aplicación web puede obtener toda la información esencial de una interacción específica. Además, esta organización de la información está especialmente diseñada para el tipo de tablas que se muestran en la interfaz web de APID Interactomes.

4.2.6.2 Optimización de las consultas textuales

Otra de las optimizaciones que se llevaron a cabo fue la de optimizar las consultas cuyo criterio de búsqueda era una cadena textual. Esto estaba dirigido fundamentalmente a mejorar los tiempos en las búsquedas iniciales de proteínas ya que en dichas operaciones el usuario de la aplicación puede enviar al servidor un texto libre con el nombre total o parcial de una proteína o de un gen.

Para este tipo de búsquedas era necesario hacer uso del operador LIKE en las sentencias SQL con la penalización en tiempo que esto conlleva. Además, en la búsqueda de listados de proteínas, dicha penalización se multiplicaba en función del número de proteínas especificadas por el usuario.

Lo que se hizo para optimizar este tipo de consultas fue implementar búsquedas de tipo FullText **(Schwartz et al. 2012)**. Para ello, se crearon índices FullText y se modificaron las sentencias SQL de tal manera que pasaron a utilizar expresiones del tipo MATCH…AGAINST para sustituir a los operadores LIKE.

4.2.6.3 Desnormalización de la base de datos

La desnormalización de bases de datos es una técnica que trata de optimizar el rendimiento de una base de datos a través de la agregación de datos redundantes. Aunque un diseño lógico normalizado es lo más deseable para evitar inconsistencias en las bases de datos relacionales, es cierto que dicha estrategia formal no tiene en cuenta el rendimiento de estas. Así, el almacenamiento físico en ficheros de la base de datos se ve normalmente penalizado por su diseño lógico.

Aunque este problema se ha paliado tradicionalmente con el uso de vistas materializadas en ficheros físicos en disco, lo más eficaz es efectuar una desnormalización de la base de datos guiada, si es posible, por el uso que una determinada aplicación hará de ella. El inconveniente de esto es que al introducir datos redundantes se corre el riesgo de que aparezcan inconsistencias.

En el caso de la base de datos de APID Interactomes, las ventajas de una desnormalización resultan evidentes ya que está diseñada específicamente para un tipo de aplicación concreto. Además, el riesgo de que aparezcan inconsistencias es nulo puesto que se trata de una base de datos de solo lectura.

Lo que se hizo, por tanto, fue efectuar una desnormalización orientada a mejorar el acceso a la información relacionada con cada interacción sobre la base de datos generada por el algoritmo de unificación de interacciones. Para ello se añadieron datos redundantes sobre las proteínas que participaban en cada interacción como la especie, el nombre del gen asociado, etc.

4.3 Herramienta *web* para generación, visualización y análisis de redes de interacción de proteínas

Cualquier conjunto de datos que represente las relaciones entre diferentes entidades puede expresarse como una red de elementos interconectados. En el caso de los datos de interacción de proteínas, su representación en forma de red aporta una capa de información adicional a través del análisis de la topología resultante. Dicho análisis permite contextualizar las interacciones individuales y caracterizar cada proteína en base a su peso en la estructura global permitiendo de esta manera extraer conclusiones en el ámbito de la biología molecular **(Vidal et al. 2011)**.

Además, la representación de una red permite visualizar en un único gráfico todas las interacciones que tienen lugar en un conjunto concreto de proteínas, o en todo el proteoma de un organismo, tratándose entonces de una visualización del interactoma de dicho organismo.

Es por ello que una parte fundamental del trabajo llevado a cabo en esta Tesis Doctoral estuvo enfocado a la creación de un visualizador de redes de interacción. Este se planteó como destino para todos los conjuntos de interacciones que un investigador pudiese generar con la plataforma bioinformática descrita en el capítulo anterior.

Así, se desarrolló un visualizador web que recibe como entrada un conjunto de interacciones y ofrece al investigador un entorno interactivo con el que poder explorar de forma gráfica la información disponible en APID Interactomes sobre dicho conjunto de datos.

4.3.1 Arquitectura del visualizador

El mayor reto de esta parte del trabajo residió en construir una interfaz compleja, pero compatible con las capacidades de los navegadores web habituales. Algunos de los objetivos que se plantearon al inicio del desarrollo fueron los siguientes:

1. Proporcionar al investigador una visualización de la red desde su navegador sin que tenga la necesidad de recurrir a ningún software adicional.
2. Tratar de expresar gráficamente la información sobre las interacciones generada por el algoritmo de unificación de APID Interactomes.
3. Utilizar dicha información como criterio para filtrar interacciones de la red inicial.

4. Ofrecer la posibilidad de aplicar diferentes *layouts* sobre la red (es decir, diferentes formas de disponer y representar en el plano visual los nodos y las relaciones).
5. Diseñar un sistema que permita al investigador conocer el *ambiente* funcional y estructural de la red visualizada mediante el uso de las anotaciones biológicas funcionales y estructurales que se conocen asociadas a dos o más de las proteínas representadas.
6. Conectar esta visualización con el resto de las funcionalidades de APID Interactomes.

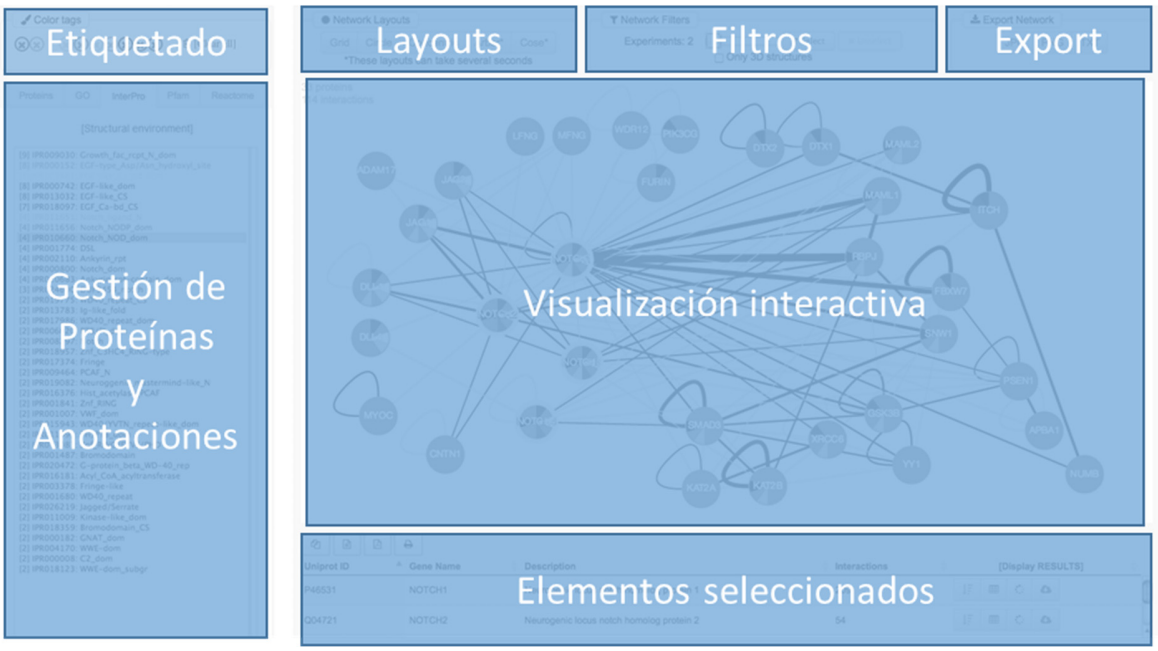


Figura 51. *Estructura de la interfaz diseñada para el visualizador de redes de APID Interactomes.*

En **Figura 51** puede verse un esquema con las diferentes áreas de la interfaz del visualizador. A continuación, se describirá cada una de estas partes y sus funcionalidades.

4.3.2 Representación interactiva de la red

Como se ha dicho anteriormente, el objetivo principal de esta parte de la aplicación web es poder mostrar la representación gráfica de la red sin que el usuario tenga que recurrir a la instalación de ningún software adicional.

Lo que se hizo fue construir una estructura de datos que representara las redes y métricas disponibles en APID Interactomes y que a su vez fuera compatible con el sistema de visualización de cytoscape.js. Para ello, se definieron conjuntos de atributos tanto para los nodos (proteínas) como para los arcos (interacciones) y se utilizaron para almacenar dicha información.

De esta manera, la red que se estaba representando contenía, no solo las relaciones entre nodos necesarias para su propia visualización, sino también la información disponible en APID Interactomes sobre ambas entidades. Gracias a ello, la aplicación maneja una capa adicional de información que posibilita, por ejemplo, utilizar el nombre del gen en la visualización para hacerla

más compatible con las aplicaciones bioinformáticas habituales o asociar a los arcos las métricas de la interacción que representan.

Esto permitió también implementar nuevas funcionalidades interactivas adaptadas a las necesidades de la aplicación, tales como generar una tabla con la información sobre las proteínas seleccionadas y ofrecer sobre estas las mismas operaciones que APID Interactomes proporciona si se buscan a través de la interfaz principal de la aplicación, o mostrar al usuario información sobre las métricas de la interacción entre dos proteínas al hacer doble *click* sobre el arco correspondiente. Ambos ejemplos pueden verse en **Figura 52**.

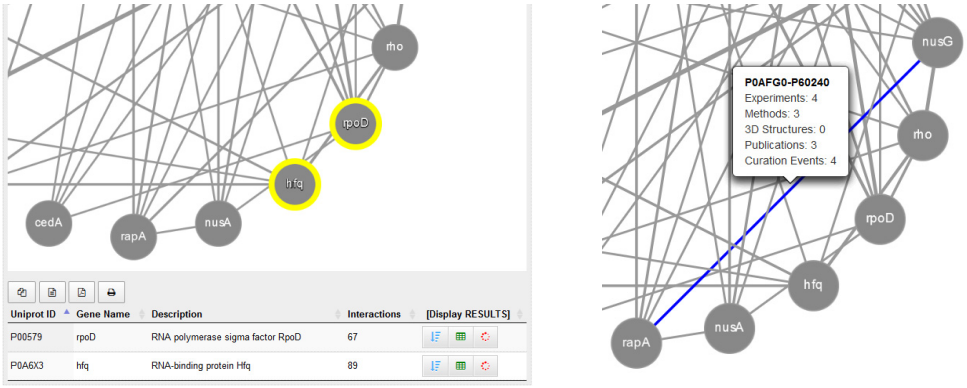


Figura 52. *Ejemplos de funcionalidades implementadas sobre el visualizador de redes. Las proteínas seleccionadas en la red aparecen debajo de esta ofreciendo la misma información y operaciones disponibles que si se hubieran buscado en el menú principal de APID Interactomes. Se muestra también la información que aparece al hacer doble click sobre cualquiera de los arcos que representan a cada interacción.*

Además, la métrica principal que APID Interactomes calcula para cada interacción, el número de experimentos, se utilizó como parámetro para asignar diferente grosor a la representación de cada arco. De esta manera, el investigador puede identificar a simple vista qué interacciones son más fiables.

Por último, se construyó un listado textual con las proteínas que componen la red ordenadas alfabéticamente y se enlazó con su visualización para que el usuario pueda seleccionarlas desde el propio listado.

4.3.3 Gestión de las anotaciones

Más allá de la información que se utiliza para construir la visualización de la red, la aplicación maneja otras estructuras de datos que contienen información asociada a cada proteína. Para ello utiliza todas las anotaciones funcionales de Gene Ontology **(Huntley et al. 2015)** (30), estructurales de InterPro **(Mitchell et al. 2015)** (35) y Pfam **(Finn et al. 2015)** (34) y sobre rutas de señalización de Reactome **(Croft et al. 2014)** (33) que estén asignadas, al menos, a dos proteínas de la red.

APID Interactomes construye, con toda esta información, cuatro espacios de anotación que se adaptan al sub-interactoma que cada red visualizada representa. Dichos espacios muestran el número de ocurrencias en la red para cada término de las diferentes ontologías permitiendo de esta

manera la asignación global de características funcionales o estructurales a dicha red con mayor o menos significatividad.

Para conseguir esto, APID Interactomes reorganiza de forma dinámica todas las anotaciones disponibles para cada red que el investigador construye. En **Figura 53** puede verse un esquema que representa dicho procesamiento, donde partiendo de un listado de proteínas con conjuntos de anotaciones asociadas se construye un nuevo listado que, en este caso, está compuesto por anotaciones con conjuntos de proteínas asociados, permitiendo así una mejor caracterización de la red.

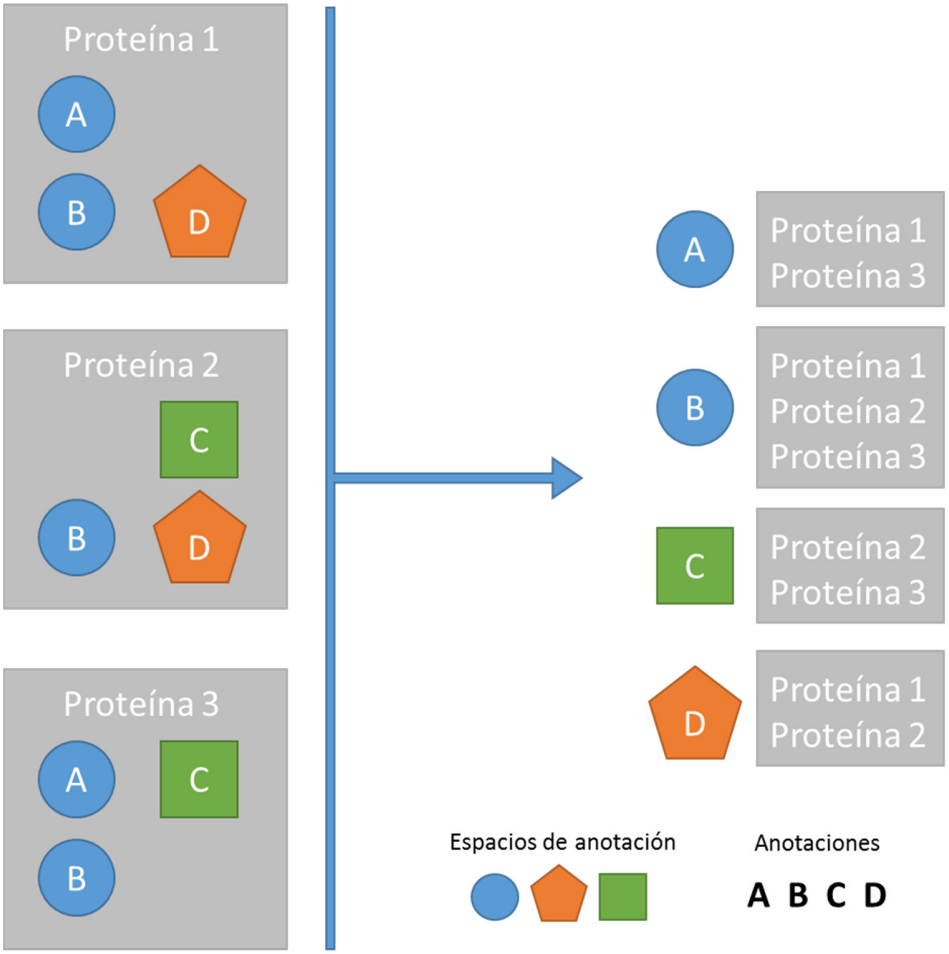


Figura 53. *Procesamiento de las anotaciones para el conjunto de proteínas que forman una red específica. Las figuras geométricas representan los diferentes espacios de anotación y las letras los términos dentro de cada espacio.*

También, gracias al uso de identificadores internos, cada término queda enlazado con los nodos que representan a las proteínas que lo tienen asignado. Esto permite, al igual que con el listado textual de proteínas mencionado anteriormente, seleccionar un término en el listado de anotaciones provocando que se seleccionen automáticamente las proteínas correspondientes en la red interactiva (**Figura 54**). Si bien esta selección automática resulta útil para visualizar las proteínas de

la red con una determinada anotación, APID Interactomes implementa un sistema de representación de anotaciones más completo tal y como se describe en el siguiente apartado.

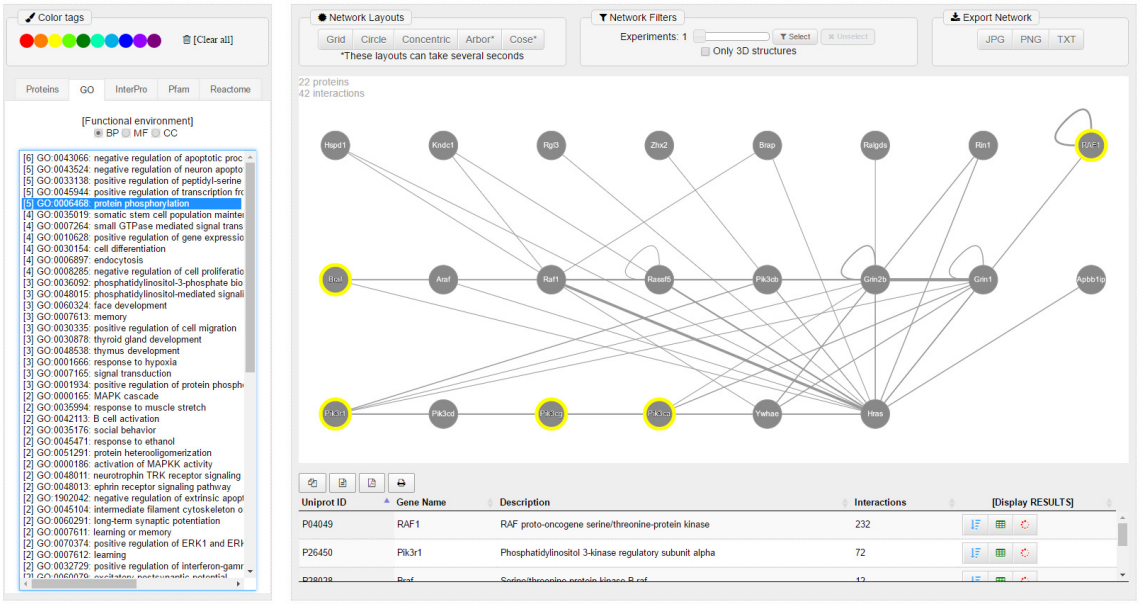


Figura 54. *Selección de un término en el listado de anotaciones funcionales que provoca la selección automática de las proteínas con dicha anotación en la red.*

4.3.4 Sistema de etiquetado por colores

Con el objetivo de aprovechar mejor la gran cantidad de anotaciones disponibles en la base de datos de APID Interactomes y obtener una visualización de la red más completa, se planteó la posibilidad de implementar un mecanismo de marcaje visual de nodos para más de una anotación.

El sistema se diseñó teniendo en cuenta dos aspectos fundamentales: **(i)** el número de marcas visuales que un nodo puede tener es limitado y **(ii)** cada investigador estará interesado en diferentes anotaciones independientemente del número de ocurrencias de estas en la red.

Así, se implementó un sistema de etiquetado de nodos interactivo y dinámico basado en el uso de colores para un máximo de 10 anotaciones (**Figura 55**). Interactivo, porque el usuario puede asignar colores a cualquiera de las anotaciones simplemente arrastrando el círculo del color correspondiente sobre el texto de la anotación. Y dinámico, porque el algoritmo se adapta al número de etiquetas en cada momento de tal manera que utiliza toda el área del nodo para distribuir los colores. Además, independientemente del número de etiquetas de color utilizadas, cada una aparece en la misma posición en todos los nodos para que estos sean comparables visualmente.

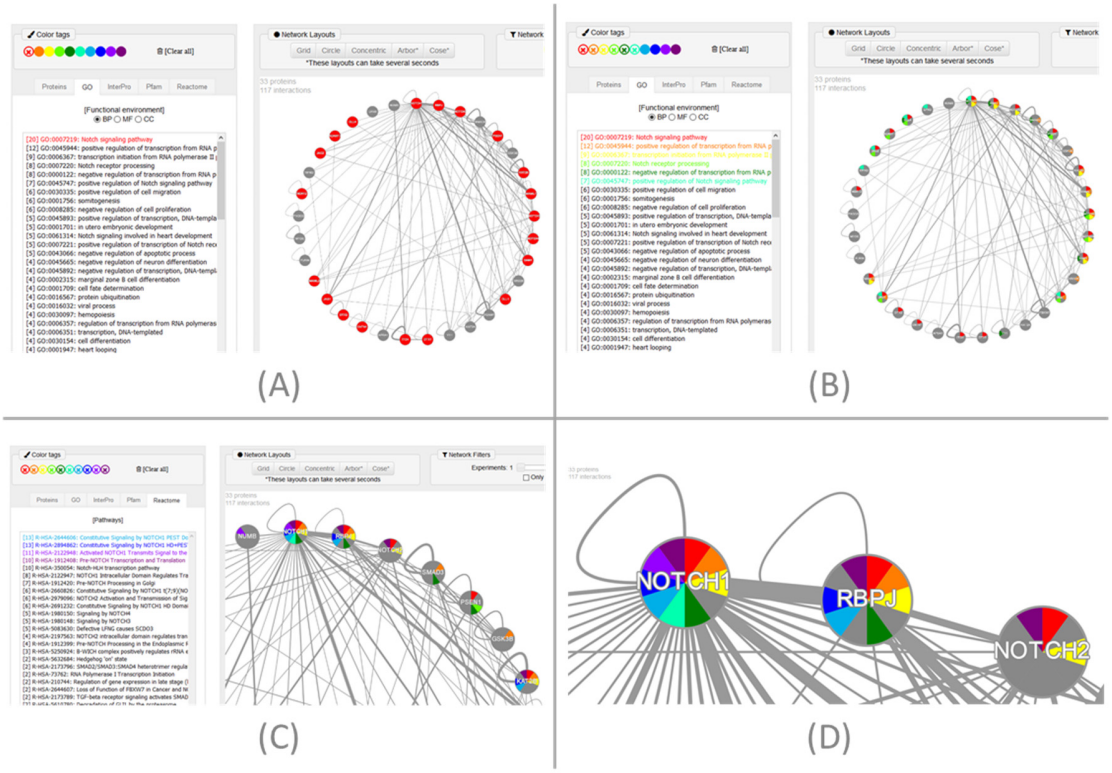


Figura 55. *Sistema de etiquetado por colores. Uso progresivo del área del nodo (A y B), combinación de diferentes espacios de anotación (C, Gene Ontology y Reactome) y detalle del* etiquetado *simultáneo de diez anotaciones con la misma posición para cada etiqueta en todos los nodos (D).*

La asignación de colores puede hacerse sobre etiquetas de cualquiera de los espacios de anotación disponibles. También pueden eliminarse de forma individual o retirando todos los colores asignados de forma simultánea.

4.3.5 Layouts

Una misma red de interacción puede tener numerosas visualizaciones dado que se trata de un grafo cuya representación gráfica en un espacio bidimensional puede estar sujeta a diferentes criterios. Si bien pueden encontrarse en la literatura recomendaciones generales sobre la estrategia a seguir a la hora de representar un grafo **(Sugiyama 2002)**, existen diversas teorías y aproximaciones a dicho problema. La representación de un grafo, dependiendo del tamaño de este y la estrategia elegida, puede ser una tarea con un alto coste en tiempo de computación.

En APID Interactomes se han implementado cinco estrategias de representación, o *layouts*, diferentes:

1. **Grid**: representa la red con los nodos en forma de malla de puntos, ocupando todo el espacio disponible y ordenados alfabéticamente. Se trata de una representación sencilla que no requiere mucho tiempo de CPU, motivo por el cual se estableció como la representación por defecto para todas las redes.

2. **Circle**: representa los nodos en una circunferencia ordenándolos por grado. Este *layout* ya utiliza una métrica de la red como criterio para ordenar los nodos, pero sigue siendo poco exigente en cuanto a recursos computacionales.
3. **Concentric**: utiliza de nuevo el grado de cada nodo, pero esta vez para establecer un patrón de circunferencias concéntricas situando en el centro los nodos con mayor grado.
4. **Arbor**: Adaptación para cytoscape.js del algoritmo de simulación física Arbor (64). Requiere efectuar un cálculo complejo que para grafos grandes puede resultar inviable.
5. **Cose**: Es un algoritmo del tipo *force-directed* **(Fruchterman and Reingold 1991)** que utiliza simulaciones físicas tal y como se describe en **(Dogrusoz et al. 2009)**. Al igual que *Arbor*, aunque puede resultar muy útil para descubrir patrones, tiene un coste computacional elevado.

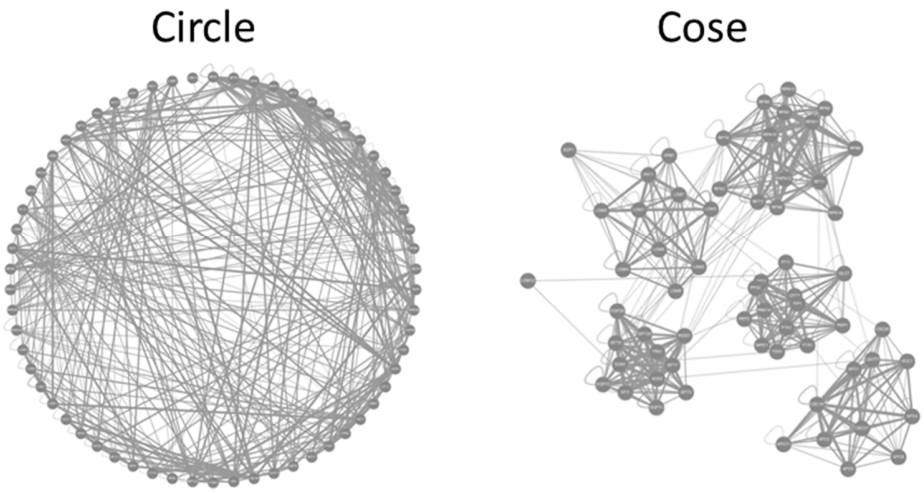


Figura 56. *Red de interacción con dos layouts diferentes, denominados Circle y Cose. Puede observarse como este último permite al investigador detectar visualmente la existencia de agrupamientos en la red.*

El uso de los diferentes *layouts* puede ayudar al investigador a extraer información significativa de la topología, como por ejemplo la existencia de agrupamientos o *clusters* tal y cómo se muestra en **Figura 56**, en este caso gracias al uso del algoritmo de representación Cose.

4.3.6 Filtros interactivos

El visualizador de redes implementado en APID Interactomes ofrece también la posibilidad de filtrar las interacciones, en tiempo real, usando como criterio el número de experimentos que las describen. Así, el investigador puede utilizar un control interactivo para establecer el número mínimo de experimentos y automáticamente las interacciones que cumplen dicho requisito aparecerán seleccionadas en la red. Si se aplica el filtro en ese momento, las interacciones restantes se eliminarán dando lugar a una nueva red con una topología distinta. Este proceso es reversible, bastará con pulsar el botón *Unselect* para que las interacciones eliminadas vuelvan a aparecer.

También se ha implementado un filtro basado en la existencia de estructuras tridimensionales de tal manera que, si se selecciona, la visualización mostrará una nueva red solo con las interacciones

que estén refrendadas por una estructura tridimensional en la base de datos PDB **(De Beer et al. 2014; Rose et al. 2015)**.

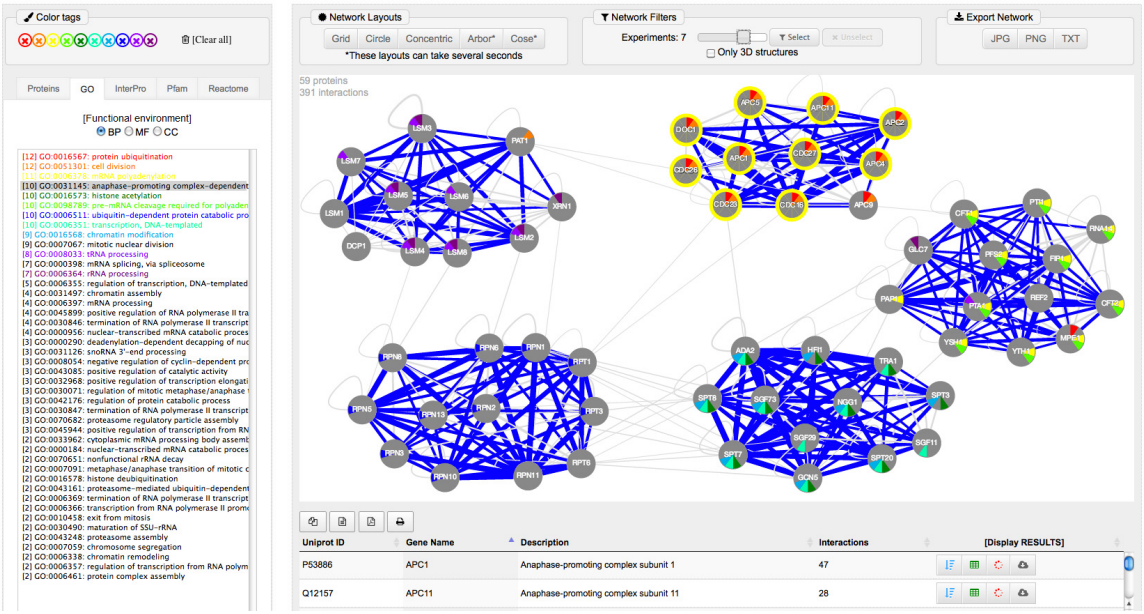


Figura 57. *Uso combinado del sistema de filtrado, el layout arbor y el sistema de etiquetado por colores para descubrir complejos proteicos.*

El uso de este sistema de filtrado, junto a los diferentes *layouts* disponibles, permite al investigador extraer conclusiones sobre la topología de la red y descubrir agrupamientos que de otra manera permanecerían ocultos. En **Figura 57** puede verse una red que contiene 59 proteínas de *Saccharomyces cerevisiae* implicadas en procesos relacionados con el DNA/RNA. Al filtrar las interacciones por número de experimentos y aplicar el *layout* arbor, la topología de la red muestra la presencia de cinco complejos de proteínas diferentes. Utilizando el sistema de etiquetado por colores se observa que las anotaciones también confirman la presencia de cinco grupos funcionales distintos.

4.4 Análisis comparativo, topológico y funcional, de algunos de los interactomas generados

Los interactomas generados por el algoritmo de APID Interactomes permiten el estudio global de todas las interacciones descritas en la literatura científica para un organismo dado. En este último capítulo de resultados se presentan, de modo breve y esquemático, un conjunto de análisis y estudios comparativos de varios interactomas procedentes de humano y de organismos modelo, para los cuales se tiene ya actualmente una gran cantidad de datos. Los interactomas estudiados en este apartado son los correspondientes a los siguientes organismos: *Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae* y *Escherichia coli*.

Tanto este análisis, como el que se presenta en el apartado **Análisis comparativo de bases de datos primarias: caso interactoma humano**, se llevaron a cabo a partir de los datos disponibles en la

versión de junio 2016 de la base de datos de APID Interactomes. En **Tabla 5** se puede observar que, como cabría esperar, dicha versión incluye más interacciones que la versión inicial y mejora ligeramente la cobertura de estas sobre el proteoma correspondiente para la mayoría de los organismos. La base de datos de APID Interactomes se actualiza aproximadamente cada 6 meses y el resultado de estas actualizaciones puede consultarse en la sección de estadísticas de la página web.

Tabla 5. *Cobertura de los interactomas analizados en la versión junio 2016 de la base de datos de APID Interactomes.*

Organismo	Proteoma	Proteínas en el interactoma	Interacciones en el interactoma	Cobertura (%)
Homo sapiens	70,615	29,701	349,144	42.06
Mus musculus	51,414	18,951	59,329	36.86
Caenorhabditis elegans	26,672	5,519	14,516	20.69
Saccharomyces cerevisiae (strain ATCC 204508 / S288c)	*6,721*	*7,276*	*131,895*	*108.26*
Escherichia coli (strain K12)	*4,254*	*3,546*	*25,332*	*83.36*

4.4.1 Niveles de calidad basados en la métrica de experimentos

Una de las ventajas de los interactomas generados durante este trabajo de investigación es que, como ya se ha explicado anteriormente, cada interacción contenida en estos lleva asociadas una serie de métricas que caracterizan su fiabilidad. Esto, unido al hecho de que los interactomas completos pueden descargarse de la plataforma web en un formato estándar, hace posible la construcción de redes con los datos calculados por el algoritmo de APID Interactomes en programas especializados en análisis de este tipo de redes como Cytoscape.

De esta manera, asociando cada métrica a su arco correspondiente, pueden establecerse filtros numéricos personalizados para generar diferentes redes en base a ellos. Esto es, el uso de las métricas de APID Interactomes en forma de parámetros de los arcos de la red proporciona un sistema de filtrado de interacciones que permite generar interactomas específicos para cualquier combinación de valores de dichas métricas.

En **Figura 58** se muestra un ejemplo de esto. Partiendo del interactoma completo para *Homo sapiens*, se genera una red en Cytoscape que contiene la métrica de experimentos calculada por el algoritmo de APID Interactomes en forma de parámetro para cada uno de los arcos de la red. A continuación, se establece un filtro de arcos basado en dicho parámetro y se utiliza para generar subredes con diferentes valores numéricos de este. Mediante este procedimiento se pueden generar diferentes versiones de un interactoma asociadas a cada umbral mínimo de experimentos o a cualquier otra de las métricas calculadas en APID Interactomes.

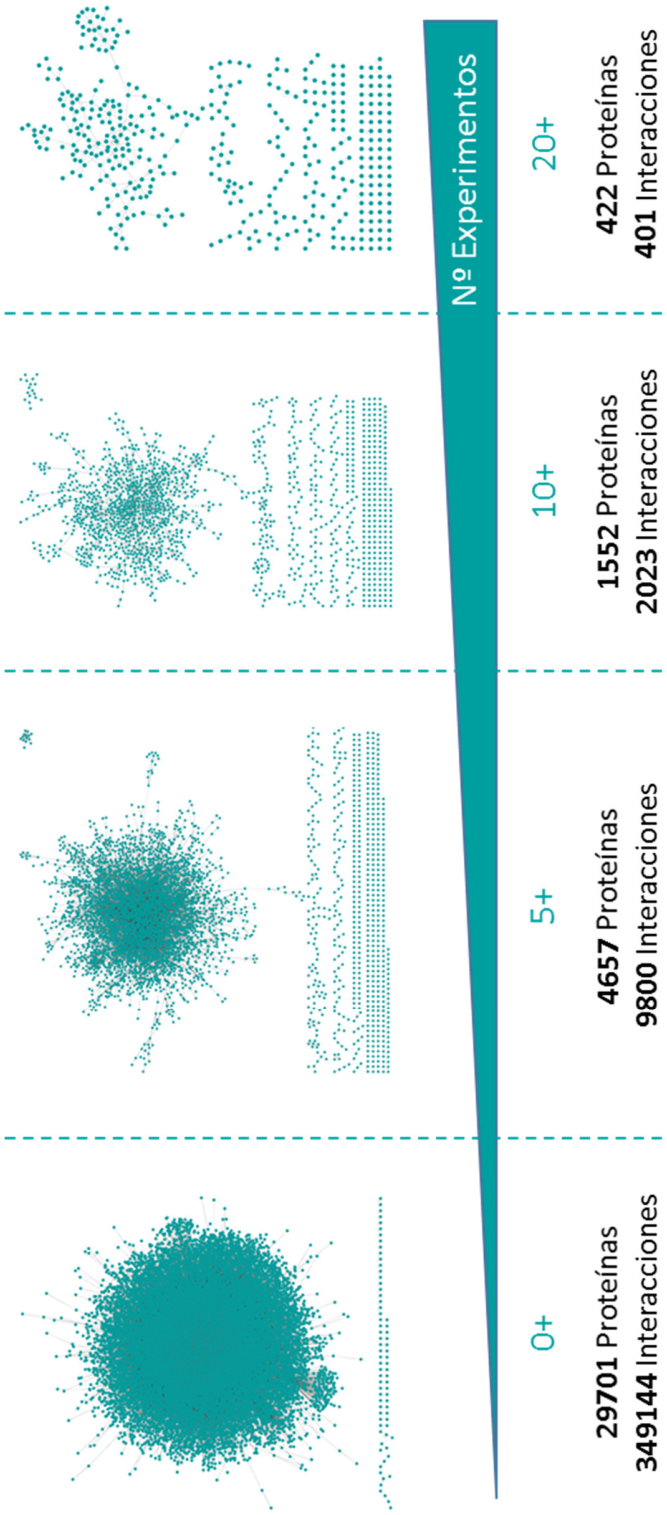


Figura 58. *Interactomas de Homo Sapiens para diferentes conjuntos de interacciones seleccionadas en base al número de experimentos que el algoritmo de APID Interactomes les ha asignado. Para cada grupo se exigen un mínimo de 0, 5, 10 o 20 experimentos por interacción respectivamente.*

La **Figura 58** representa cuatro redes obtenidas de esta manera para el interactoma humano que contienen respectivamente: 29701 proteínas (sin filtro), 4657 proteínas (con un filtro de al menos 5 experimentos que validen las interacciones), 1552 proteínas (con un filtro de al menos 10 experimentos) y 422 proteínas (con un filtro de al menos 20 experimentos). Está claro que los dos últimos filtros son muy astringentes y no dejan casi datos, pero el primero produce un interactoma humano fiable que puede resultar útil en diversos escenarios.

4.4.2 Estudio comparativo de los interactomas de cinco organismos modelo

A continuación, se presenta una caracterización topológica de las redes generadas con los datos de los interactomas de *Homo sapiens* y otros cuatro organismos modelo: *Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae* y *Escherichia coli.*

Para poder analizar dichas redes, estas se generaron en Cytoscape a partir de los datos disponibles en APID Interactomes y se analizaron con la herramienta NetworkAnalizer **(Doncheva et al. 2012)**. Esta aplicación permite calcular diversos parámetros que caracterizan la topología de la red. En **Figura 59** se muestran los siguientes:

1. **Número de nodos**: Representa el número total de proteínas contenidas en el interactoma.
2. **Número de arcos**: Representa el número de interacciones moleculares conocidas entre dichas proteínas.
3. **Diámetro de la red**: Distancia máxima entre dos pares de nodos. Caracteriza la dimensión y cohesión global de la red.
4. **Coeficiente de agrupamiento medio (*clustering coefficient*)**: Valor medio para el coeficiente de agrupamiento de todos los nodos de la red. Dicho coeficiente representa, para un nodo dado, el ratio entre el número de arcos existentes entre sus vecinos directos y el número máximo de arcos posibles entre estos **(Barabasi et al. 2004; Watts and Strogatz 1999)**.
5. **Centralidad de la red**: Valor entre 0 y 1 donde las topologías similares a la forma de estrella tendrán valores cercanos a 1 **(Dong and Horvath 2007)**.
6. **Número medio de vecinos (*degree or connectivity*)**: El grado medio o conectividad media de todos los nodos de la red que se define, para cada nodo, como el número de nodos con los que está conectado. Caracteriza la densidad de conexiones que presenta la red.

Además, se muestra un gráfico para cada una de las redes con la distribución del grado de todos sus nodos. En este gráfico se representa el número de nodos (eje Y) que tienen los diferentes valores de grado (eje X), lo cual permite conocer la distribución de estos últimos a lo largo de toda la red. Para verificar matemáticamente que los interactomas son redes libres de escala, tal y como se espera de cualquier red biológica **(Barabasi et al. 2004)**, se ajustó esta distribución a una ley de potencias de tipo $y=ax^b$ obteniendo en todos los casos valores significativos, como se muestra en **Figura 59**.

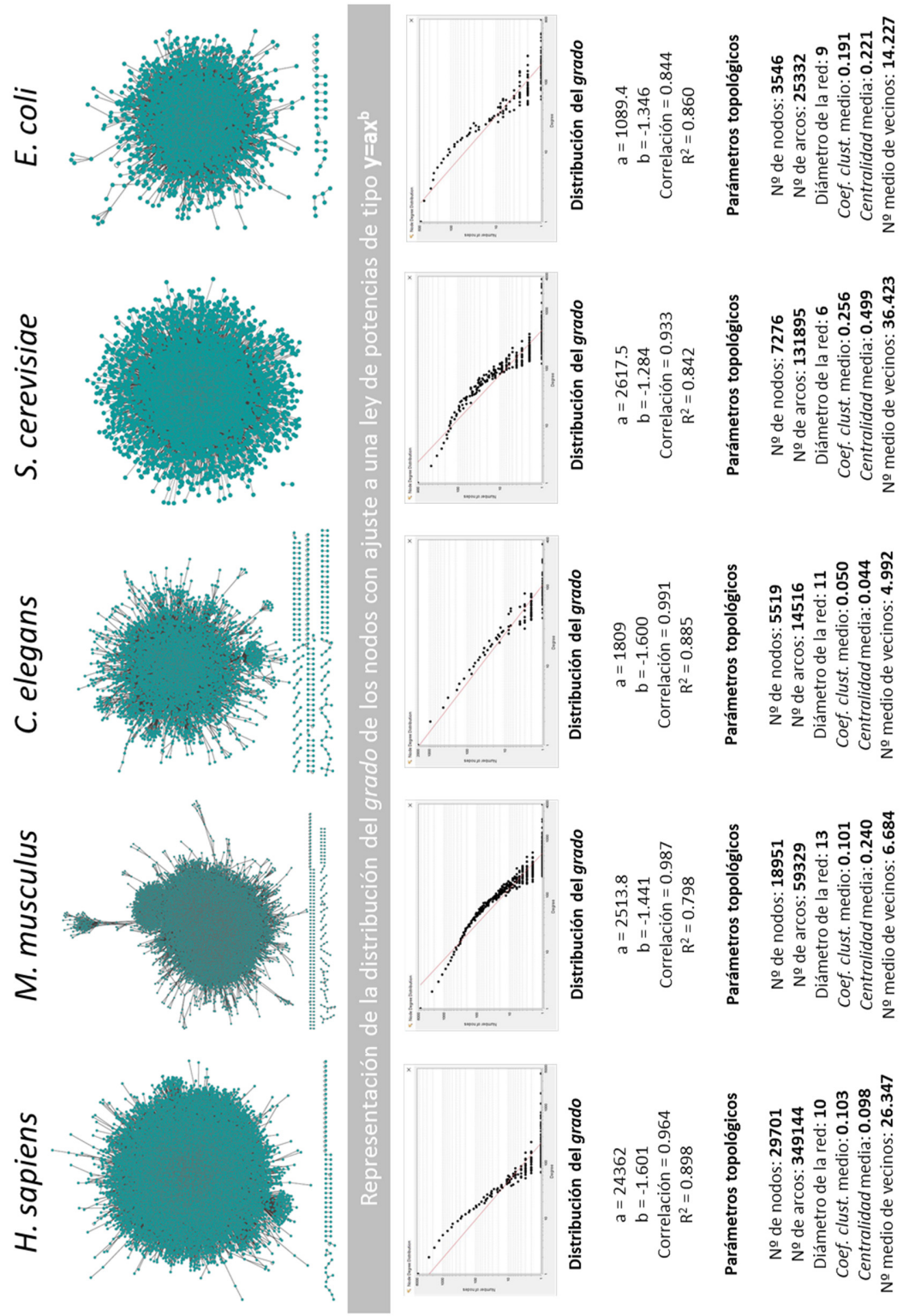


Figura 59. *Análisis comparativo de los parámetros topológicos de las redes generadas a partir de los interactomas de los organismos Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae y Escherichia coli.*

4.4.3 Análisis funcional de los interactomas de *H. sapiens* y *S. cerevisiae*

Con el objetivo de obtener información sobre los agrupamientos principales de cada red se utilizó la herramienta MCODE **(Bader and Hogue 2003)**, que analiza la red en busca de regiones altamente conectadas y permite extraer módulos a partir de ellas. Así, usando el algoritmo implementado por esta herramienta, se analizaron los interactomas de *H. sapiens* y *S. cerevisiae* para un mínimo de 2 experimentos por interacción y se extrajeron los 10 agrupamientos con mejor *score* para cada una de las dos redes. De esta manera, se obtuvieron 10 grupos de proteínas para cada organismo de forma no supervisada y partiendo simplemente de la topología de sus interactomas.

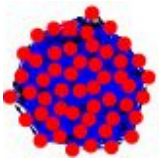
A continuación, para tratar de caracterizar biológicamente estos grupos de proteínas interconectadas, se efectuaron análisis de enriquecimiento funcional para los genes asociados a cada uno de ellos. Los resultados obtenidos se muestran en **Tabla 6** y **Tabla 7**. Como puede observarse, la mayoría de los grupos presentaron un enriquecimiento significativo en funciones esenciales para la célula.

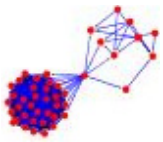
En casi todos los casos el algoritmo de GeneTerm Linker ofreció como resultado un metagrupo combinando términos de anotaciones en los espacios de GO, KEGG e InterPro con un valor p ajustado significativo y una cohesión de grupo específica expresada a través del parámetro *silhouette* tal y como se describe en **(Fontanillo et al. 2011)**.

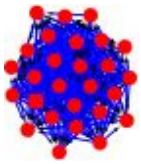
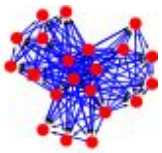
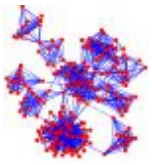
En el caso de los grupos *Hs10* y *Sc6*, GeneTerm Linker no consiguió generar metagrupos por lo que se utilizaron los algoritmos de Enrichr **(Kuleshov et al. 2016)** y GeneCodis **(Carmona-Saez et al. 2007; Nogales-Cadenas et al. 2009; Tabas-Madrid et al. 2012)** respectivamente y ambos presentaron enriquecimientos significativos.

Por último, en el caso de los grupos *Sc5* y *Sc10* no se encontraron anotaciones suficientes para efectuar un análisis de enriquecimiento pero se utilizó el sistema de búsqueda de proteínas de la base de datos Complex Portal **(Meldal et al. 2015)** para confirmar que ambos grupos correspondían a complejos descritos en la literatura previamente.

Tabla 6. *Agrupamientos o módulos obtenidos a partir del análisis del interactoma humano con el algoritmo MCODE* ***(Bader and Hogue 2003)****. Se muestra también el resultado del análisis de enriquecimiento funcional para los genes asociados a cada grupo. Dicho análisis está calculado en la mayoría de los casos con la herramienta GeneTerm Linker* ***(Fontanillo et al. 2011)****, a partir de cuyos resultados se extraen* ***(i)*** *el número de genes en cada metagrupo respecto al total de genes en el módulo MCODE,* ***(ii)*** *el número de genes del genoma de referencia que se asignarían a ese grupo funcional respecto al total de genes en el genoma,* ***(iii)*** *la significación del enriquecimiento y* ***(iv)*** *el coeficiente silhouette, valor numérico entre 0 y 1 que representa el grado de cohesión funcional de cada metagrupo. Para los grupos que no presentaban enriquecimiento funcional en GeneTermLinker se utilizaron las herramientas Enrichr* ***(Kuleshov et al. 2016)*** *o GeneCodis* **(Carmona-Saez et al. 2007; Nogales-Cadenas et al. 2009; Tabas-Madrid et al. 2012)***, extrayendo en ambos casos el grado de significación del enriquecimiento.*

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Hs1	 *Score*: 46,92 Nodos: 51 Arcos: 1173	RPL10 RPL10A RPL11 RPL12 RPL13 RPL14 RPL15 RPL18 RPL18A RPL19 RPL21 RPL22 RPL23 RPL23A RPL24 RPL3 RPL30 RPL31 RPL37A RPL4 RPL5 RPL6 RPL7 RPL7A RPL8 RPLP0 RPS11 RPS12 RPS13 RPS14 RPS15A RPS16 RPS19 RPS2 RPS20 RPS23 RPS25 RPS26 RPS3 RPS3A RPS4X RPS5 RPS6 RPS7 RPS8 RPS9 RPSA Rrbp	GO:0006414 translational elongation (BP) GO:0005829 cytosol (CC) GO:0031018 endocrine pancreas development (BP) GO:0044267 cellular protein metabolic process (BP) GO:0016070 RNA metabolic process (BP) GO:0016032 viral reproduction (BP) GO:0016071 mRNA metabolic process (BP) GO:0006415 translational termination (BP) GO:0019058 viral infectious cycle (BP) GO:0019083 viral transcription (BP) GO:0003735 structural constituent of ribosome (MF) 03010 Ribosome GO:0005840 ribosome (CC) GO:0005622 intracellular (CC) GO:0022625 cytosolic large ribosomal subunit (CC)
		Genes del módulo en metagrupo (genes totales en el módulo): **48 (50)** Genes del genoma en metagrupo (genes totales en el genoma)**: 155 (34208)** pValor ajustado: **1.13044e-113** Coeficiente Silhouette: **0.2366**	

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos
Módulo Hs2	 *Score*: 26,186 Nodos: 44 Arcos: 564	ADRM1 PSMA1 PSMA5 PSMB2 PSMC1 PSMC2 PSMC3 PSMC4 PSMC5 PSMC6 PSMD14 PSMD6 PSMD8 UCHL5	GO:0000502 proteasome complex (CC) GO:0006915 apoptosis (BP) GO:0002474 antigen processing and presentation of peptide antigen via MHC class I (BP) GO:0000209 protein polyubiquitination (BP) GO:0000278 mitotic cell cycle (BP) GO:0000082 G1/S transition of mitotic cell cycle (BP) GO:0000084 S phase of mitotic cell cycle (BP) GO:0042981 regulation of apoptosis (BP) GO:0051436 negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle (BP) GO:0051437 positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle (BP) GO:0034641 cellular nitrogen compound metabolic process (BP) GO:0031145 anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process (BP) GO:0051439 regulation of ubiquitin-protein ligase activity during mitotic cell cycle (BP) GO:0006977 DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest (BP) GO:0000216 M/G1 transition of mitotic cell cycle (BP) GO:0006521 regulation of cellular amino acid metabolic process (BP) GO:0000075 cell cycle checkpoint (BP) GO:0006508 proteolysis (BP) 03050 Proteasome GO:0016787 hydrolase activity (MF) GO:0016887 ATPase activity (MF) GO:0006200 ATP catabolic process (BP) GO:0005524 ATP binding (MF) GO:0000166 nucleotide binding (MF) IPR003593 ATPase, AAA+ type, core GO:0030163 protein catabolic process (BP) IPR005937 26S proteasome subunit P45 GO:0017111 nucleoside-triphosphatase activity (MF)
		Genes del módulo en metagrupo (genes totales en el módulo): **14 (43)** Genes del genoma en metagrupo (genes totales en el genoma)**: 18 (34208)** pValor ajustado: **6.95539e-39** Coeficiente Silhouette: **0. 4186**	

<table>
<tr><th></th><th>Módulo MCODE</th><th>Metagrupo
GeneTerm
Linker</th><th>Términos</th></tr>
<tr><td rowspan="2">Módulo Hs3</td><td rowspan="2"> 
Score: 24,154
Nodos: 27
Arcos: 314</td><td>MED1
MED10
MED12
MED13
MED14
MED17
MED21
MED24
MED26
MED27
MED30
MED4
MED6
MED7</td><td>GO:0001104 RNA polymerase II transcription cofactor activity
GO:0016592 mediator complex (CC)
GO:0006367 transcription initiation from RNA polymerase II promoter (BP)
GO:0003713 transcription coactivator activity (MF)
GO:0004872 receptor activity (MF)
GO:0045893 positive regulation of transcription, DNA-dependent (BP)
GO:0003712 transcription cofactor activity (MF)
GO:0030374 ligand-dependent nuclear receptor transcription coactivator activity (MF)
GO:0030521 androgen receptor signaling pathway (BP)
GO:0046966 thyroid hormone receptor binding (MF)
GO:0030518 steroid hormone receptor signaling pathway (BP)
GO:0042809 vitamin D receptor binding (MF)
GO:0045944 positive regulation of transcription from RNA polymerase II promoter (BP)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 14 (27)
Genes del genoma en metagrupo (genes totales en el genoma): 59 (34208)
pValor ajustado: 7.6372e-33
Coeficiente Silhouette: 0.4813</td></tr>
<tr><td rowspan="2">Módulo Hs4</td><td rowspan="2"> 
Score: 13,5
Nodos: 21
Arcos: 135</td><td>ARPC1B
CAPZA2
MYO5C
Myo1c
SPTAN1
SPTBN1</td><td>GO:0005516 calmodulin binding (MF)
GO:0051693 actin filament capping (BP)
GO:0005200 structural constituent of cytoskeleton (MF)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 6 (21)
Lista referencia: 63 (34208)
pValor ajustado: 1.62225e-12
Coeficiente Silhouette: 0.5059</td></tr>
<tr><td rowspan="2">Módulo Hs5</td><td rowspan="2"> 
Score: 12,622
Nodos: 197
Arcos: 1240</td><td>BRD8
DMAP1
KAT5
MORF4L1
Ppp2ca
Ruvbl1
TRRAP
YEATS4</td><td>GO:0035267 NuA4 histone acetyltransferase complex (CC)
GO:0040008 regulation of growth (BP)
GO:0043967 histone H4 acetylation (BP)
GO:0043968 histone H2A acetylation (BP)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 8 (181)
Genes del genoma en metagrupo (genes totales en el genoma): 38 (34208)
pValor ajustado: 2.24534e-11
Coeficiente Silhouette: 0.7353</td></tr>
</table>

<table>
<tr><th></th><th>Módulo MCODE</th><th>Metagrupo GeneTerm Linker</th><th>Términos</th></tr>
<tr><td rowspan="2">Módulo Hs6</td><td rowspan="2"> 
Score: 8
Nodos: 8
Arcos: 28</td><td>COG1 COG2
COG3 COG4
COG5 COG6
COG7 COG8</td><td>GO:0005794 Golgi apparatus (CC)
GO:0017119 Golgi transport complex (CC)
GO:0016020 membrane (CC)
GO:0000139 Golgi membrane (CC)
GO:0015031 protein transport (BP)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 8 (8)
Genes del genoma en metagrupo (genes totales en el genoma): 10 (34208)
pValor ajustado: 9.68427e-31
Coeficiente Silhouette: 0.6098</td></tr>
<tr><td rowspan="2">Módulo Hs7</td><td rowspan="2"> 
Score: 6,825
Nodos: 64
Arcos: 222</td><td>CCNH CDK7
ERCC2 ERCC3
MCM2 MCM3
MCM4 MCM5
MCM6 MCM7
MNAT1</td><td>GO:0000075 cell cycle checkpoint (BP)
GO:0000082 G1/S transition of mitotic cell cycle (BP)
GO:0000084 S phase of mitotic cell cycle (BP)
GO:0000216 M/G1 transition of mitotic cell cycle (BP)
04110 Cell cycle
GO:0006271 DNA strand elongation during DNA replication (BP)
03030 DNA replication
GO:0006270 DNA replication initiation (BP)
GO:0042555 MCM complex (CC)
IPR001208 Mini-chromosome maintenance, DNA-dependent ATPase
GO:0004386 helicase activity (MF)
GO:0004003 ATP-dependent DNA helicase activity (MF)
GO:0006268 DNA unwinding during replication (BP)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 11 (62)
Genes del genoma en metagrupo (genes totales en el genoma): 71 (34208)
pValor ajustado: 6.38576e-19
Coeficiente Silhouette: 0.6567</td></tr>
<tr><td rowspan="2">Módulo Hs8</td><td rowspan="2"> 
Score: 6,64
Nodos: 26
Arcos: 84</td><td>ACTL6A EP400
EPC1 ING3</td><td>GO:0043967 histone H4 acetylation (BP)
GO:0043968 histone H2A acetylation (BP)
GO:0035267 NuA4 histone acetyltransferase complex (CC)
GO:0040008 regulation of growth (BP))</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 4 (22)
Genes del genoma en metagrupo (genes totales en el genoma): 11 (34208)
pValor ajustado: 4.21915e-11
Coeficiente Silhouette: 0.6207</td></tr>
</table>

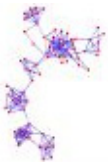
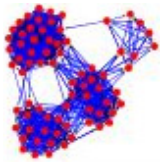
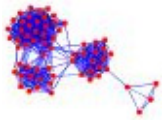
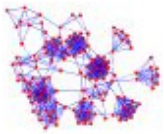
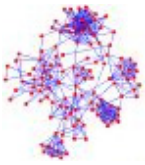
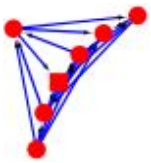
<table>
<tr><th></th><th>Módulo MCODE</th><th>Metagrupo GeneTerm Linker</th><th>Términos</th></tr>
<tr><td rowspan="2">Módulo Hs9</td><td rowspan="2"> 
Score: 6,225
Nodos: 81
Arcos: 249</td><td>AP2A2 POLR2B POLR2C POLR2D POLR2E POLR2F POLR2H POLR2I TAF7 TBP</td><td>05016 Huntington's disease
GO:0006367 transcription initiation from RNA polymerase II promoter (BP)
GO:0006368 RNA elongation from RNA polymerase II promoter (BP)
00230 Purine metabolism
GO:0006289 nucleotide-excision repair (BP)
GO:0006370 mRNA capping (BP)
GO:0050434 positive regulation of viral transcription (BP)
GO:0006283 transcription-coupled nucleotide-excision repair (BP)
GO:0003899 DNA-directed RNA polymerase activity (MF)
GO:0005665 DNA-directed RNA polymerase II, core complex (CC)
00240 Pyrimidine metabolism
03020 RNA polymerase
GO:0006383 transcription from RNA polymerase III promoter (BP)</td></tr>
<tr><td colspan="2">Genes del módulo en metagrupo (genes totales en el módulo): 10 (74)
Genes del genoma en metagrupo (genes totales en el genoma): 54 (34208)
pValor ajustado: 2.64142e-17
Coeficiente Silhouette: 0.6470</td></tr>
<tr><td rowspan="2">Módulo Hs10</td><td rowspan="2"> 
Score: 5,714
Nodos: 8
Arcos: 20</td><td>Emc6
Emc1
EMC3
EMC7
EMC4
EMC10
EMC1
MMGT1</td><td>* Enrichr
GO:0072546 ER membrane protein complex</td></tr>
<tr><td colspan="2">pValor: 2.838e-24</td></tr>
</table>

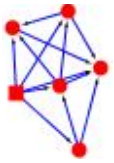
Tabla 7. *Agrupamientos o módulos obtenidos a partir del análisis del interactoma de S. cerevisiae con el algoritmo MCODE* ***(Bader and Hogue 2003)****. Se muestra también el resultado del análisis de enriquecimiento funcional para los genes asociados a cada grupo. Dicho análisis está calculado en la mayoría de los casos con la herramienta GeneTerm Linker* ***(Fontanillo et al. 2011)****, a partir de cuyos resultados se extraen* ***(i)*** *el número de genes en cada metagrupo respecto al total de genes en el módulo MCODE,* ***(ii)*** *el número de genes del genoma de referencia que se asignarían a ese grupo funcional respecto al total de genes en el genoma,* ***(iii)*** *la significación del enriquecimiento y* ***(iv)*** *el coeficiente silhouette, valor numérico entre 0 y 1 que representa el grado de cohesión funcional de cada metagrupo. Para los grupos que no presentaban enriquecimiento funcional en GeneTermLinker se utilizaron las herramientas Enrichr* ***(Kuleshov et al. 2016)*** *o GeneCodis* **(Carmona-Saez et al. 2007; Nogales-Cadenas et al. 2009; Tabas-Madrid et al. 2012)***, extrayendo en ambos casos el grado de significación del enriquecimiento. Por último, para los metagrupos cuyos genes no presentan suficientes anotaciones funcionales, se utilizó la base de datos Complex Portal* ***(Meldal et al. 2015)*** *para confirmar que dichos grupos correspondían a complejos proteicos descritos en la literatura previamente.*

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Sc1	 *Score*: 21,644 Nodos: 74 Arcos: 790	BLM10 CDC6 CIC1 ECM29 HSM3 NAS6 PRE1 RAD23 RPN1 RPN10 RPN11 RPN13 RPN14 RPN2 RPN7 RPT1 RPT2 RPT3 RPT4 RPT5 RPT6 SEM1 SGF73 UBP6 UBP	GO:0000502 proteasome complex (CC) 03050 Proteasome GO:0008540 proteasome regulatory particle, base subcomplex (CC) GO:0070682 proteasome regulatory particle assembly (BP) GO:0030163 protein catabolic process (BP) GO:0006511 ubiquitin-dependent protein catabolic process (BP) GO:0017111 nucleoside-triphosphatase activity (MF) GO:0016887 ATPase activity (MF) IPR003593 ATPase, AAA+ type, core IPR003959 ATPase, AAA-type, core IPR003960 ATPase, AAA-type, conserved site IPR005937 26S proteasome subunit P45 GO:0045899 positive regulation of RNA polymerase II transcriptional preinitiation complex assembly (BP)
		Genes del módulo en metagrupo (genes totales en el módulo): **25 (74)** Genes del genoma en metagrupo (genes totales en el genoma)**: 85 (7109)** pValor ajustado: **4.19783e-31** Coeficiente Silhouette: **0.4965**	

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Sc2	 *Score*: 16,889 Nodos: 64 Arcos: 532	BMS1 BUD21 CRM1 MPP10 NAN1 NOC4 NOP4 NOP7 NUG1 PWP2 RRP9 SPB4 TIF6 UTP10 UTP15 UTP22 UTP4 UTP6 UTP7	GO:0030686 90S preribosome (CC) 03008 Ribosome biogenesis in eukaryotes GO:0032040 small-subunit processome (CC) GO:0000462 maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (BP) GO:0030515 snoRNA binding (MF) GO:0000447 endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (BP) GO:0000472 endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (BP) IPR001680 WD40 repeat IPR019782 WD40 repeat 2 IPR017986 WD40-repeat-containing domain IPR019781 WD40 repeat, subgroup GO:0000480 endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (BP) GO:0045943 positive regulation of transcription from RNA polymerase I promoter (BP) GO:0034455 t-UTP complex (CC)
		Genes del módulo en metagrupo (genes totales en el módulo): **19 (64)** Genes del genoma en metagrupo (genes totales en el genoma)**: 100 (7109)** pValor ajustado: **5.74942e-21** Coeficiente Silhouette: **0.6000**	
Módulo Sc3	 *Score*: 10,64 Nodos: 151 Arcos: 799	APC1 APC11 APC2 APC4 APC5 APC9 ASR1 CDC16 CDC23 CDC26 CDC27 DOC1 MND2 PPH22 PRP19 SWM1 TPK2	GO:0007067 mitosis (BP) GO:0051301 cell division (BP) 04113 Meiosis - yeast GO:0005680 anaphase-promoting complex (CC) GO:0031145 anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process (BP) GO:0016567 protein ubiquitination (BP) 04111 Cell cycle - yeast 04120 Ubiquitin mediated proteolysis GO:0004842 ubiquitin-protein ligase activity (MF) GO:0030071 reg. of mitotic metaphase/anaphase transition (BP) GO:0031497 chromatin assembly (BP)
		Genes del módulo en metagrupo (genes totales en el módulo): **17 (151)** Genes del genoma en metagrupo (genes totales en el genoma)**: 124 (7109)** pValor ajustado: **7.27886e-10** Coeficiente Silhouette: **0.7204**	

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Sc4	 *Score*: 7,339 Nodos: 125 Arcos: 455	RET1 RPA12 RPA14 RPA190 RPA34 RPA43 RPA49 RPC10 RPC17 RPC25 RPC31 RPC34 RPC37 RPC53 RPC82 RPO31	GO:0003899 DNA-directed RNA polymerase activity (MF) 00230 Purine metabolism 00240 Pyrimidine metabolism 03020 RNA polymerase GO:0005666 DNA-directed RNA polymerase III complex (CC) GO:0042797 tRNA transcription from RNA polymerase III promoter (BP) GO:0005736 DNA-directed RNA polymerase I complex (CC) GO:0042790 transcription of nuclear rRNA large RNA polymerase I transcript (BP) GO:0016740 transferase activity (MF) GO:0016779 nucleotidyltransferase activity (MF)
		Genes del módulo en metagrupo (genes totales en el módulo): **16 (125)** Genes del genoma en metagrupo (genes totales en el genoma)**: 32 (7109)** pValor ajustado: **1.48349e-20** Coeficiente Silhouette: **0.5523**	
Módulo Sc5	 *Score*: 6,333 Nodos: 7 Arcos: 19	POP7 POP6 POP4 POP5 POP8 POP1 RMP1	** Complex Portal Nucleolar ribonuclease P complex
		(Complex Portal EBI-2877618) https://www.ebi.ac.uk/intact/complex/details/EBI-2877618	
Módulo Sc6	 *Score*: 6 Nodos: 7 Arcos: 18	ELO1 TSC13 GAS3 YHR140W ERG25 PHO86 GPI2	*** GeneCodis GO:0016021: integral to membrane (CC) GO:0016020: membrane (CC) GO:0005783: endoplasmic reticulum (CC)
		pValor: **7.34296e-06**	

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Sc7	 *Score*: 5,758 Nodos: 67 Arcos: 190	CHL4 CSE4 CTF3 HTA1 IML3 MCM21 MCM22 NSL1 PSH1 SWC4	GO:0034087 establishment of mitotic sister chromatid cohesion (BP) GO:0007067 mitosis (BP) GO:0051301 cell division (BP) GO:0007049 cell cycle (BP) GO:0007126 meiosis (BP) GO:0000775 chromosome, centromeric region (CC) GO:0005694 chromosome (CC) GO:0000776 kinetochore (CC) GO:0000777 condensed chromosome kinetochore (CC) GO:0007059 chromosome segregation (BP)
		Genes del módulo en metagrupo (genes totales en el módulo): **10 (65)** Genes del genoma en metagrupo (genes totales en el genoma)**: 138 (7109)** pValor ajustado: **3.95154e-07** Coeficiente Silhouette: **0.8144**	
Módulo Sc8	 *Score*: 5,444 Nodos: 55 Arcos: 147	COG1 COG2 COG3 COG4 COG5 COG6 COG7 COG8 STV1	GO:0000139 Golgi membrane (CC) GO:0005794 Golgi apparatus (CC) GO:0006891 intra-Golgi vesicle-mediated transport (BP) GO:0032258 CVT pathway (BP) GO:0017119 Golgi transport complex (CC) GO:0000301 retrograde transport, vesicle recycling within Golgi (BP)
		Genes del módulo en metagrupo (genes totales en el módulo): **9 (55)** Genes del genoma en metagrupo (genes totales en el genoma)**: 68 (7109)** pValor ajustado: **1.74407e-09** Coeficiente Silhouette: **0.7639**	
Módulo Sc9	 *Score*: 5,2 Nodos: 16 Arcos: 39	MRPL10 MRPL16 MRPL28 MRPL7	GO:0003735 structural constituent of ribosome (MF) GO:0030529 ribonucleoprotein complex (CC) GO:0005840 ribosome (CC) GO:0005739 mitochondrion (CC) GO:0032543 mitochondrial translation (BP) GO:0005762 mitochondrial large ribosomal subunit (CC
		Genes del módulo en metagrupo (genes totales en el módulo): **4 (16)** Genes del genoma en metagrupo (genes del genoma)**: 41 (7109)** pValor ajustado: **1.64872e-06** Coeficiente Silhouette: **0.7965**	

	Módulo MCODE	Metagrupo GeneTerm Linker	Términos funcionales
Módulo Sc10	 *Score*: 5,2 Nodos: 6 Arcos: 13	ELP3 ELP4 ELP6 IKI1 IKI3 KTI12	** Complex Portal Elongator holoenzyme complex
		(Complex Portal EBI-1254644) https://www.ebi.ac.uk/intact/complex/details/EBI-1254644	

Finalmente, tal y como se muestra en **Figura 60**, se establecieron las equivalencias entre los grupos encontrados para ambos organismos obteniendo un alto índice de solapamiento en las funciones biológicas para las que estos mostraron un enriquecimiento significativo.

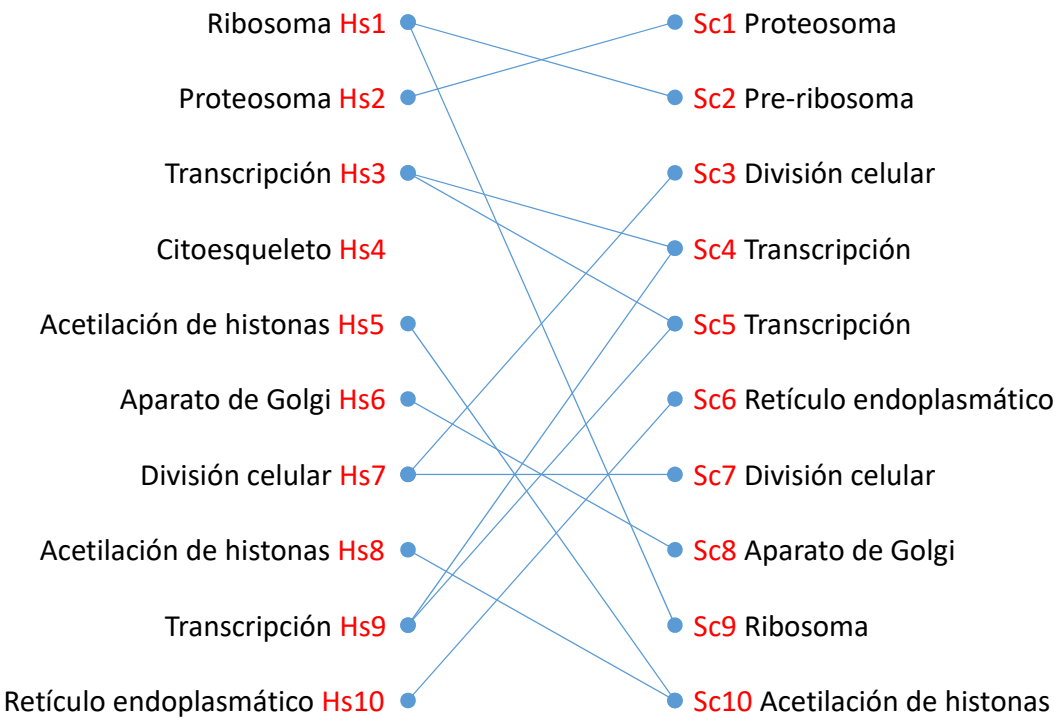


Figura 60. *Equivalencias entre las funciones biológicas principales obtenidas a partir del análisis del enriquecimiento funcional llevado a cabo con las listas de genes asociados a cada módulo resultante del algoritmo MCODE para los interactomas de H. sapiens y S. cerevisiae.*

4.4.4 Cálculo del interactoma binario de *Homo Sapiens*

El último análisis llevado a cabo con los interactomas generados por APID Interactomes está relacionado con los métodos de detección de interacciones empleados por la comunidad científica. Para el caso del interactoma de *H. sapiens,* a lo largo de todas las interacciones procesadas por el algoritmo de integración descrito en esta Tesis Doctoral, se detectaron un total de 160 métodos diferentes de detección de interacciones, todos ellos asociados a su término correspondiente en la ontología diseñada por el grupo HUPO PSI-MI.

Estos métodos, como ya se explicó anteriormente, no siempre detectan la interacción directa entre 2 proteínas, sino que en algunos casos detectan interacciones múltiples o *co-complex* **(De Las Rivas and Fontanillo 2010)** que deben ser expandidas, tal y como se mostraba en **Figura 29**.

Desde hace un tiempo, se están llevando a cabo proyectos que tratan de identificar lo que sería el interactoma humano utilizando solo interacciones directas entre dos proteínas o binarias, es decir, aquellas detectadas a través de métodos binarios. Entre estos proyectos cabe destacar *HuRI: The Human Reference Protein Interactome Mapping Project* (72), a partir de los datos publicados en **(Rual et al. 2005; Rolland et al. 2014)** e *Interactome3D* **(Mosca et al. 2012)** (73), que añade además datos estructurales.

En este contexto, se estudiaron los diferentes métodos de detección de interacciones procesados por APID Interactomes para establecer cuáles de ellos eran métodos binarios tal y como se muestra en **Tabla 8**. Este análisis resultó en un total de 39 métodos binarios, es decir, aproximadamente un 25% del total de los métodos empleados para detectar las interacciones contenidas en la base de datos de APID Interactomes.

Tabla 8. *Listado de métodos de detección de interacciones procesados por el algoritmo de integración de APID Interactomes para el interactoma de H. sapiens. Aparecen marcados en azul los métodos binarios.*

PSI-MI ID	Descripción del término	PSI-MI ID	Descripción del término
MI:0004	affinity chromatography technology	MI:0424	protein kinase assay
MI:0006	anti bait coimmunoprecipitation	MI:0425	kinase scintillation proximity assay
MI:0007	anti tag coimmunoprecipitation	MI:0426	light microscopy
MI:0008	array technology	MI:0428	imaging technique
MI:0010	beta galactosidase complementation	MI:0434	phosphatase assay
MI:0011	beta lactamase complementation	MI:0435	protease assay
MI:0012	bioluminescence resonance energy transfer	MI:0437	protein three hybrid
MI:0013	biophysical	MI:0440	saturation binding
MI:0016	circular dichroism	MI:0492	in vitro
MI:0017	classical fluorescence spectroscopy	MI:0493	in vivo
MI:0018	two hybrid	MI:0510	homogeneous time resolved fluorescence
MI:0019	coimmunoprecipitation	MI:0512	zymography

MI:0020	transmission electron microscopy	MI:0515	methyltransferase assay
MI:0027	cosedimentation	MI:0516	methyltransferase radiometric assay
MI:0028	cosedimentation in solution	MI:0588	three hybrid
MI:0029	cosedimentation through density gradient	MI:0605	enzymatic footprinting
MI:0030	cross-linking study	MI:0655	lambda repressor two hybrid
MI:0031	protein cross-linking with a bifunctional reagent	MI:0663	confocal microscopy
MI:0034	display technology	MI:0676	tandem affinity purification
MI:0038	dynamic light scattering	MI:0678	antibody array
MI:0040	electron microscopy	MI:0686	unspecified method
MI:0045	experimental interaction detection	MI:0726	reverse two hybrid
MI:0047	far western blotting	MI:0727	lexa b52 complementation
MI:0048	filamentous phage display	MI:0728	gal4 vp16 complementation
MI:0049	filter binding	MI:0729	luminescence based mammalian interactome mapping
MI:0051	fluorescence technology	MI:0807	comigration in gel electrophoresis
MI:0052	fluorescence correlation spectroscopy	MI:0808	comigration in sds page
MI:0053	fluorescence polarization spectroscopy	MI:0809	bimolecular fluorescence complementation
MI:0054	fluorescence-activated cell sorting	MI:0813	proximity ligation assay
MI:0055	fluorescent resonance energy transfer	MI:0814	protease accessibility laddering
MI:0065	isothermal titration calorimetry	MI:0825	x-ray fiber diffraction
MI:0066	lambda phage display	MI:0826	x ray scattering
MI:0067	light scattering	MI:0841	phosphotransferase assay
MI:0069	mass spectrometry studies of complexes	MI:0858	immunodepleted coimmunoprecipitation
MI:0071	molecular sieving	MI:0870	demethylase assay
MI:0077	nuclear magnetic resonance	MI:0872	atomic force microscopy
MI:0081	peptide array	MI:0880	atpase assay
MI:0084	phage display	MI:0889	acetylase assay
MI:0089	protein array	MI:0892	solid phase assay
MI:0090	protein complementation assay	MI:0905	amplified luminescent proximity homogeneous assay
MI:0091	chromatography technology	MI:0920	ribonuclease assay
MI:0095	proteinchip(r) on a surface-enhanced laser desorption/ionization	MI:0921	surface plasmon resonance array
MI:0096	pull down	MI:0943	detection by mass spectrometry
MI:0097	reverse ras recruitment system	MI:0944	mass spectrometry study of hydrogen/deuterium exchange

MI:0099	scintillation proximity assay	MI:0947	bead aggregation assay
MI:0104	static light scattering	MI:0949	gdp/gtp exchange assay
MI:0107	surface plasmon resonance	MI:0953	polymerization
MI:0108	t7 phage display	MI:0964	infrared spectroscopy
MI:0111	dihydrofolate reductase reconstruction	MI:0968	biosensor
MI:0112	ubiquitin reconstruction	MI:0969	bio-layer interferometry
MI:0114	x-ray crystallography	MI:0979	oxidoreductase assay
MI:0115	yeast display	MI:0982	electrophoretic mobility-based method
MI:0226	ion exchange chromatography	MI:0984	deaminase assay
MI:0227	reverse phase chromatography	MI:0990	cleavage assay
MI:0231	mammalian protein protein interaction trap	MI:0991	lipid cleavage assay
MI:0232	transcriptional complementation assay	MI:0997	ubiquitinase assay
MI:0254	genetic interference	MI:0998	deubiquitinase assay
MI:0256	rna interference	MI:1000	hydroxylase assay
MI:0276	blue native page	MI:1010	neddylase assay
MI:0369	lex-a dimerization assay	MI:1016	fluorescence recovery after photobleaching
MI:0370	tox-r dimerization assay	MI:1017	rna immunoprecipitation
MI:0397	two hybrid array	MI:1019	protein phosphatase assay
MI:0398	two hybrid pooling approach	MI:1024	scanning electron microscopy
MI:0399	two hybrid fragment pooling approach	MI:1037	Split renilla luciferase complementation
MI:0400	affinity technology	MI:1038	silicon nanowire field-effect transistor
MI:0401	biochemical	MI:1086	equilibrium dialysis
MI:0402	chromatin immunoprecipitation assay	MI:1103	solution state nmr
MI:0404	comigration in non denaturing gel electrophoresis	MI:1104	solid state nmr
MI:0405	competition binding	MI:1112	two hybrid prey pooling approach
MI:0406	deacetylase assay	MI:1138	decarboxylation assay
MI:0410	electron tomography	MI:1147	ampylation assay
MI:0411	enzyme linked immunosorbent assay	MI:1203	split luciferase complementation
MI:0412	electrophoretic mobility supershift assay	MI:1204	split firefly luciferase complementation
MI:0413	electrophoretic mobility shift assay	MI:1235	thermal shift binding
MI:0415	enzymatic study	MI:1247	mst
MI:0416	fluorescence microscopy	MI:1313	proximity labelling technology
MI:0417	footprinting	MI:1314	proximity-dependent biotin identification
MI:0419	gtpase assay	MI:1354	lipase assay

MI:0420	kinase homogeneous time resolved fluorescence	MI:1356	validated two hybrid
MI:0423	in-gel kinase assay	MI:2189	avexis

Usando esta información se extrajeron de la base de datos todas las interacciones entre proteínas humanas, no interespecie, detectadas mediante el uso de cualquiera de esos 39 métodos binarios. A continuación, se generó un fichero tabulado y se importó a Cytoscape para construir así la red correspondiente al interactoma binario de *H. sapiens,* que puede verse en **Figura 61**.

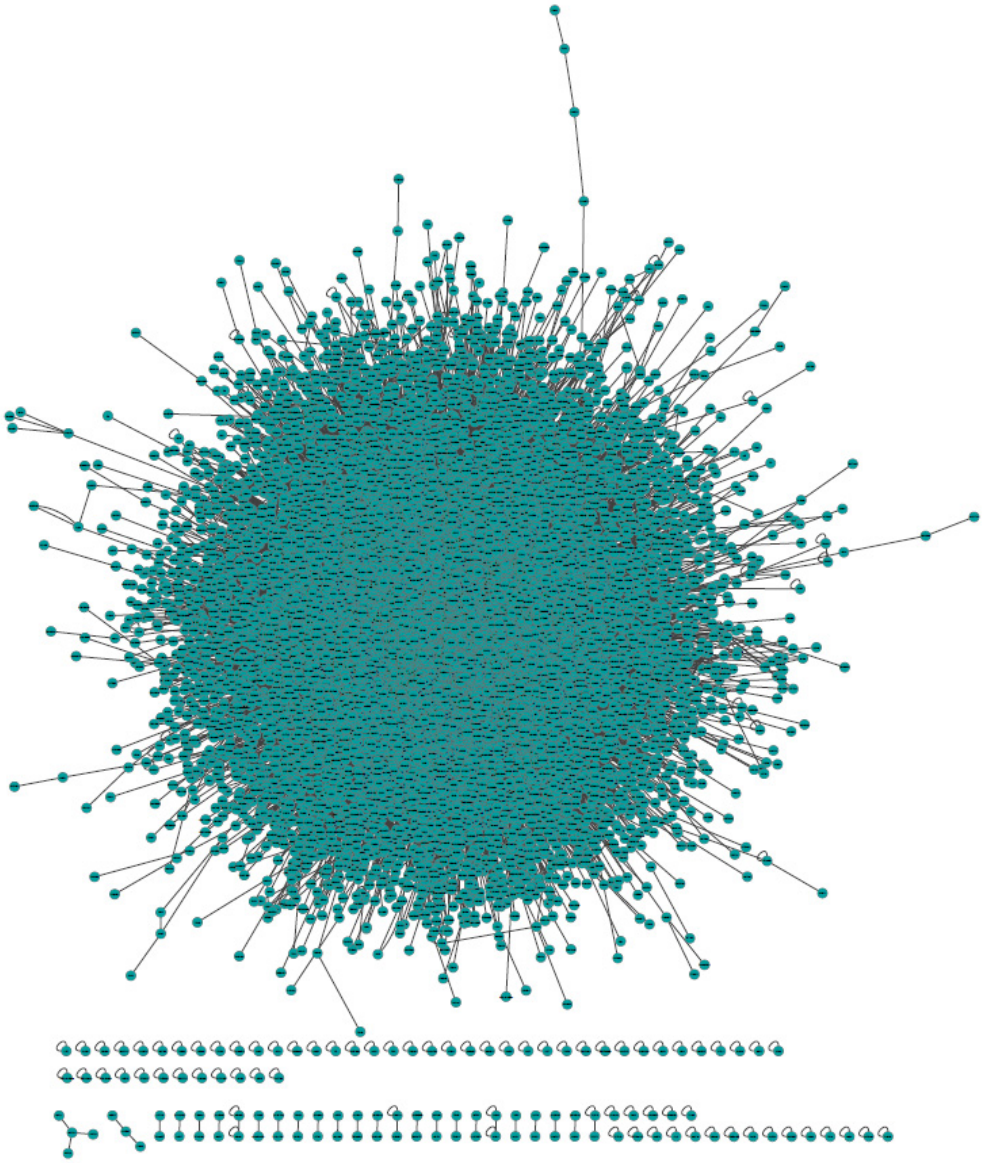


Figura 61. *Red correspondiente al interactoma humano, con interacciones descritas solo mediante métodos binarios y sin incluir aquellas con proteínas de otras especies (interacciones interespecie).*

El interactoma binario de *H. sapiens* que se generó a partir de los datos contenidos en APID Interactomes comprendía un total de 12137 proteínas y 64194 interacciones únicas entre estas. En **Tabla 9** puede verse la comparativa entre el interactoma humano binario calculado en esta Tesis Doctoral y los disponibles en *Interactome3D* y *HuRI*.

Tabla 9. *Tabla comparativa con el número de proteínas e interacciones únicas contenidas en los interactomas humanos binarios de APID Interactomes, Interactome3D y HuRI.*

	Proteínas	**Interacciones únicas**
APID Interactomes **(Alonso-López et al. 2016)**	12137	64194
Interactome3D **(Mosca et al. 2012)**	12839	66452
HuRI **(Rual et al. 2005; Rolland et al. 2014)**	8392	49226

5 DISCUSIÓN

El capítulo anterior de Resultados ha presentado en detalle la arquitectura, los desarrollos y los contenidos presentes en la plataforma bioinformática APID Interactomes, tanto a nivel de estructura y tipo de datos como de la información biológica que se puede obtener por el análisis de los interactomas construidos.

En este marco, la Discusión que cierra esta Tesis Doctoral, no pretende hacer un comentario más o menos ilustrado de los Resultados ya explicados, sino que se plantea como un análisis complementario de las distintas bases de datos de interacciones proteína-proteína que existen, comparando sus características, coberturas y los criterios que sirven para dar validez o confianza a las interacciones o asociaciones entre proteínas que reportan (que a menudo no son sólo de tipo PPIs). Pensamos que esta aproximación da más valor y realce a los Resultados y complementa su significación.

5.1 Características diferenciales de APID frente a otras plataformas similares

Como se indicó en la sección **Introducción y estado del arte**, existen múltiples bases de datos que proporcionan interacciones y otras asociaciones funcionales entre proteínas. En **Tabla 10** pueden observarse, de forma resumida, algunas de las diferencias más importantes entre las diferentes plataformas bioinformáticas.

***Tabla 10**. Tabla comparativa entre diferentes plataformas de datos de interacciones de proteínas.*

	Interactomas	*Visualizador de redes*	*Solo datos experimentales*	*Score*
APID Interactomes **(Alonso-López et al. 2016)**	Si (829 organismos)	Si	Si	No
Mentha **(Calderone et al. 2013)**	No	Si	Si	Si
iRefWeb **(Turner et al. 2010)**	No	No	No	No
HINT **(Das and Yu 2012)**	Si (12 organismos)	No	No	No
STRING **(von Mering et al. 2003)**	No	Si	No	Si
GeneMania **(Warde-Farley et al. 2010)**	No	Si	No	Si

Si se toma como criterio comparativo la funcionalidad de estas meta-bases de datos, Mentha aparece como el proyecto más cercano a APID Interactomes debido a su concepción y objetivos. Aun así, existen algunas diferencias importantes entre ambos tal y como se detalla a continuación:

1. Mentha tiene menos interacciones en su base de datos y estas están sujetas a la conversión que cada base de datos primaria haya hecho de sus datos para adaptarlos al formato PSI-MI TAB usado por PSICQUIC. Un ejemplo de esto sería la expansión de complejos ya que este formato, al ser tabulado, solo puede almacenar interacciones binarias. Como consecuencia, disponen de un proceso de actualización mucho menos costoso computacional y temporalmente, pero obtienen menos cobertura y menos detalle para cada interacción. Los creadores de Mentha diseñaron un proceso automatizado para conseguir actualizar semanalmente su base de datos. Esta frecuencia de actualización resulta inviable en una integración no automatizada como la que se lleva a cabo en APID Interactomes ya que el proceso de cálculo y anotación es muy complejo. La estrategia de Mentha para conseguirlo consiste en obtener sus datos, no directamente de las bases de datos primarias, sino de la interfaz que PSICQUIC proporciona. Este proceso se lleva a cabo exactamente de la misma manera que obtiene los datos, por ejemplo, la aplicación Cytoscape (con su opción de importar datos de servicios web, tal y como se mostraba en **Figura 5**). Mentha, además, solo utiliza los datos de 5 de las 36 bases de datos presentes en PSICQUIC: MINT, IntAct, DIP, MatrixDB y BioGRID. Estas fuentes de datos son muy similares a las de APID Interactomes, pero APID Interactomes obtiene sus datos a partir de los ficheros originales en formato PSI-MI XML consiguiendo así una mejor cobertura, tal y como se discutirá en el próximo apartado.

2. En el proceso de integración de interacciones implementado en Mentha no se actualizan los identificadores únicos de las proteínas, tal y como se hace en APID Interactomes. Esto provoca redundancia e inconsistencias en la base de datos ya que una misma proteína aparece con identificadores diferentes que, en consecuencia, se tratan como entidades biológicas diferentes. En **Figura 62** puede verse un ejemplo de este problema donde, al buscar el gen CT45A5, se nos ofrecen dos proteínas aparentemente diferentes según sus identificadores: Q6NSH3 y P0DMU8. Si se comprueban estos identificadores en UniProt puede verse que Q6NSH3 es un identificador obsoleto que hace referencia a la misma proteína que el identificador actual P0DMU8. Incluso UniProt nos da la posibilidad de examinar la historia de sus identificadores, pudiendo comprobarse a través de dicha historia para el caso de P0DMU8, la fecha exacta en la que este identificador reemplazó a Q6NSH3.

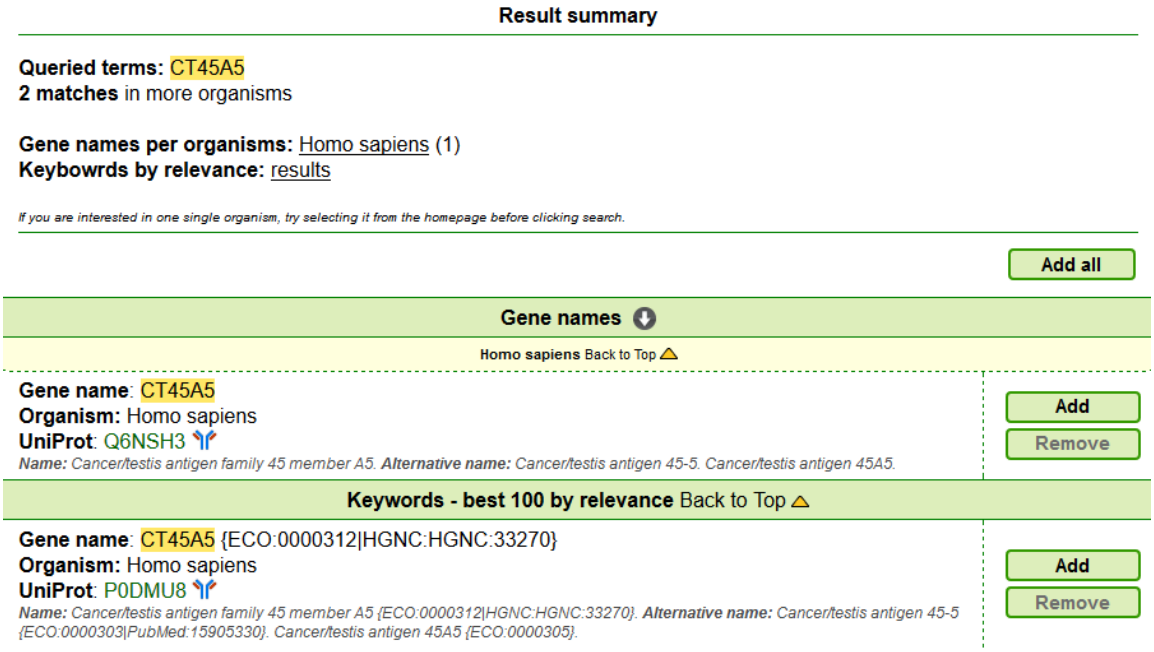
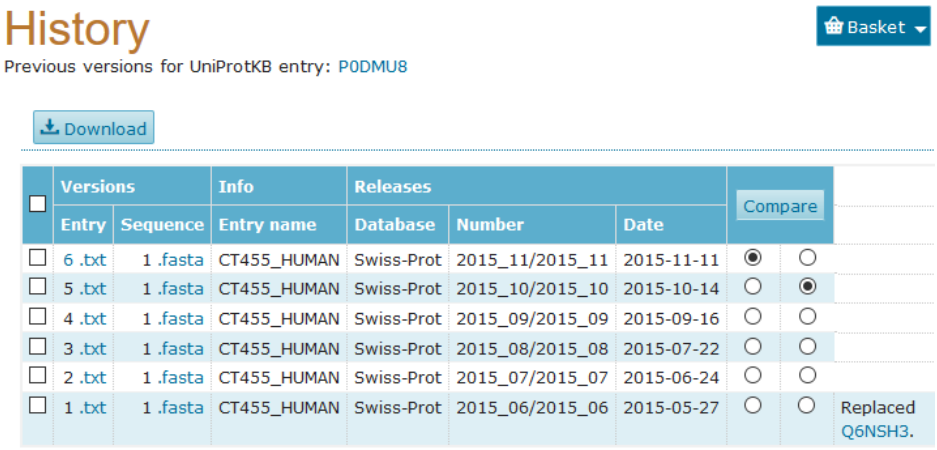


Figura 62. *Captura de pantalla de una búsqueda en el servidor web de Mentha. Puede comprobarse que para la búsqueda del gen CT45A5 se devuelven resultados que contienen identificadores obsoletos (Q6NSH3).*

En **Figura 63** puede observarse la tabla extraída de la web de UniProt donde aparece la historia del identificador P0DMU8. Allí aparecen cinco actualizaciones sobre el contenido del registro de esta proteína desde mediados del año 2015 y justo a continuación se detalla la fecha (27-05-2015) en que este identificador pasó a sustituir al anterior (Q6NSH3).



History

Basket

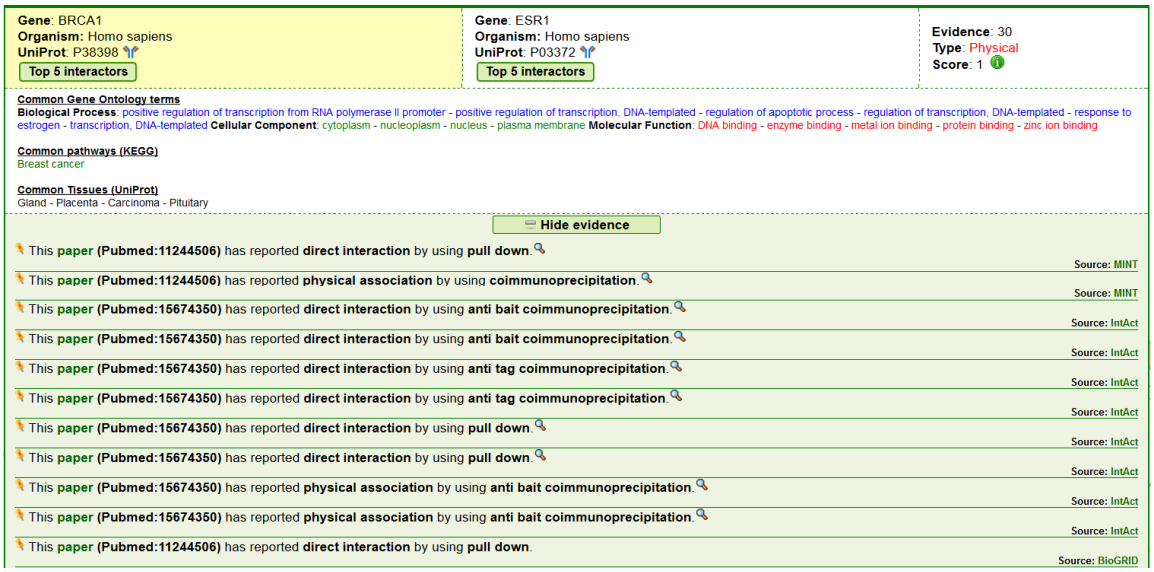
Previous versions for UniProtKB entry: P0DMU8

Download

	Versions		Info	Releases			Compare		
	Entry	Sequence	Entry name	Database	Number	Date			
☐	6 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_11/2015_11	2015-11-11	◉	○	
☐	5 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_10/2015_10	2015-10-14	○	◉	
☐	4 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_09/2015_09	2015-09-16	○	○	
☐	3 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_08/2015_08	2015-07-22	○	○	
☐	2 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_07/2015_07	2015-06-24	○	○	
☐	1 .txt	1 .fasta	CT455_HUMAN	Swiss-Prot	2015_06/2015_06	2015-05-27	○	○	Replaced Q6NSH3.

Figura 63. *Captura de pantalla de la web de UniProt con la historia de identificador P0DMU8.*

3. Mentha calcula un *score* para dar un valor de calidad a cada interacción basado en el número de evidencias que soportan dicha interacción. El problema es que consideran evidencia al mero hecho de que una interacción esté registrada en una base de datos primaria, sin entrar a valorar si es única o no. En APID Interactomes, el hecho de que una interacción esté registrada en una base de datos primaria se considera un *curation event* y, para tratar de asignar una determinada calidad a las interacciones, se utiliza el concepto de *experiment*, que contabiliza el número de veces que una interacción entre dos proteínas ha sido demostrada experimentalmente por un método de detección concreto y en una publicación concreta (es decir, un *experiment* es un evento que combina un método de detección y un artículo). En **Figura 64** se muestra una captura de pantalla de la aplicación de Mentha para la interacción entre las proteínas BRCA1_HUMAN y ESR1_HUMAN donde se puede observar un ejemplo de esto.



Gene: BRCA1
Organism: Homo sapiens
UniProt: P38398
Top 5 interactors

Gene: ESR1
Organism: Homo sapiens
UniProt: P03372
Top 5 interactors

Evidence: 30
Type: Physical
Score: 1

Common Gene Ontology terms
Biological Process: positive regulation of transcription from RNA polymerase II promoter - positive regulation of transcription, DNA-templated - regulation of apoptotic process - regulation of transcription, DNA-templated - response to estrogen - transcription, DNA-templated **Cellular Component**: cytoplasm - nucleoplasm - nucleus - plasma membrane **Molecular Function**: DNA binding - enzyme binding - metal ion binding - protein binding - zinc ion binding

Common pathways (KEGG)
Breast cancer

Common Tissues (UniProt)
Gland - Placenta - Carcinoma - Pituitary

Hide evidence

This **paper** **(Pubmed:11244506)** has reported **direct interaction** by using **pull down**. Source: MINT
This **paper** **(Pubmed:11244506)** has reported **physical association** by using **coimmunoprecipitation**. Source: MINT
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **anti bait coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **anti bait coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **anti tag coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **anti tag coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **pull down**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **direct interaction** by using **pull down**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **physical association** by using **anti bait coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:15674350)** has reported **physical association** by using **anti bait coimmunoprecipitation**. Source: IntAct
This **paper** **(Pubmed:11244506)** has reported **direct interaction** by using **pull down**. Source: BioGRID

Figura 64. *Extracto de los resultados del servidor web de Mentha para la interacción BRCA1-ESR1 donde puede comprobarse como el mismo método de detección de interacciones (pull down) descrito en la misma publicación (Pubmed:*

11244506) es contabilizado como dos evidencias diferentes solo por el hecho de estar registrado en dos bases de datos primarias (MINT y BioGRID). En APID Interactomes, esto correspondería a dos curations events y un único experimento o evidencia.

4. El visualizador de redes de Mentha está escrito en JAVA y, por tanto, depende de software adicional para su ejecución, estando así sujeto a las diferentes combinaciones de versiones de sistema operativo, navegador y complementos JAVA. En APID Interactomes se ha implementado un visualizador de redes basado en JavaScript que funciona directamente en el navegador sin necesidad de ningún otro software, reduciendo así los problemas de compatibilidad.
5. Mentha ofrece la posibilidad de trabajar con las interacciones de proteínas únicas o listados de estas, pero no ofrece interactomas completos. APID Interactomes, sin embargo, ofrece interactomas pre-calculados en tres diferentes niveles de calidad para todas las especies que contiene.

En cuanto al resto de bases de datos mencionadas inicialmente, cuya concepción es diferente a APID Interactomes, cabría destacar lo siguiente:

1. **IrefWeb:** Esta plataforma incluye unos desarrollos similares a APID Interactomes en algunos aspectos, pero su objetivo principal es diferente ya que trata de recopilar el máximo de información textual y ofrecer un potente sistema de filtrado, pero no ofrece ningún tipo de visualización de redes ni interactomas precalculados.
2. **HINT:** Aunque se trata de un proyecto de cobertura mucho menor, no comparable a otros de los que aquí se mencionan, cabe destacar que es el único que ofrece, como APID Interactomes, la posibilidad de obtener interactomas completos desde su aplicación web.
3. **STRING:** En este caso, se trata de una meta-base de datos con gran cantidad de información y una interfaz muy bien desarrollada capaz de presentar redes de modo interactivo; pero que incluye sobre todo predicción de interacciones por métodos computacionales y también mezcla las PPIs con datos de coexpresión génica derivados de estudios transcriptómicos, alejándose de esta manera del objetivo de APID Interactomes y otras bases de datos como Mentha o HINT.
4. **GeneMania:** Al igual que ocurre con STRING, se trata de una meta-base de datos con gran cantidad de información y con un visualizador de redes muy potente, pero que mezcla las PPIs con muchos datos de coexpresión génica derivados de estudios transcriptómicos y también con mera inferencia de asociación entre proteínas por estar en las mismas vías metabólicas o biológicas funcionales; de este modo tiene un objetivo muy diferente al de APID Interactomes.

5.2 Cobertura de interacciones respecto a servidores bioinformáticos similares

Las únicas bases de datos que siguen una aproximación similar a APID Interactomes, en cuanto a que integran datos experimentales de interacciones, y que estaban actualizadas en 2016, son iRefWeb **(Turner et al. 2010)** (17) y Mentha **(Calderone et al. 2013)** (15).

En **Tabla 11** puede verse una comparativa entre APID Interactomes, iRefWeb y Mentha donde se muestra el número total de interacciones contenidas en cada una de estas bases de datos para ocho de los más importantes organismos modelo. En cada caso se calcula la cobertura del interactoma correspondiente sobre el proteoma de dicho organismo según la base de datos Uniprot Proteomes y se calcula la media para los ocho organismos. Todos los datos de esta comparativa están recogidos en enero de 2016. Los datos de Mentha se obtuvieron de la sección de estadísticas disponible en su página web (15). En el caso de iRefWeb, los datos se obtuvieron utilizando su sistema de búsqueda avanzada en su versión 13 (17). Para que estos últimos fueran comparables a los datos disponibles en APID Interactomes se seleccionaron solo las interacciones experimentales físicas descritas en una o más publicaciones científicas.

El resultado de esta comparativa indica que la base de datos de APID Interactomes proporciona en todos los casos (excepto en *E. Coli* respecto a Mentha) una cobertura superior a las otras dos bases de datos siendo la cobertura media de los interactomas casi un 10% superior. Como ya se comentó anteriormente, en el caso del organismo *S. cerevisiae (strain ATCC 204508 / S288c)*, la cobertura en APID Interactomes excede el 100% porque su interactoma incluye interacciones con proteínas de TrEMBL no incluidas en el proteoma de Uniprot Proteomes.

Como comentario adicional cabe destacar que, durante todo este trabajo, se han considerado los proteomas contenidos en la base de datos Uniprot Proteomes, que incluyen tanto proteínas revisadas (*Reviewed*) como no revisadas (*Unreviewed*). Resulta evidente que, por ejemplo, para humano, existen actualmente unas 20000 proteínas bien caracterizadas, aunque Uniprot incluye registros distintos de otras 50000 proteínas humanas que corresponden a fragmentos o a secuencias de polipéptidos identificados de distintos modos. En esta Tesis Doctoral no se aborda la validez de dichos registros, sino que simplemente se han integrado en la base de datos de APID Interactomes para poder referirse a ellos cuando algún trabajo experimental publicado les cita.

***Tabla 11**. Comparativa de la cobertura de los interactomas disponibles en APID Interactomes, iRefWeb y Mentha para algunos de los organismos modelo más importantes.*

		APID Interactomes (enero 2106)			iRefWeb (v13.0)			Mentha (enero 2016)		
	Proteínas en el PROTEOMA	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el PROTEOMA (%)	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el PROTEOMA (%)	Proteínas en el INTERACTOMA	Interacciones en el INTERACTOMA	Cobertura sobre el PROTEOMA (%)
Homo sapiens	69 986	28 749	334 817	**41.08**	18 841	222 069	**26.92**	17 390	223 978	**24.85**
Mus musculus	50 189	17 478	53 676	**34.82**	9 118	30 117	**18.17**	8 777	24 878	**17.49**
Arabidopsis thaliana	31 477	10 394	40 018	**33.02**	7 879	21 454	**25.03**	8 346	27 722	**26.51**
Rattus norvegicus	29 885	5 556	8 797	**18.59**	3 330	8 286	**11.14**	2 388	4 057	**7.99**
Caenorhabditis elegans	26 596	5 569	14 762	**20.94**	5 506	14 102	**20.70**	5 502	13 965	**20.69**
Drosophila melanogaster	22 005	11 692	59 977	**53.13**	9 509	44 996	**43.21**	10 509	43 794	**47.76**
S. cerevisiae (cepa ATCC 204508 / S288c)	6 721	6 823	128 740	**101.52**	6 083	117 029	**90.51**	6 159	100 948	**91.64**
Escherichia coli (strain K12)	4 252	3 547	25 335	**83.42**	3 351	15 269	**78.81**	4 181	26 033	**98.33**
Cobertura media sobre el PROTEOMA (%)				**48.32**			**39.31**			**41.91**

Para tratar de explicar la mejora en la cobertura que el algoritmo de APID Interactomes consigue, podrían plantearse las siguientes posibilidades:

1. Los conjuntos de datos que se comparan están generados a partir de diferentes versiones de los datos primarios. Esta posibilidad, aunque viable, es bastante improbable puesto que todos los datos fueron recogidos en la misma fecha (enero 2016) y las tres bases de datos declaraban una actualización reciente en aquel momento.
2. APID Interactomes incluye fuentes de datos adicionales. Como se muestra en **Tabla 12**, APID Interactomes incluye dos bases de datos adicionales respecto a Mentha y una respecto a iRefWeb. Por contraste, respecto a APID Interactomes, Mentha incluye una base de datos adicional y iRefWeb un total de cinco adicionales.

Tabla 12. *Fuentes de datos primarias que procesa cada base de datos integrativa.*

	iRefWeb	**Mentha**	**APID Interactomes**
***IntAct* (Kerrien et al. 2012)**	Si	Si	Si
***DIP* (Salwinski et al. 2004)**	Si	Si	Si
***MINT* (Licata et al. 2012)**	Si	Si	Si
***BioGRID* (Chatr-Aryamontri et al. 2015)**	Si	Si	Si
***BIND* (Bader et al. 2003)**	Si	-	-
***CORUM* (Ruepp et al. 2010)**	Si	-	-
***MPact* (Güldener et al. 2006)**	Si	-	-
***MPPI* (Mewes et al. 2011)**	Si	-	-
***OPHID* (Brown and Jurisica 2005)**	Si	-	-
***MatrixDB* (Chautard et al. 2011)**	-	Si	-
***BioPlex* (Huttlin et al. 2015)**	-	-	Si
***HPRD* (Keshava Prasad et al. 2009)**	Si	-	Si

Se da la circunstancia de que las dos bases de datos adicionales presentes en APID Interactomes solo contienen datos de proteínas humanas. Además, la parte publicada de BioPlex, que contiene unas 24.000 interacciones **(Huttlin et al. 2015)**, está incluida en la base de datos BioGrid (que tanto Mentha como iRefWeb incluyen). Si bien es cierto que APID Interactomes va más allá e incluye un total de 56.000 a partir de los datos públicos que BioPlex ofrece en su web, esta diferencia no justifica por si sola el incremento general en la cobertura de APID Interactomes frente a las otras dos meta-bases de datos, especialmente en organismos distintos del humano.

3. Los ficheros procedentes de las bases de datos primarias que se procesan no son los mismos en los tres casos. Esto, que a priori parecería improbable, es quizás la causa

fundamental de las diferencias en cobertura entre los tres proyectos que se comparan. APID Interactomes procesa ficheros crudos en formato PSI-MI XML en todos los casos excepto en el de BioPlex ya que esta solo proporciona un fichero tabulado. Estos ficheros XML requieren un procesamiento técnicamente complejo, pero contienen toda la información que las bases de datos alojan. Es habitual, entre las bases de datos que procesan interacciones, trabajar con ficheros tabulados de tipo PSI-MI MITAB, que son más sencillos de procesar, pero pueden no contener toda la información disponible en la base de datos. En el caso de Mentha, además, sus autores indican explícitamente en la publicación que sus datos provienen de un proceso automatizado que utiliza el protocolo de PSICQUIC **(Aranda et al. 2011; del-Toro et al. 2013)**, lo cual limita su cobertura de forma significativa al excluir cualquier interacción que no esté contenida en el consorcio IMEx **(Orchard et al. 2012)**.

Teniendo en cuenta lo expuesto en estos tres puntos, cabe atribuir la diferencia de cobertura al origen técnico de los datos y a su posterior procesamiento. Lo que resulta especialmente interesante es que el procesamiento que el algoritmo de APID Interactomes hace de estos datos crudos está orientado de forma clara a asegurar su calidad y evitar ambigüedades, lo que lógicamente supone una reducción en el número de interacciones que finalmente resultan aceptadas y, por tanto, una pérdida de cobertura final.

Valga un ejemplo real de procesamiento de datos en APID Interactomes para ilustrar de forma cuantitativa esta pérdida de interacciones. En **Figura 65** puede verse un resumen esquemático de parte del procesamiento de las interacciones contenidas en las bases de datos primarias para el organismo *Mus musculus*: partiendo de 120086 registros de interacciones entre las proteínas de dicho organismo que han sido descritas en publicaciones científicas, obtenemos finalmente 53676 interacciones únicas para almacenar en APID Interactomes. Este conjunto de interacciones conforma el interactoma que APID Interactomes publica y que, respecto al proteoma de UniProt, ofrece una cobertura del 34.82% tal y como se indica en **Tabla 11**.

Durante el procesamiento de dichos registros procedentes de las bases de datos primarias se descartan aproximadamente un 3% de interacciones porque alguno de sus interactores no es una proteína y un 32% de las restantes debido a procesos como la asignación y actualización de identificadores o la unificación de isoformas. Así, cuando finalmente se agrupan los registros por pares de proteínas únicos, se obtienen 53676 interacciones frente a las 79389 que se obtendrían si no se procesaran los datos.

Cabe destacar, por tanto, que aunque la pérdida de datos que el algoritmo de APID Interactomes considera ambiguos o no identificables es significativa, dicho proceso genera interactomas con una cobertura superior a la obtenida por los algoritmos de Mentha o iRefWeb. Esto se cumple en todos los organismos estudiados excepto en las interacciones de *E. Coli* en Mentha. En este último caso, se ha buscado una causa que explique esa diferencia, pero no hay información publicada al respecto. También se ha comprobado que no hay error en el procesamiento realizado en este trabajo, quedando como última posibilidad que Mentha incluya alguna fuente primaria de datos no declarada.

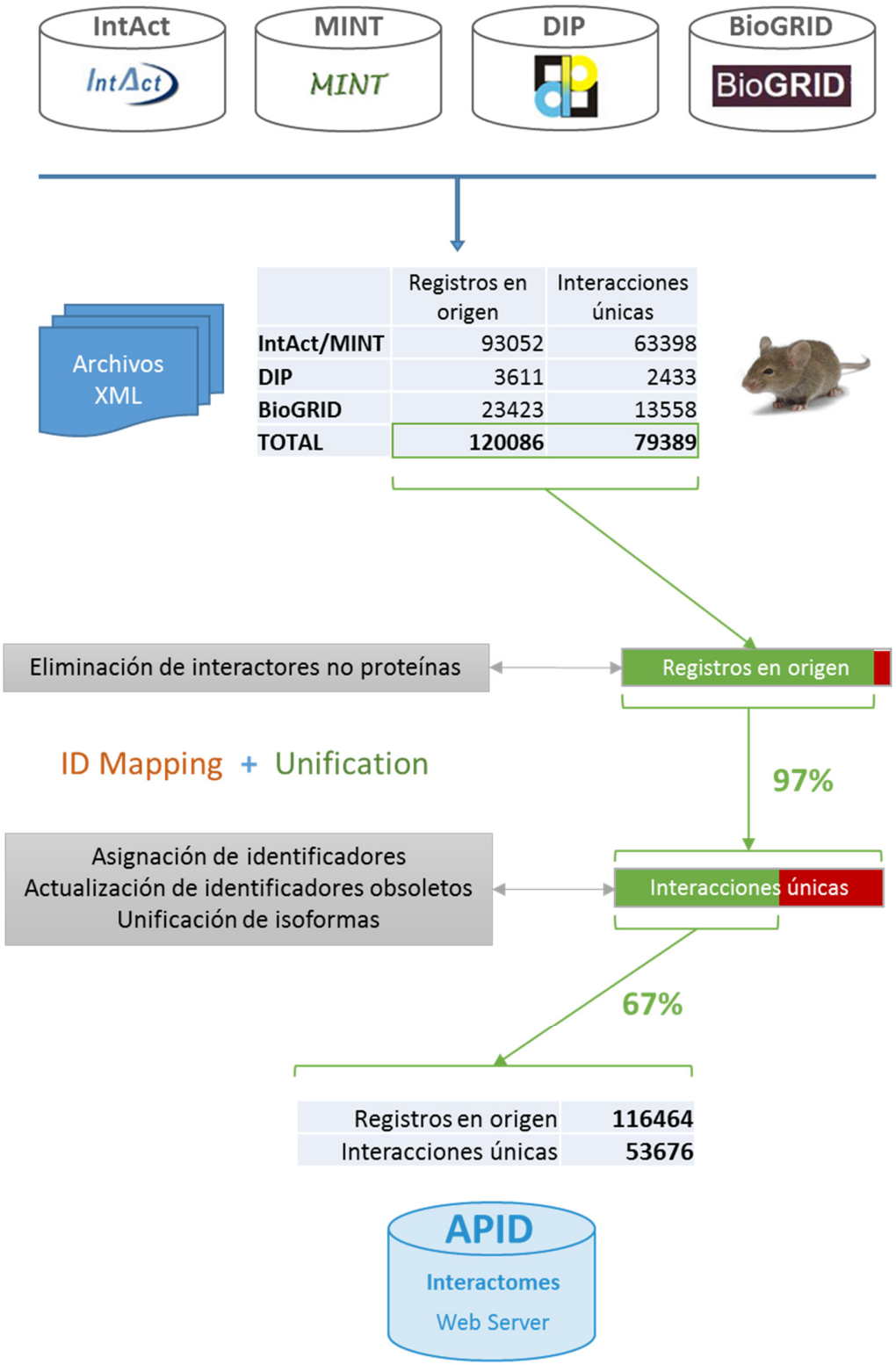


Figura 65. *Proceso esquemático de filtrado de registros iniciales en las bases de datos primarias para el interactoma del organismo Mus musculus.*

5.3 Análisis comparativo de bases de datos primarias: caso interactoma humano

Con el objetivo de caracterizar cuantitativamente y comparar el aporte de interacciones desde cada base de datos primaria a APID Interactomes, se utilizó el interactoma humano, sin interacciones interespecie, como caso de estudio.

Así, se extrajeron de la base de datos integrada en APID Interactomes todos los conjuntos de interacciones entre proteínas humanas que se habían importado originalmente de cada una de las bases de datos consideradas primarias (es decir, de BioGRID, BioPlex, DIP, HPRD, IntAct y PDB) y se combinaron para encontrar tanto los solapamientos entre ellas como aquellas interacciones exclusivas de cada base de datos. Con el resultado de dicho análisis se construyó el diagrama de Venn que se muestra en **Figura 66**. Al tratarse de seis conjuntos distintas de datos la visualización es compleja, pero la figura construida muestra varios resultados relevantes: **(i)** BioGRID es la fuente que reporta más PPIs de humano propias, no presentes en otras bases de datos (107010 PPIs), **(ii)** las siguientes fuentes que aportan más PPIs de modo exclusivo son BioPlex (36024 PPIs) e IntAct (34122 PPIs), **(iii)** la mayor intersección se da entre BioGRID e IntAct (56181 PPIs) y **(iv)** la intersección de todas sólo se da para 79 PPIs.

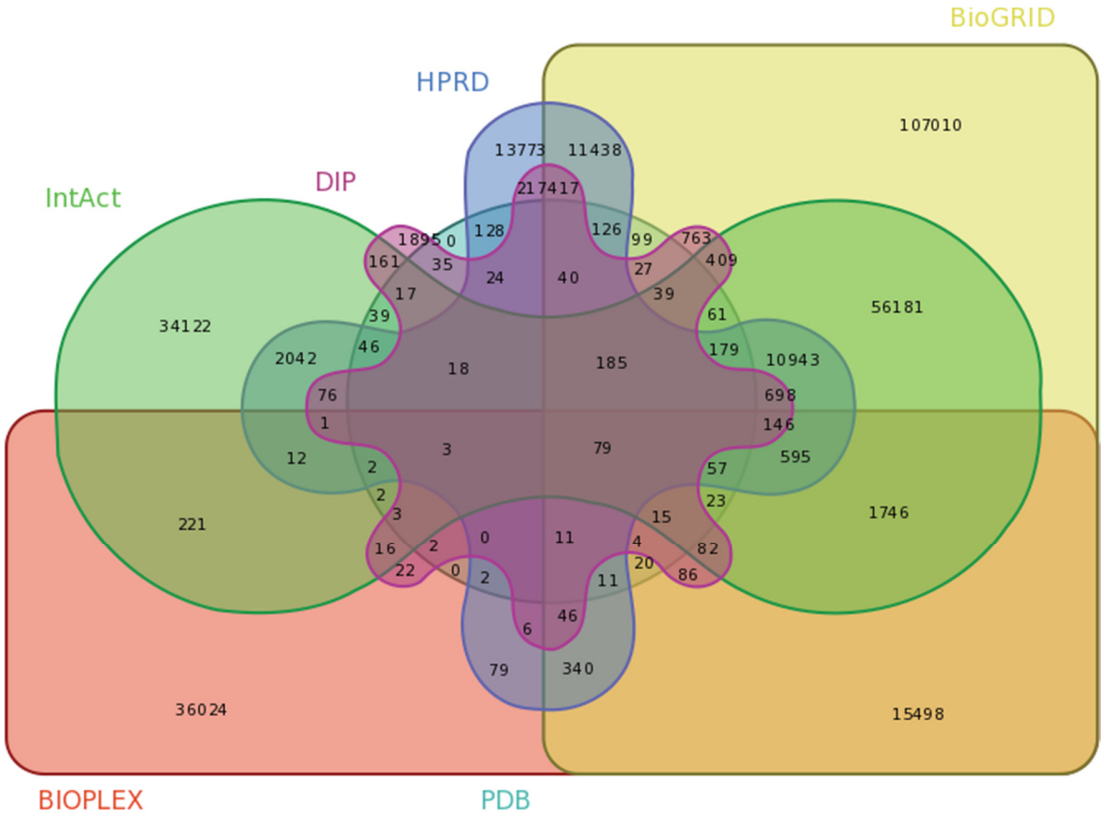


Figura 66. *Solapamiento de información procedente de las diferentes bases de datos primarias para las interacciones contenidas en el interactoma humano, no interespecie y con el uso de cualquier método de detección.*

Este mismo análisis se hizo para el subconjunto de interacciones asociado al interactoma humano *binario*, que incluye sólo interacciones proteína-proteína que han sido demostradas por métodos experimentales de tipo binario (como se ha descrito en el apartado Cálculo del interactoma binario de *Homo Sapiens*). De este modo, los interactomas binarios derivados de cada base de datos fueron comparados obteniendo el diagrama de Venn que puede verse en **Figura 67**.

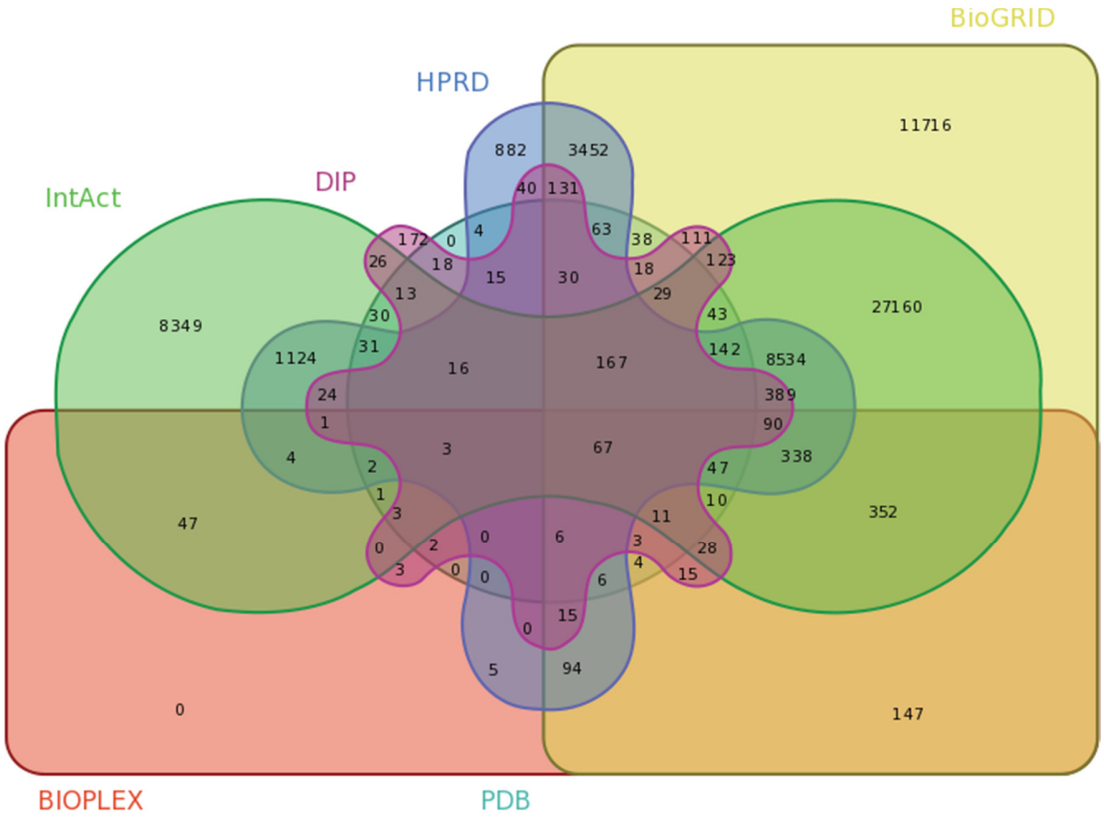


***Figura 67**. Solapamiento de información procedente de las diferentes bases de datos primarias para las interacciones contenidas en el interactoma humano, no interespecie y solo mediante el uso de método de detección binarios (Interactoma humano binario).*

En esta comparativa puede comprobarse como BioPlex no aporta ninguna PPI al interactoma binario debido a que todas sus interacciones están detectadas con un método no binario (MI:0004, *affinity chromatography technology*). Si ocurre que varias de las interacciones descritas en BioPlex están descritas también en otras bases de datos mediante el uso de métodos binarios, como puede verse en los diferentes solapamientos. Otra observación relevante de este análisis es que todas las bases de datos ven reducidos sus números de modo drástico, siendo este efecto más llamativo en el caso de BioGRID que ahora pasa a aportar tan sólo 11716 PPIs binarias respecto a las 107010 PPIs que se incluyeron en el interactoma global.

También puede verse que, tanto en el interactoma general como en el binario, la base de datos PDB no aporta ninguna interacción por sí misma. Esto es lo esperado ya que las estructuras tridimensionales extraídas de PDB solo son asignadas a las interacciones que el algoritmo de APID

Interactomes previamente ha obtenido de alguna de las otras cinco bases de datos (que son las realmente consideradas como fuentes *primarias*). Así, PDB es utilizada como una fuente de validación y de aporte de información estructural sobre las interfaces moleculares de interacción. En todo caso, es relevante incluirla y mostrarla en estos diagramas con la intención de que puedan observarse los solapamientos del resto de fuentes primarias de PPIs con la base de datos de estructuras 3D de proteínas.

En **Figura 68** y **Figura 69** pueden observarse los diagramas de Venn correspondientes a los interactomas humanos, general y binario, teniendo en cuenta solo las bases de datos primarias, es decir, sin PDB. BioGRID e IntAct siguen siendo las que más aportan y las que más solapan entre sí.

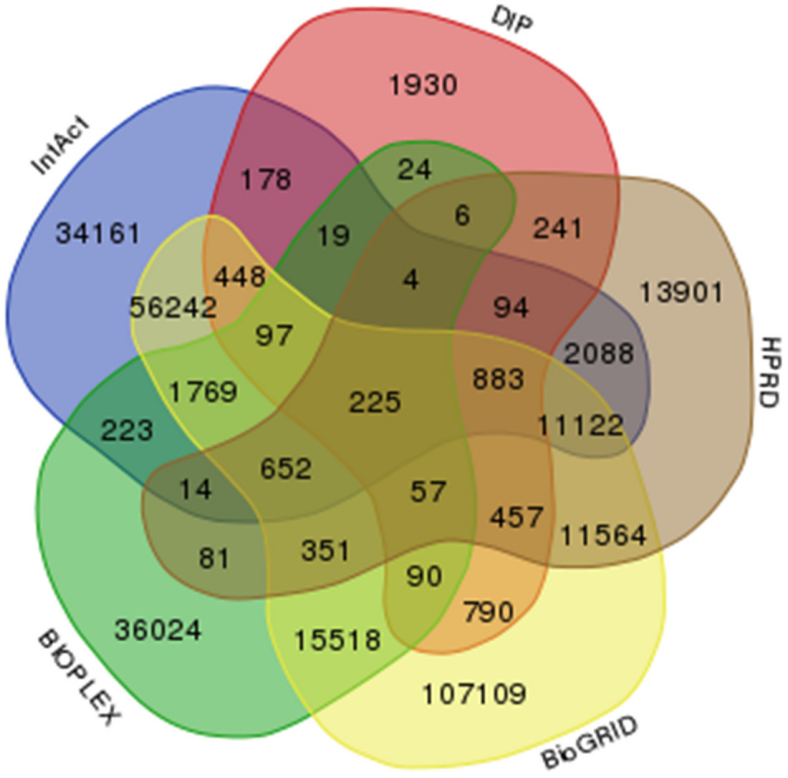


Figura 68. *Solapamiento de información procedente de las diferentes bases de datos primarias, excepto PDB, para las interacciones contenidas en el interactoma humano, no interespecie y con el uso de cualquier método de detección.*

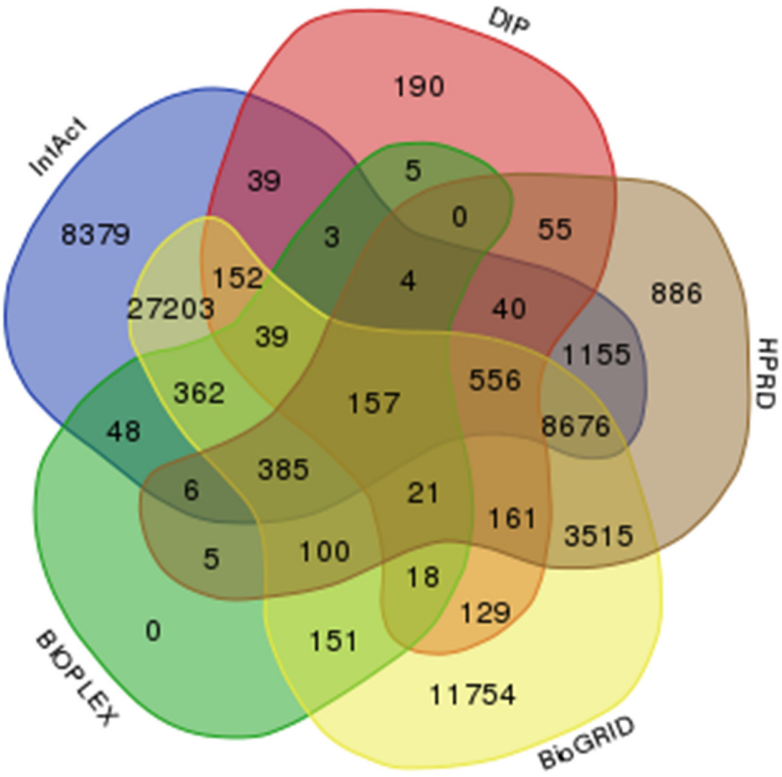


Figura 69. *Solapamiento de información procedente de las diferentes bases de datos primarias, excepto PDB, para las interacciones contenidas en el interactoma humano, no interespecie y solo mediante el uso de método de detección binarios (Interactoma humano binario).*

Por último, en **Tabla 13** se desglosan, para el caso de las proteínas humanas, las diferentes aportaciones de cada base de datos al interactoma general global y al interactoma binario. En dicha tabla se muestra el número de interacciones específicas que cada una de las bases de datos comparada aporta de forma única y, a partir de este dato, se calculan los porcentajes respecto al total de las interacciones aportadas por la misma base de datos. Este cálculo se hace primero teniendo en cuenta todos los métodos de detección de interacciones y luego teniendo en cuenta sólo los métodos binarios. Por último, se calcula la proporción de métodos de detección binarios en cada base de datos y en el total de interacciones almacenadas en APID Interactomes. De este modo se puede ver que el número de interacciones binarias que incluye APID es un 21.66 % del total, lo que supone un conjunto de 64194 PPIs. Este tamaño para el interactoma binario humano es mucho más concorde con HuRI ***(Rual et al. 2005; Rolland et al. 2014)*** (72), que pretende proporcionar un interactoma humano de referencia validado experimentalmente por métodos binarios.

Tabla 13. *Número y porcentaje de los diferentes aportes de cada base de datos primaria a los interactomas humanos general y binario generados en APID Interactomes.*

Interactoma humano (PPIs no interespecie)								
	Todos los métodos			Métodos binarios				
	Totales	Específicas	% de PPIs exclusivas respecto al total de PPIs en su BBDD	% de PPIs exclusivas respecto al total de PPIs en APID	Totales	Específicas	% de PPIs exclusivas respecto al total de PPIs en su BBDD	% de PPIs exclusivas respecto al total de PPIs en APID	% de PPIs binarias
Intact	**108219**	34122	31.53%	11.51%	**47204**	8349	17.69%	13.01%	**43.62%**
DIP	**5543**	1895	34.19%	0.64%	**1569**	172	10.96%	0.27%	**28.31%**
HPRD	**41740**	13773	33.00%	4.65%	**15722**	882	5.61%	1.37%	**37.67%**
BioGRID	**207374**	107010	51.60%	36.11%	**53379**	11716	21.95%	18.25%	**25.74%**
BioPlex	**55154**	36024	65.32%	12.16%	**1304**	0	0.00%	0.00%	**2.36%**
PDB	**1297**	-	-	-	**822**	-	-	-	**63.38%**
Interacciones únicas totales	**296362**	192824		65.06%	**64194**	21119		32.90%	**21.66%**

5.4 Uso de métricas transparentes frente a *scores* integrados para las interacciones

El uso de sistemas de puntuación o *scores* está muy extendido en estudios de redes para medir o cuantificar la asociación o interacción entre nodos. Esto se debe a que proporciona un criterio rápido para filtrar y ordenar resultados según un único valor de peso o puntuación calculado a partir de diversos parámetros que miden la validez, grado de confianza o fuerza de cada relación.

En el ámbito del estudio de las interacciones entre proteínas, estos sistemas de puntuación se utilizan para lograr una modulación en el grado de interacción, asignando normalmente un determinado nivel de fiabilidad o certeza a cada interacción. Como ejemplos de dicho uso pueden citarse los *scores* calculados por Mentha **(Calderone et al. 2013)** o GeneMania **(Zuberi et al. 2013)**, que establecen fórmulas matemáticas para asignar un valor a cada interacción en función de las evidencias registradas. El problema del uso de *scores* es que los parámetros que deben considerarse en determinados contextos experimentales no siempre pueden expresarse de forma integrativa y transparente con un único valor. Es decir, a pesar de que dicha simplificación tenga ventajas a nivel práctico, se corre el riesgo de que esta no represente la realidad de forma correcta.

En el caso de las interacciones proteína-proteína, cuando el *score* quiere reflejar el grado de confianza o certeza que se tiene, no es fácil definir el grado de influencia que debe tener cada parámetro en el resultado final, ya que los criterios pueden ser variables en función del contexto experimental. Así, resulta prácticamente imposible diseñar una fórmula o modelo que integre todos los parámetros de fiabilidad de una interacción y se adapte a los diferentes escenarios donde un investigador trata de decidir si una interacción concreta es más o menos fiable que otra.

***Tabla 14**. Parámetros calculados en APID Interactomes para las interacciones entre las proteínas de Homo Sapiens HRAS-SOS1 y HRAS-RAF1.*

	Experiments	**Methods**	**Publications**	**3D Structures**	**Curation Events**
HRAS - SOS1	7	5	7	18	12
HRAS - RAF1	36	8	30	3	40

Por ejemplo, si se trata de comparar la interacción entre las proteínas de *Homo Sapiens* HRAS y SOS1 frente a la interacción entre HRAS y RAF1, basándose en los parámetros presentados en **Tabla 14**, resulta complicado decidir qué interacción es más fiable. Por ejemplo, en las dos interacciones comparadas, existe una diferencia clara entre el número de experimentos que demuestran la primera frente a la segunda: 7 *experiments* para HRAS-SOS1 frente a 36 *experiments* para HRAS-RAF1. Esto daría mucha más validez y confianza a la interacción HRAS-RAF1. Sin embargo, en el caso de HRAS-SOS1, existen publicadas una gran cantidad de estructuras 3D donde se han co-cristalizado ambas proteínas y se conoce la interfaz de contactos fisicoquímicos a nivel molecular entre ellas (18

estructuras 3D). Sin embargo, sólo existen 3 estructuras 3D para la pareja de proteínas RASH-RAF1. ¿Qué parámetro confirma mejor que esa interacción existe? ¿Cuál es más importante? La respuesta a dichas preguntas no es clara porque depende del contexto experimental e, incluso, del criterio personal del investigador. Es por esto que en APID Interactomes se ha decidido no establecer una fórmula única para asignar fiabilidad a las interacciones, sino que de modo directo se proporcionan los números de todo el conjunto de parámetros que permiten valorar dicha fiabilidad desde diferentes perspectivas.

Así, en función del contexto, puede darse más o menos fiabilidad a una interacción concreta basándose en los parámetros objetivos que se tienen para cada interacción:

1. ***Experiments***: El número de experimentos distintos que han sido reportados que demuestran una interacción concreta entre dos proteínas (definiendo como "experimento" cada vez que un método de determinación experimental concreto se aplica a una pareja de proteínas y se publica en un artículo distinto).
2. ***Methods***: El número de métodos experimentales distintos utilizados para demostrar una interacción concreta entre dos proteínas.
3. ***Publications***: El número de publicaciones, bien referenciadas, obtenidas de la literatura científica en las que una interacción concreta ha sido reportada.
4. ***3D structures***: El número de estructuras 3D, derivado de la base de datos PDB, en las cuales un par de proteínas han sido co-cristalizadas y tienen una interfaz de interacción, con contactos moleculares bien definidos, que es distinta de la simple inclusión de ambas en la caja del cristal (*crystal lattice*).
5. ***Curation Events***: El número de veces que una interacción entre dos proteínas derivada de un artículo científico ha sido registrada en distintas bases de datos de PPIs primarias (es decir, el número de veces que un equipo de expertos o *curators* ha registrado dicha interacción, de modo independiente o coordinado, en bases de datos distintas).

Todos estos parámetros bien definidos proporcionan una serie de métricas que permiten al investigador establecer filtros para priorizar unas interacciones sobre otras, pero sin perder la transparencia a través de un cálculo de *score* concreto único que, inevitablemente, implica ciertas decisiones que pueden no ser las más adecuadas para el contexto del estudio específico o del análisis que se quiera realizar.

5.5 Número de experimentos como métrica de confianza frente a *curation events*

En APID Interactomes, la métrica principal de una interacción es el valor denominado *experiments*, que como se ha indicado representa el número de demostraciones diferentes por método y publicación que se han reportado para una interacción dada. Esta métrica expresa el grado de estudio experimental y demostración que se ha dado hasta la fecha de una interacción independientemente de su grado de registro en bases de datos. Es decir, el número de *curation events* no tiene por qué coincidir con el número de *experiments*. Esto es esperable ya que un mismo

experimento puede estar registrado en más de una base de datos de las que se usan como fuente en un estudio. Está claro que el número de experimentos está ligado al "grado de conocimiento y de estudio" que se tiene de una proteína concreta, que es reflejado en el número de publicaciones y en las veces que distintos equipos científicos han probado sus interacciones; pero el conocimiento no aumenta por el hecho de que diversas bases de datos concurrentes hayan registrado y reportado el mismo experimento. Por ello, se deben contar de modo distinto los *experiments* ("experimentos") y los *curations events* ("registros") sin mezclarlos. Esto es lo que se hace en APID Interactomes, donde cada interacción proteína-proteína se reporta de modo independiente con X número de métodos experimentales e Y número de registros distintos en bases de datos. Sin embargo, esto es algo que no se hace en otras bases de datos integrativas donde la métrica principal es el número de *curation events*. El ejemplo más claro, tal y como se explica en el primer apartado de esta discusión, aparece en Mentha **(Calderone et al. 2013)** (15), donde se denominan evidencias a los *curation events* y se utilizan para calcular un *score* de fiabilidad para cada interacción.

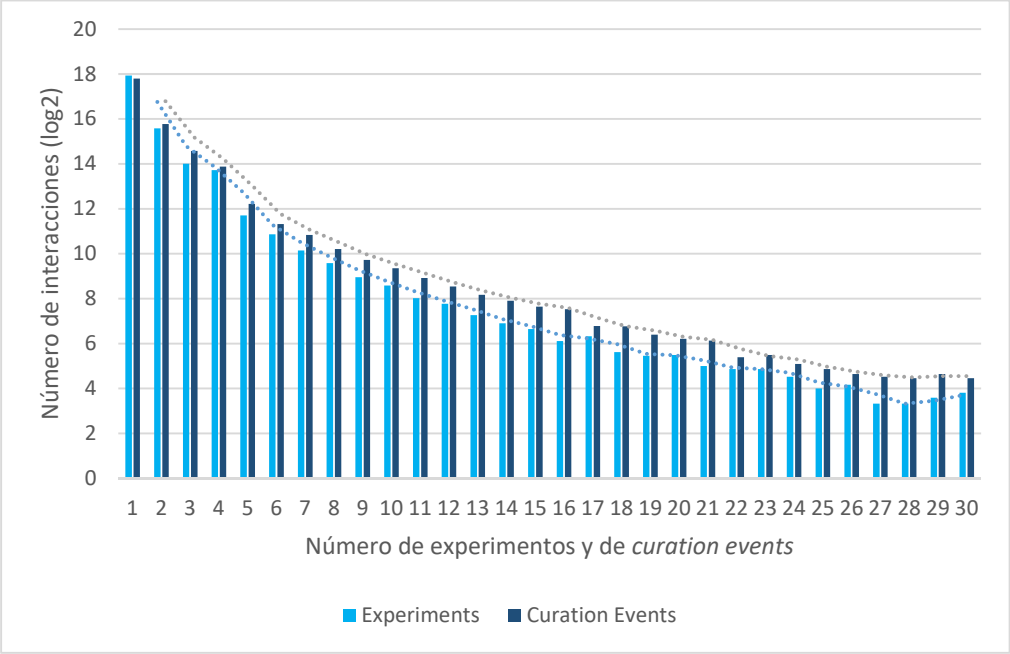


Figura 70. *Representación gráfica del número de interacciones existentes en APID Interactomes para diferentes valores de experimentos y registros en bases de datos primarias (curation events). El número interacciones para cada nivel de filtrado está expresado en escala logarítmica (log2) para que pueda visualizarse mejor.*

El uso de los registros en bases de datos primarias como métrica principal es problemático porque se sobreestima la fiabilidad de las interacciones. En **Figura 70** puede verse cómo el uso del número de *curation events* como criterio de filtrado proporciona un número de interacciones más alto en cada nivel frente al uso del número de experimentos tal y como se definen en APID Interactomes. Esto sucede en todos los niveles de fiabilidad excepto para el valor 1 puesto que cuando una interacción presenta un único registro entre todas las bases de datos primarias, este no puede estar repetido.

Las diferencias calculadas para cada nivel de fiabilidad entre el uso de uno u otro criterio de filtrado pueden verse en **Figura 71**. De nuevo, excepto para el valor 1, la sobreestimación es clara cuando se usa como criterio el número de *curation events*. De hecho, para ciertos niveles de fiabilidad, se obtienen más del doble de interacciones, dándose un incremento artificial entre un ≈10 % y hasta un ≈170 % cuando se cuentan *curation events*. De este modo, el tamaño de los interactomas que se obtienen si se filtra a un cierto nivel de fiabilidad se ve afectado de forma significativa por el criterio descrito.

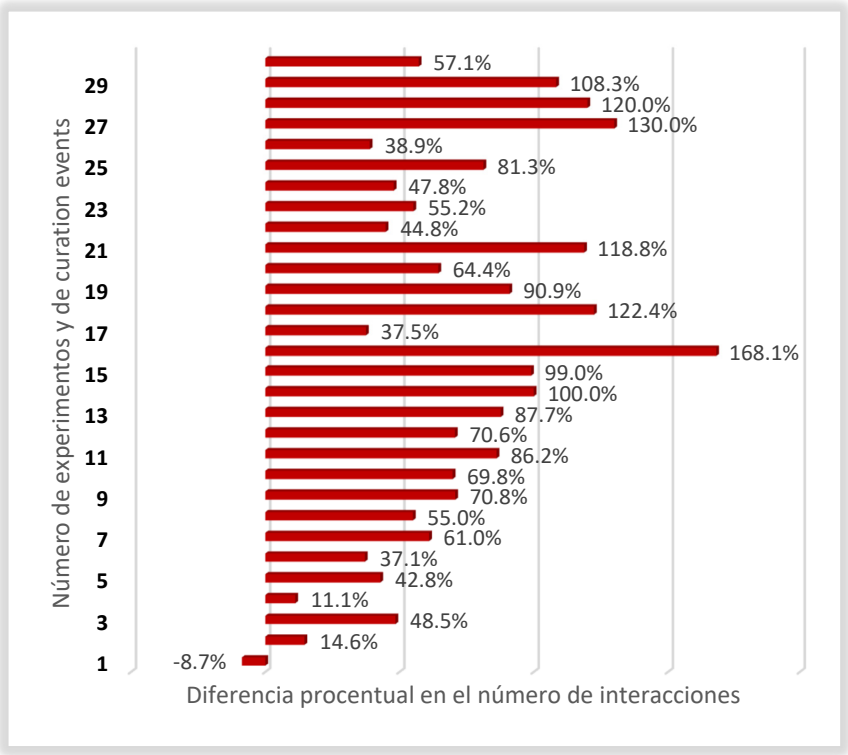


***Figura 71**. Diferencia porcentual en el número de interacciones existentes en APID Interactomes filtrando por un mismo valor de experimentos y de registros en base de datos primaria (curation events).*

En conclusión, el uso de los registros en bases de datos primarias como evidencias de una interacción provoca una sobre-estimación en el tamaño de los interactomas y, en algunos casos, un sesgo en los cálculos de *scores* que pretender cuantificar la fiabilidad de las diferentes interacciones.

6 CONCLUSIONES

A continuación, se enuncian las principales conclusiones de esta Tesis Doctoral, asociadas a los resultados obtenidos, aunque buscando trascender lo meramente fenomenológico y proponer perspectivas para la investigación futura en el campo de estudio.

1. Se ha construido una plataforma bioinformática de acceso libre que integra y ordena miles de datos referenciados de interacción molecular física entre proteínas, organizados por proteomas, y permite hacer consultas específicas con listados de proteínas. La plataforma, denominada APID Interactomes, también permite la construcción de redes de interacción molecular de proteínas que pueden ser visualizadas, analizadas y exportadas para su estudio con otros programas.

2. La identificación de interacciones moleculares binarias entre proteínas demostradas experimentalmente permite construir redes complejas, que establecen mapas de relación donde se pueden reconocer módulos como grupos de proteínas asociadas. Estos módulos reflejan la arquitectura biomolecular de las células, y corresponden en primer lugar a los complejos de proteínas como máquinas moleculares esenciales para los sistemas biológicos. Si las interacciones no tienen una calidad y fiabilidad alta, las redes y mapas de relación derivados no reflejan bien la biología.

3. El mapeo experimental de todas las interacciones binarias entre las proteínas de un organismo es condición necesaria para lograr conocer su interactoma completo. Este trabajo aporta una colección bastante amplia de interactomas de proteínas que, para humano y para un conjunto amplio de organismos modelo (desde el ratón a la levadura), tiene la mejor cobertura contrastada con bases de datos y repositorios actuales.

4. El estudio de las redes de interacción de proteínas se puede realizar por técnicas de teoría de grafos y métodos computacionales derivados. Las redes de proteínas que se han construido en este trabajo son complejas ya que incluyen normalmente miles de nodos y decenas de miles de relaciones. No obstante, implementando herramientas de búsqueda, filtrado y contextualización, su análisis permite revelar proteínas centrales, módulos interconectados, relaciones con funciones biológicas u otro tipo de características topológicas o funcionales, todos ellos aspectos relevantes para la biología del sistema en estudio.

5. La integración y unificación de datos de alta dimensión procedentes de distintas bases de datos biológicas primarias, realizada en este estudio, no es trivial y nos ha llevado a descubrir incoherencias, ambigüedades y errores en las fuentes de datos y en el uso de los formatos establecidos. La bioinformática debe avanzar ante estos retos y en este trabajo hemos propuesto una serie de soluciones que han resultado prácticas y eficaces para la obtención de un compendio amplio de interactomas de proteínas que incluye controles de calidad y guarda los estándares reclamados por la *HUPO Proteomics Standards Initiative*.

7 REFERENCIAS BIBLIOGRÁFICAS

Alanis-Lobato, G, Andrade-Navarro, MA and Schaefer MH. 2017. **HIPPIE v2.0: Enhancing Meaningfulness and Reliability of Protein-Protein Interaction Networks.** *Nucleic Acids Research* 45(D1): D408-D414. doi:10.1093/nar/gkw985.

Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. 2016. **APID Interactomes: Providing Proteome-Based Interactomes with Controlled Quality for Multiple Species and Derived Networks**. *Nucleic Acids Research* 44 (W1): W529-35. doi:10.1093/nar/gkw363.

Aloy P, Russell RB. 2006. **Structural Systems Biology: Modelling Protein Interactions.** *Nature Reviews Molecular Cell Biology* 7 (3): 188–97. doi:10.1038/nrm1859.

Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, Dana JM et al. 2011. **PSICQUIC and PSISCORE: Accessing and Scoring Molecular Interactions.** *Nature Methods* 8 (7): 528–29. doi:10.1038/nmeth.1637.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, et al. 2000. **Gene Ontology: Tool for the Unification of Biology**. *Nature Genetics* 25 (1): 25–29. doi:10.1038/75556.

Bader GD, Betel D, Hogue CWV. 2003. **BIND: The Biomolecular Interaction Network Database.** *Nucleic Acids Research* 31 (1): 248–50. doi:10.1093/NAR/GKG056.

Bader GD, Hogue CWV. 2003. **An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks.** *BMC Bioinformatics* 4 (1): 2. doi:10.1186/1471-2105-4-2.

Barabási AL, Réka A. 1999. **Emergence of Scaling in Random Networks**. *Science* 286 (5439): 509-12. doi:10.1126/science.286.5439.509.

Barabasi AL, Oltvai ZN. 2004. **Network Biology: Understanding the Cell's Functional Organization**. *Nature Reviews Genetics* 5 (2): 101–13. doi:10.1038/nrg1272.

Ben-Hur A, Noble WS. 2005. **Kernel Methods for Predicting Protein-Protein Interactions**. *Bioinformatics* 21 Suppl. 1: i38-46. doi:10.1093/bioinformatics/bti1016.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. **GenBank**. *Nucleic Acids Research* 41 (Database issue): D36-42. doi:10.1093/nar/gks1195.

Berggård T, Linse S, James P. 2007. **Methods for the Detection and Analysis of Protein–protein Interactions**. *Proteomics* 7 (16): 2833–42. doi:10.1002/pmic.200700131.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. **The Protein Data Bank.** *Nucleic Acids Research* 28 (1): 235–42. doi:10.1093/NAR/28.1.235.

Bernhardt A, Lechner E, Hano P, Schade V, Dieterle M, Anders M, Dubin MJ et al. 2006. **CUL4 Associates with DDB1 and DET1 and Its Downregulation Affects Diverse Aspects of Development in Arabidopsis Thaliana**. *Plant Journal* 47 (4): 591–603. doi:10.1111/j.1365-313X.2006.02810.x.

Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov D et al. 2015. **Gene Ontology Consortium: Going Forward**. *Nucleic Acids Research* 43 (Database issue): D1049-56. doi:10.1093/nar/gku1179.

Boriack-Sjodin PA, Margarit SM, Bar-Sagi D, Kuriyan J. 1998. **The Structural Basis of the Activation of Ras by Sos.** *Nature* 394 (6691): 337–43. doi:10.1038/28548.

Brown KR, Jurisica I. 2005. **Online Predicted Human Interaction Database.** *Bioinformatics* 21 (9): 2076–82. doi:10.1093/bioinformatics/bti273.

Calderone A, Castagnoli L, Cesareni G. 2013. **Mentha: A Resource for Browsing Integrated Protein-Interaction Networks.** *Nature Methods* 10 (8): 690. doi:10.1038/nmeth.2561.

Carbon S, Ireland A, Mungall CJ, Shu SQ, Marshall B, Lewis S, Lomax J et al. 2009. **AmiGO: Online Access to Ontology and Annotation Data**. *Bioinformatics* 25 (2): 288–89. doi:10.1093/bioinformatics/btn615.

Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. 2007. **GENECODIS: A Web-Based Tool for Finding Significant Concurrent Annotations in Gene Lists.** *Genome Biology* 8 (1): R3. doi:10.1186/gb-2007-8-1-r3.

Ceol A, Chatr-Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. 2009. **MINT, the Molecular Interaction Database: 2009 Update**. *Nucleic Acids Research* 38 (SUPPL.1): D532-9. doi:10.1093/nar/gkp983.

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C et al. 2015. **The BioGRID Interaction Database: 2015 Update**. *Nucleic Acids Research* 43 (Database issue): D470–8. doi:10.1093/nar/gku1204.

Chatr-aryamontri A, Ceol A, Licata L, Cesareni G. 2008. **Protein Interactions: Integration Leads to Belief**. *Trends in Biochemical Sciences* 33 (6): 241–42. doi:10.1016/j.tibs.2008.04.002.

Chautard E, Fatoux-Ardore M, Ballut L, Thierry-Mieg N, Ricard-Blum S. 2011. **MatrixDB, the Extracellular Matrix Interaction Database.** *Nucleic Acids Research* 39 (Database issue): D235-40. doi:10.1093/nar/gkq830.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R et al. 2007. **Integration of Biological Networks and Gene Expression Data Using Cytoscape**. *Nature Protocols* 2 (10): 2366–82. doi:10.1038/nprot.2007.324.

Côté R, Jones P, Martens L, Apweiler R, Hermjakob H. 2008. **The Ontology Lookup Service: More Data and Better Tools for Controlled Vocabulary Queries.** *Nucleic Acids Research* 36 (Web Server issue): 97. doi:10.1093/nar/gkn252.

Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H. 2010. **The Ontology Lookup Service: Bigger and Better.** *Nucleic Acids Research* 38 (Web Server issue): W155-60. doi:10.1093/nar/gkq331.

Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J. 2012. **PINA v2.0: Mining Interactome Modules**. *Nucleic Acids Research* 40 (Database issue): D862–65. doi:10.1093/nar/gkr967.

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M et al. 2014. **The Reactome Pathway Knowledgebase**. *Nucleic Acids Research* 42 (Database issue): D472-7. doi:10.1093/nar/gkt1102.

Dandekar T, Snel B, Huynen M, Bork P. 1998. **Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact.** *Trends in Biochemical Sciences* 23 (9): 324–28. http://www.ncbi.nlm.nih.gov/pubmed/9787636.

Das J, Yu H. 2012. **HINT: High-Quality Protein Interactomes and Their Applications in Understanding Human Disease.** *BMC Systems Biology* 6 (1): 92. doi:10.1186/1752-0509-6-92.

De Beer TA, Berka K, Thornton JM, Laskowski RA. 2014. **PDBsum Additions**. *Nucleic Acids Research* 42 (Database issue): D292–96. doi:10.1093/nar/gkt940.

Del-Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J et al. 2013. **A New Reference Implementation of the PSICQUIC Web Service.** *Nucleic Acids Research* 41 (Web Server issue): 601–6. doi:10.1093/nar/gkt392.

Dogrusoz U, Giral E, Cetintas A, Civril A, Demir E. 2009. **A Layout Algorithm for Undirected Compound Graphs**. *Information Sciences* 179 (7): 980–94. doi:10.1016/j.ins.2008.11.017.

Doncheva NT, Assenov Y, Domingues FS, Albrecht M. 2012. **Topological Analysis and Interactive Visualization of Biological Networks and Protein Structures.** *Nature Protocols* 7 (4): 670–85. doi:10.1038/nprot.2012.004.

Dong J, Horvath S. 2007. **Understanding Network Concepts in Modules.** *BMC Systems Biology* 1 (1): 24. doi:10.1186/1752-0509-1-24.

Dumousseau M, Koch M, Shrivastava A, Sullivan J, Yehudi Y, Micklem G, Alonso-López D, De Las Rivas J, Orchard S. 2017. **JAMI: A Java Library for Molecular Interactions**. *(en preparación)*.

Dumousseau M, Koch M, Shrivastava A, Sullivan J, Yehudi Y, Micklem G, Alonso-López D, De Las Rivas J, Orchard S et al. 2017. **Encompassing New Use Cases - Level 3.0 of the HUPO-PSI Format for Molecular Interactions**. *(en preparación)*.

ENCODE Project Consortium. 2012. **An Integrated Encyclopedia of DNA Elements in the Human Genome.** *Nature* 489 (7414): 57–74. doi:10.1038/nature11247.

Enright AJ, Iliopoulos I, Kyrpides NC, and Ouzounis CA. 1999. **Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events.** *Nature* 402 (6757): 86–90. doi:10.1038/47056.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC et al. 2015. **The Pfam Protein Families Database: Towards a More Sustainable Future.** *Nucleic Acids Research* 44 (D1): D279-85. doi:10.1093/nar/gkv1344.

Fontanillo C, Nogales-Cadenas R, Pascual-Montano A, De Las Rivas J. 2011. **Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms**. *PLoS ONE* 6 (9): e24289. doi:10.1371/journal.pone.0024289.

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. 2015. **Cytoscape.js: A Graph Theory Library for Visualisation and Analysis**. *Bioinformatics* 32 (2): 309–11. doi:10.1093/bioinformatics/btv557.

Fruchterman TMJ, Reingold EM. 1991. **Graph Drawing by Force-Directed Placement**. *Software: Practice and Experience* 21 (11): 1129–64. doi:10.1002/spe.4380211102.

Griffiths AJF, Gelbart WM, Lewontin RC, Miller JH. 1999. **Modern Genetic Analysis**. WH Freeman. New York. ISBN-10: 0-7167-3118-5

Guelzim N, Bottani S, Bourgine P, Kepes F. 2002. **Topological and Causal Structure of the Yeast Transcriptional Regulatory Network.** *Nature Genetics* 31 (1): 60–63. doi:10.1038/ng873.

Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V. 2006. **MPact: The MIPS Protein Interaction Resource on Yeast.** *Nucleic Acids Research* 34 (Database issue): D436-41. doi:10.1093/nar/gkj003.

Hakes L, Robertson DL, Oliver SG, Lovell SC. 2006. **Protein Interactions from Complexes: A Structural Perspective**. *Comparative and Functional Genomics* 2007:49356. doi:10.1155/2007/49356

Hermjakob H, Montecchi-Palazzi L, Bader GD, Wojcik J, Salwinski L, Ceol A, Moore S et al. 2004. **The HUPO PSI's Molecular Interaction Format--a Community Standard for the Representation of Protein Interaction Data.** *Nature Biotechnology* 22 (2): 177–83. doi:10.1038/nbt926.

Human Genome Sequencing Consortium, International. 2004. **Finishing the Euchromatic Sequence of the Human Genome**. *Nature* 431 (7011): 931–45. doi:10.1038/nature03001.

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. 2015. **The GOA Database: Gene Ontology Annotation Updates for 2015**. *Nucleic Acids Research* 43 (Database issue): D1057–63. doi:10.1093/nar/gku1113.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S et al. 2015. **The BioPlex Network: A Systematic Exploration of the Human Interactome**. *Cell* 162 (2): 425–40. doi:10.1016/j.cell.2015.06.043.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. **A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome.** *Proceedings of the National Academy of Sciences of the USA* 98 (8): 4569–74. doi:10.1073/pnas.061034498.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. 2003. **A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data.** *Science* 302 (5644): 449–53. doi:10.1126/science.1087361.

Jeong H, Mason SP, Barabási AL, Oltvai ZN. 2001. **Lethality and Centrality in Protein Networks.** *Nature* 411 (6833): 41–42. doi:10.1038/35075138.

Jeong H, Tombor B, Albert R, Oltvai ZA, Barabasi AL. 2000. **The Large-Scale Organization of Metabolic Networks**. *Nature* 407 (6804): 651–54. doi:10.1038/35036627.

Kalathur RKR, Pinto JP, Hernández-Prieto MA, MacHado RSR, Almeida D, Chaurasia G, Futschik ME. 2014. **UniHI 7: An Enhanced Database for Retrieval and Interactive Analysis of Human Molecular Interaction Networks**. *Nucleic Acids Research* 42 (Database issue): D408–14. doi:10.1093/nar/gkt1100.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M et al. 2012. **The IntAct Molecular Interaction Database in 2012**. *Nucleic Acids Research* 40 (Database issue): D841–6. doi:10.1093/nar/gkr1088.

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD et al. 2007. **Broadening the Horizon--Level 2.5 of the HUPO-PSI Format for Molecular Interactions.** *BMC Biology* 5: 44. doi:10.1186/1741-7007-5-44.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D et al. 2009. **Human Protein Reference Database - 2009 Update**. *Nucleic Acids Research* 37:767–72. doi:10.1093/nar/gkn892.

Kodama Y, Shumway M, Leinonen R. 2012. **The Sequence Read Archive: Explosive Growth of Sequencing Data**. *Nucleic Acids Research* 40 (Database issue): D54–56. doi:10.1093/nar/gkr854.

Kostyuchenko VA, Leiman PG, Chipman PR, Kanamaru S, van Raaij MJ, Arisaka F, Mesyanzhinov VV, Rossmann MG. 2003. **Three-Dimensional Structure of Bacteriophage T4 Baseplate.** *Nature Structural Biology* 10 (9): 688–93. doi:10.1038/nsb970.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S et al. 2016. **Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.** *Nucleic Acids Research* 44 (W1): W90-7. doi:10.1093/nar/gkw377.

Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA. 2003. **Myriads of Protein Families, and Still Counting.** *Genome Biology* 4 (2): 401. doi:10.1186/gb-2003-4-2-401.

De Las Rivas J, Fontanillo C. 2010. **Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks.** *PLoS Computational Biology* 6 (6): e1000807. doi:10.1371/journal.pcbi.1000807.

Leinonen R, Sugawara H, Shumway M. 2011. **The Sequence Read Archive**. *Nucleic Acids Research* 39(Database issue): D19-21. doi:10.1093/nar/gkq1019.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F et al. 2012. **MINT, the Molecular Interaction Database: 2012 Update**. *Nucleic Acids Research* 40 (Database issue): D857-61. doi:10.1093/nar/gkr930.

López Y, Nakai K, Patil A. 2015. **HitPredict Version 4: Comprehensive Reliability Scoring of Physical Protein-Protein Interactions from More than 100 Species**. *Database* 2015 (1): bav117. doi:10.1093/database/bav117.

Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM. 2007. **Protein Interactions: Is Seeing Believing?** *Trends in Biochemical Sciences* 32 (12): 530–31. doi:10.1016/j.tibs.2007.09.006.

Mani R, St Onge RP, Hartman JL, Giaever J, Roth FP. 2008. **Defining Genetic Interaction**. *Proceedings of the National Academy of Sciences of the USA* 105 (9): 3461–66. doi:10.1073/pnas.0712255105.

Meldal B, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS et al. 2015. **The Complex Portal - An Encyclopaedia of Macromolecular Complexes**. *Nucleic Acids Research* 43 (D1): D479–84. doi:10.1093/nar/gku975.

Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. 2003. **STRING: A Database of Predicted Functional Associations between Proteins**. *Nucleic Acids Research* 31(1): 258-61. doi:10.1093/nar/gkg034.

Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K et al. 2011. **MIPS: Curated Databases and Comprehensive Secondary Data Resources in 2010.** *Nucleic Acids Research* 39 (Database issue): D220-4. doi:10.1093/nar/gkq1157.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C et al. 2015. **The InterPro Protein Families Database: The Classification Resource after 15 Years**. *Nucleic Acids Research* 43 (Database issue): D213–21. doi:10.1093/nar/gku1243.

Mosca R, Céol A, Aloy P. 2012. **Interactome3D: Adding Structural Details to Protein Networks**. *Nature Methods* 10 (1): 47–53. doi:10.1038/nmeth.2289.

Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A. 2009. **GeneCodis: Interpreting Gene Lists through Enrichment Analysis and Integration of Diverse Biological Information**. *Nucleic Acids Research* 37 (Web Server issue): W317-22. doi:10.1093/nar/gkp416.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 27 (1): 29-34. doi:10.1093/nar/27.1.29.

Omar AM, Mahran MA, Ghatge MS, Bamane FHA, Ahmed MH, El-Araby ME, Abdulmalik O, Safo MK. 2016. **Aryloxyalkanoic Acids as Non-Covalent Modifiers of the Allosteric Properties of Hemoglobin**. *Molecules* 21 (8): 1057. doi:10.3390/molecules21081057.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH et al. 2014. **The MIntAct Project - IntAct as a Common Curation Platform for 11 Molecular Interaction Databases**. *Nucleic Acids Research* 42 (Database issue): D358–63. doi:10.1093/nar/gkt1115.

Orchard S, Kerrien S, Abbani S, Aranda N, Bhate J, Bidwell S, Bridge A et al. 2012. **Protein Interaction Data Curation: The International Molecular Exchange (IMEx) Consortium.** *Nature Methods* 9 (4): 345–50. doi:10.1038/nmeth.1931.

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A et al. 2007. **The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx).** *Nature Biotechnology* 25 (8): 894–98. doi:10.1038/nbt1324.

Osumi-Sutherland D. 2010. **OBO Format**. *Ontogenesis*, January. http://ontogenesis.knowledgeblog.org/245.

Patil A, Nakai K, Nakamura H. 2011. **HitPredict: A Database of Quality Assessed Protein-Protein Interactions in Nine Species**. *Nucleic Acids Research* 39 (Database issue): D744-9. doi:10.1093/nar/gkq897.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. **Correlated Mutations Contain Information about Protein-Protein Interaction.** *Journal of Molecular Biology* 271 (4): 511–23. doi:10.1006/jmbi.1997.1198.

Pazos F, Valencia A. 2001. **Similarity of Phylogenetic Trees as Indicator of Protein-Protein Interaction.** *Protein Engineering* 14 (9): 609–14.

Prieto C, De Las Rivas J. 2006. **APID: Agile Protein Interaction DataAnalyzer**. *Nucleic Acids Research* 34 (Web Server issue): W298-302. doi:10.1093/nar/gkl128.

Razick S, Magklaras G, Donaldson IM. 2008. **iRefIndex: A Consolidated Protein Interaction Database with Provenance.** *BMC Bioinformatics* 9: 405. doi:10.1186/1471-2105-9-405.

Rolland T, Taşan M, Charloteaux M, Pevzner SJJ, Zhong Q, Sahni N, Yi S et al. 2014. **A Proteome-Scale Map of the Human Interactome Network**. *Cell* 159 (5): 1212–26. doi:10.1016/j.cell.2014.10.050.

Rose PW, Prli A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK et al. 2015. **The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education**. *Nucleic Acids Research* 43 (Database issue): D345–56. doi:10.1093/nar/gku1214.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF et al. 2005. **Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network**. *Nature* 437 (7062): 1173–78. doi:10.1038/nature04209.

Ruepp A, Waegele B, Lechner M, Brauner M, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2010. **CORUM: The Comprehensive Resource of Mammalian Protein Complexes--2009.** *Nucleic Acids Research* 38 (Database issue): D497-501. doi:10.1093/nar/gkp914.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. **The Database of Interacting Proteins: 2004 Update.** *Nucleic Acids Research* 32 (Database issue): D449-51. doi:10.1093/nar/gkh086.

Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. 2012. **Hippie: Integrating Protein Interaction Networks with Experiment Based Quality Scores**. Edited by Charlotte M. Deane. *PLoS ONE* 7 (2): e31826. doi:10.1371/journal.pone.0031826.

Schwartz B, Zaitsev P, Tkachenko V. 2012. **High Performance MySQL**. *O'Reilly*. ISBN: 978-1-4493-1428-6

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Research* 13 (11): 2498–2504. doi:10.1101/gr.1239303.

Shneiderman B. 1996. **The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations**. In *Visual Languages*, 1996. Proceedings, IEEE Symposium on, pp. 336-343. doi:10.1109/VL.1996.545307.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ et al. 2007. **The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration.** *Nature Biotechnology* 25 (11): 1251–55. doi:10.1038/nbt1346.

Sugiyama K. 2002. **Graph Drawing and Applications for Software and Knowledge Engineers**. doi:10.1142/4902.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M et al. 2015. **STRING v10: Protein-Protein Interaction Networks, Integrated over the Tree of Life**. *Nucleic Acids Research* 43 (Database issue): D447–52. doi:10.1093/nar/gku1003.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A et al. 2016. **The STRING Database in 2017: Quality-Controlled Protein–protein Association Networks, Made Broadly Accessible**. *Nucleic Acids Research* 45(Database issue): D362-D368. doi:10.1093/nar/gkw937.

Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. 2012. **GeneCodis3: A Non-Redundant and Modular Enrichment Analysis Tool for Functional Genomics**. *Nucleic Acids Research* 40 (Web Server issue): W478-83. doi:10.1093/nar/gks402.

The UniProt Consortium. 2014. **UniProt: A Hub for Protein Information**. *Nucleic Acids Research* 43 (Database issue): D204-12. doi:10.1093/nar/gku989.

Tong A. 2004. **Global Mapping of the Yeast Genetic Interaction Network**. *Science* 303 (5659): 808–13. doi:10.1126/science.1091317.

Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ. 2010. **iRefWeb: Interactive Analysis of Consolidated Protein Interaction Data and Their Supporting Evidence.** *Database* 2010: baq023. doi:10.1093/database/baq023.

Valencia A, Pazos F. 2008. **Computational Methods to Predict Protein Interaction Partners**. In: Protein-protein Interactions and Networks, pp. 67-81. Springer London. doi:10.1007/978-1-84800-125-1_4.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO et al. 2001. **The Sequence of the Human Genome**. *Science* 291 (5507): 1304-51. doi:10.1126/science.1058040.

Vidal M, Cusick ME, Barabási AL. 2011. **Interactome Networks and Human Disease**. *Cell* 144 (6): 986–98. doi:10.1016/j.cell.2011.02.016.

Volkman BF, Alam SL, Satterlee JD, Markley JL. 1998. **Solution Structure and Backbone Dynamics of Component IV Glycera Dibranchiata Monomeric Hemoglobin-CO**. *Biochemistry* 37 (31): 10906–19. doi:10.1021/bi980810b.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, et al. 2010. **The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function.** *Nucleic Acids Research* 38 (Web Server issue): W214-20. doi:10.1093/nar/gkq537.

Ware C. 2004. **Information Visualization: Perception for Design**. *Morgan Kaufman (Elsevier).* ISBN: 978-0123814647. doi:10.1016/B978-0-12-381464-7.00018-1.

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. 2011. **Towards the Prediction of Protein Interaction Partners Using Physical Docking.** *Molecular Systems Biology* 7 (1): 469. doi:10.1038/msb.2011.3.

Watts DJ, Strogatz SH. 1999. **Collective Dynamics of 'Small-World' Networks.** *Nature* 393 (6684): 440–42. doi:10.1038/30918.

Wetlaufer DB. 1973. **Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins**. *Proceedings of the National Academy of Sciences of the USA* 70 (3):697–701. doi:10.1073/pnas.70.3.697.

Wiemann S, Pennacchio C, Hu Y, Hunter P, Harbers M, Amiet A, Bethel G et al. 2016. **The ORFeome Collaboration: A Genome-Scale Human ORF-Clone Resource**. *Nature Methods* 13 (3): 191–92. doi:10.1038/nmeth.3776.

Wodak SJ, Pu S, Vlasblom J, Séraphin B. 2009. **Challenges and Rewards of Interaction Proteomics.** *Molecular & Cellular Proteomics* 8 (1): 3–18. doi:10.1074/mcp.R800014-MCP200.

Yoshida M, Muneyuki E, Hisabori T. 2001. **ATP Synthase — a Marvellous Rotary Engine of the Cell**. *Nature Reviews Molecular Cell Biology* 2 (9): 669–77. doi:10.1038/35089509.

Young KH. 1998. **Yeast Two-Hybrid: So Many Interactions, (in) so Little Time...** *Biology of Reproduction* 58 (2): 302–11. doi:10.1095/BIOLREPROD58.2.302.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T et al. 2008. **High-Quality Binary Protein Interaction Map of the Yeast Interactome Network.** *Science* 322 (5898): 104–10. doi:10.1126/science.1158684.

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B et al. 2012. **Structure-Based Prediction of Protein–protein Interactions on a Genome-Wide Scale**. *Nature* 490 (7421): 556–60. doi:10.1038/nature11503.

Zhu X, Gerstein M, Snyder M. 2007. **Getting Connected: Analysis and Principles of Biological Networks**. *Genes and Development* 21 (9): 1010-24. doi:10.1101/gad.1528707.

Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q. 2013. **GeneMANIA Prediction Server 2013 Update.** *Nucleic Acids Research* 41 (Web Server issue): W115–22. doi:10.1093/nar/gkt533.

8 LISTA DE DIRECCIONES WEB

1. **HUPO Proteomics Standards Initiative.** [En línea] http://www.psidev.info/.

2. **Human Proteome Organization.** [En línea] https://www.hupo.org/.

3. **The International Molecular Exchange Consortium (IMEx).** [En línea] http://www.imexconsortium.org/.

4. **Especificación PSI-MI MITAB 2.7.** [En línea] https://code.google.com/archive/p/psimi/wikis/PsimiTab27Format.wiki.

5. **Especificación PSI-MI XML.** [En línea] http://www.psidev.info/mif.

6. **Ontology Lookup Service (OLS).** [En línea] http://www.ebi.ac.uk/ols/index.

7. **UniProt (Universal Protein Resource).** [En línea] http://www.uniprot.org/.

8. **Reducing proteome redundancy.** [En línea] http://www.uniprot.org/help/proteome_redundancy.

9. **IntAct Molecular Interaction Database.** [En línea] http://www.ebi.ac.uk/intact/.

10. **MINT, the Molecular INTeraction database.** [En línea] http://mint.bio.uniroma2.it/.

11. **Human Protein Reference Database.** [En línea] http://www.hprd.org/.

12. **Database of Interacting Proteins.** [En línea] http://dip.doe-mbi.ucla.edu/dip/.

13. **Biological General Repository for Interaction Datasets.** [En línea] https://thebiogrid.org/.

14. **BioPlex (biophysical interactions of ORFeome-based complexes).** [En línea] http://wren.hms.harvard.edu/bioplex/.

15. **Mentha.** [En línea] http://mentha.uniroma2.it/.

16. **MatrixDB: The Extracellular Matrix Interaction Database.** [En línea] http://matrixdb.univ-lyon1.fr/.

17. **iRefWeb.** [En línea] http://wodaklab.org/iRefWeb/.

18. **CORUM: The comprehensive resource of mammalian protein complex.** [En línea] http://mips.helmholtz-muenchen.de/corum/.

19. **The MIPS Mammalian Protein-Protein Interaction Database.** [En línea] http://mips.helmholtz-muenchen.de/proj/ppi/.

20. **Interologous interaction database.** [En línea] http://ophid.utoronto.ca/ophidv2.204/.

21. **HINT.** [En línea] http://hint.yulab.org/.

22. **STRING.** [En línea] http://string-db.org/.

23. **GeneMania.** [En línea] http://genemania.org/.

24. **Hippie.** [En línea] http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/.

25. **UniHI.** [En línea] http://www.unihi.org/.

26. **PINA2.** [En línea] http://cbg.garvan.unsw.edu.au/pina.

27. **HitPredict.** [En línea] http://hintdb.hgc.jp/htp.

28. **RCSB Protein data bank.** [En línea] http://www.rcsb.org/pdb/home/home.do.

29. **PDBsum: Pictorial database of 3D structures in the Protein Data Bank.** [En línea] http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html.

30. **Gene ontology consortium.** [En línea] http://www.geneontology.org/.

31. **AmiGO 2.** [En línea] http://amigo.geneontology.org/amigo.

32. **KEGG: Kyoto Encyclopedia of Genes and Genomes.** [En línea] http://www.genome.jp/kegg/.

33. **Reactome: a curated pathway database.** [En línea] http://www.reactome.org/.

34. **Pfam database.** [En línea] http://pfam.xfam.org/.

35. **InterPro: protein sequence analysis & classification.** [En línea] https://www.ebi.ac.uk/interpro/.

36. **Cytoscape: Network Data Integration, Analysis, and Visualization in a Box.** [En línea] http://www.cytoscape.org/.

37. **IntAct FTP.** [En línea] ftp://ftp.ebi.ac.uk/pub/databases/intact/current.

38. **Licencia Apache 2.0.** [En línea] http://www.apache.org/licenses/LICENSE-2.0.

39. **Punto de acceso a la descarga de archivos en HPRD.** [En línea] http://www.hprd.org/download.

40. **Punto de acceso a la descarga de archivos en BioGRID.** [En línea] http://thebiogrid.org/download.php.

41. **Punto de acceso a la descarga de archivos en BioPlex.** [En línea] http://wren.hms.harvard.edu/bioplex/downloadInteractions.php.

42. **Esquema XSD del formato XML usado en la base de datos UniProt.** [En línea] http://www.uniprot.org/docs/uniprot.xsd.

43. **Librería Simple API for XML (SAX).** [En línea] http://www.saxproject.org/.

44. **NCBI Taxonomy FTP.** [En línea] ftp://ftp.ncbi.nih.gov/pub/taxonomy/.

45. **OBO Foundry.** [En línea] http://www.obofoundry.org/.

46. **UniProt Proteomes.** [En línea] http://www.uniprot.org/proteomes/.

47. **Servicios REST de UniProt.** [En línea] http://www.uniprot.org/help/programmatic_access.

48. **Apache Log4j 2.** [En línea] http://logging.apache.org/log4j/2.x/.

49. **Librerías JDBC.** [En línea] http://docs.oracle.com/javase/tutorial/jdbc/.

50. **Framework JAMI.** [En línea] https://github.com/MICommunity/psi-jami.

51. **ORM Hibernate.** [En línea] http://hibernate.org/.

52. **ORM Doctrine.** [En línea] http://www.doctrine-project.org/.

53. **Sistema gestor de bases de datos MySQL.** [En línea] https://www.mysql.com/.

54. **Sistema gestor de bases de datos de grafos Neo4j.** [En línea] https://neo4j.com/.

55. **Estándares HTML y CSS (W3C).** [En línea] https://www.w3.org/standards/webdesign/htmlcss.

56. **HTML Kickstart .** [En línea] http://www.99lime.com/elements/.

57. **Librería jQuery.** [En línea] https://jquery.com/.

58. **Librería qTip2.** [En línea] http://qtip2.com/.

59. **Paquete de iconos vectoriales Font Awesome.** [En línea] http://fontawesome.io/.

60. **Librería Charts de Google Developers.** [En línea] https://developers.google.com/chart/.

61. **Librería DataTables.** [En línea] https://datatables.net/.

62. **Librería cytoscape.js.** [En línea] http://js.cytoscape.org/.

63. **Licencia MIT.** [En línea] https://opensource.org/licenses/MIT.

64. **Librería arbor.js.** [En línea] http://arborjs.org/.

65. **Extensión arbor para cytoscape.js.** [En línea] https://github.com/cytoscape/cytoscape.js-arbor.

66. **GeneTerm Linker.** [En línea] http://gtlinker.cnb.csic.es/.

67. **GeneCodis.** [En línea] http://genecodis.cnb.csic.es/.

68. **Enrichr.** [En línea] http://amp.pharm.mssm.edu/Enrichr/.

69. **Complex Portal.** [En línea] https://www.ebi.ac.uk/intact/complex/.

70. **UniProt REST .** [En línea] http://www.uniprot.org/help/programmatic_access.

71. **UniProt ID Mapping.** [En línea] http://www.uniprot.org/mapping/.

72. **HuRI: The Human Reference Protein Interactome Mapping Project.** [En línea] http://interactome.baderlab.org/.

73. **Interactome3D.** [En línea] http://interactome3d.irbbarcelona.org/.

9 LISTA DE FIGURAS

10 LISTA DE TABLAS

APÉNDICE: PUBLICACIONES CIENTÍFICAS

Parte del trabajo de investigación descrito en esta Tesis Doctoral ha sido publicado en el siguiente artículo científico (cuyo contenido completo se adjunta al final de este documento):

Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. (2016). **APID Interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks**. *Nucleic Acids Research*, 44(W1): W529-35. https://doi.org/10.1093/nar/gkw363

Además, fruto de las colaboraciones llevadas a cabo durante este trabajo de investigación con el grupo HUPO PSI-MI, están en proceso de publicación los siguientes trabajos:

Dumousseau M, Koch M, Shrivastava A, Sullivan J, Yehudi Y, Micklem G, **Diego Alonso-López**, Javier De Las Rivas, Orchard S (2017). **JAMI: A Java Library for Molecular Interactions** (en preparación)

Dumousseau M et al. (2017). **Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions** (en preparación)

También, durante este periodo doctoral se han realizado otros trabajos de bioinformática, no descritos en este documento, llevados a cabo en estrecha colaboración con distintos grupos de investigación experimentales que se han publicado en los siguientes artículos científicos:

Vicente-Dueñas C, Romero-Camarero I, González-Herrero I, Alonso-Escudero E, Abollo-Jiménez F, Jiang X, Gutierrez NC, Orfao A, Marín N, Villar LM, Criado MC, Pintado B, Flores T, **Alonso-López D**, De Las Rivas J, Jiménez R, Criado FJ, Cenador MB, Lossos IS, Cobaleda C, Sánchez-García I. (2012). **A novel molecular mechanism involved in multiple myeloma development revealed by targeting MafB to haematopoietic progenitors**. *The EMBO Journal*, *31*(18), 3704–3717. https://doi.org/10.1038/emboj.2012.227

Castellanos-Martín A, Castillo-Lluva S, Sáez-Freire Mdel M, Blanco-Gómez A, Hontecillas-Prieto L, Patino-Alonso C, Galindo-Villardon P, Pérez Del Villar L, Martín-Seisdedos C, Isidoro-Garcia M, Abad-Hernández MM, Cruz-Hernández JJ, Rodríguez-Sánchez CA, González-Sarmiento R, **Alonso-López D**, De Las Rivas J, García-Cenador B, García-Criado J, Lee DY, Bowen B, Reindl W, Northen T, Mao JH, Pérez-Losada J. (2015). **Unravelling heterogeneous susceptibility and the evolution of breast cancer using a systems biology approach**. *Genome Biology*, *16*(1), 40. https://doi.org/10.1186/s13059-015-0599-z

Martín-Lorenzo A, Hauer J, Vicente-Dueñas C, Auer F, González-Herrero I, García-Ramírez I, Ginzel S, Thiele R, Constantinescu SN, Bartenhagen C, Dugas M, Gombert M, Schäfer D, Blanco O, Mayado A, Orfao A, **Alonso-López D**, Rivas J de L, Cobaleda C, García-Cenador MB, García-Criado FJ, Sánchez-García I, Borkhardt A. (2015). **Infection Exposure is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility**. *Cancer Discovery*, *5*(12), 1328–1343. https://doi.org/10.1158/2159-8290.CD-15-0892

Ordóñez JL, Amaral AT, Carcaboso AM, Herrero D, García-Macías MC, Sevillano V, **Alonso-López D**, Pascual-Pasto G, San-Segundo L, Vila-Ubach M, Rodrigues T, Fraile S, Teodosio C, Mayo-Iscar A, Aracil M, Galmarini CM, Tirado OM, Mora J, De Álava E. (2015). **The PARP inhibitor olaparib enhances the sensitivity of Ewing sarcoma to trabectedin**. *Oncotarget*, 6(22), 18875–18890. https://doi.org/10.18632/oncotarget.4303

Cañueto J, Cardeñoso-Álvarez E, García-Hernández JL, Galindo-Villardón P, Vicente-Galindo P, Vicente-Villardón JL, **Alonso-López D**, De Las Rivas J, Valero J, Moyano-Sáez E, Fernández-López E, Mao JH, Castellanos-Martín A, Román-Curto C, Pérez-Losada J. (2016). **miR-203 and miR-205 expression patterns identify subgroups of prognosis in cutaneous squamous cell carcinoma**. *British Journal of Dermatology. https://doi.org/10.1111/bjd.15236*

García-Ramírez I, Tadros S, González-Herrero I, Martín-Lorenzo A, Rodríguez-Hernández G, Duval R, Moore D, Ruiz-Roca L, Blanco O, **Alonso-López D**, Hartert K, De Las Rivas J, Klinkibiel D, Bast M, Greiner T, Vose J, Lunning M, Rodrigues-Lima F, Jiménez R, García-Criado F, García-Cenador M, Brindle P, Vicente-Dueñas C, Alizadeh A, Sánchez-García I, Green M. (2017). **Crebbp loss cooperates with Bcl2 over-expression to promote lymphoma in mice**. *Blood.* https://doi.org/10.1182/blood-2016-08-733469

Published online 30 April 2016 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* ***W529–W535***
doi: 10.1093/nar/gkw363

APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks

Diego Alonso-López, Miguel A. Gutiérrez, Katia P. Lopes, Carlos Prieto, Rodrigo Santamaría and Javier De Las Rivas*

Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC) and Universidad de Salamanca (USAL), 37007 Salamanca, Spain

Received February 13, 2016; Revised April 15, 2016; Accepted April 23, 2016

ABSTRACT

APID (Agile Protein Interactomes DataServer) is an interactive web server that provides unified generation and delivery of protein interactomes mapped to their respective proteomes. This resource is a new, fully redesigned server that includes a comprehensive collection of protein interactomes for more than 400 organisms (25 of which include more than 500 interactions) produced by the integration of only experimentally validated protein–protein physical interactions. For each protein–protein interaction (PPI) the server includes currently reported information about its experimental validation to allow selection and filtering at different quality levels. As a whole, it provides easy access to the interactomes from specific species and includes a global uniform compendium of 90,379 distinct proteins and 678,441 singular interactions. APID integrates and unifies PPIs from major primary databases of molecular interactions, from other specific repositories and also from experimentally resolved 3D structures of protein complexes where more than two proteins were identified. For this purpose, a collection of 8,388 structures were analyzed to identify specific PPIs. APID also includes a new graph tool (based on Cytoscape.js) for visualization and interactive analyses of PPI networks. The server does not require registration and it is freely available for use at http://apid.dep.usal.es.

INTRODUCTION

Identification of all the specific connections between the elements that comprise a cellular system is crucial to unraveling its molecular architecture and mechanics. In this context, physical molecular interactions between protein pairs (called protein–protein interactions, PPIs) constitute an essential part of the cellular architecture in all living organisms. Genome-wide technologies have provided, over the last two decades, a compendium of the biomolecular entities that configurate many living systems, i.e., all the genes encoded in the genomes of specific organisms and the corresponding derived proteome. Once all these elements became known, the need for comprehensive maps of the molecular physical interactions that occur between such elements was evident, and systematic proteome-scale mapping of specific interactomes began (1,2). Combined global identification of the molecular elements and their physical interactions opened a new avenue for depicting cellular networks and understanding the biomolecular processes that occur in living systems (3,4).

It is clear that over the last decade there has been a great deal of effort to build biological databases and resources providing detailed information about the 'molecular interactions' (MI) determined in thousands of experimental studies in different biological systems, performed either using small-scale or large-scale technologies and reported in thousands of publications. Within these efforts it is worth mentioning the work of international consortiums such as IMEx (http://www.imexconsortium.org) (5) which include many primary databases as partners (such as DIP, IntAct, MINT) (6–8) or observers (such as BioGRID) (9), who have made important contributions toward creating well established standards for molecular interactions (10), as well as important collaborative efforts for providing integrated access to multiple types of molecular interactions from many resources (11–13).

As of January 2016, a search in PubMed (www.ncbi.nlm.nih.gov/pubmed) with the term 'protein–protein interaction' revealed 9,687 research articles, most published in the last five years. This indicates that current biomolecular research is highly interested in finding the molecular partners of the proteins or the gene products that are

*To whom correspondence should be addressed. Tel: +34 923 294819; Fax: +34 923 294743; Email: jrivas@usal.es; APID (Agile Protein Interactomes DataServer): http://apid.dep.usal.es

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

studied in very different biological scenarios. Such interest demands an easy way to provide and visualize interacting proteins in a proteome-wide context. There are many bioinformatics tools and servers that provide information about protein interactions and protein functional associations. An extensive list can be found on Pathguide (http://www.pathguide.org/) (14) which includes more than five hundred biological pathway related resources and molecular interaction related resources. Moreover, there is a group of online resources which provides integration of both experimentally known and computationally predicted interactions, aiming for thorough comprehensiveness and coverage. These include STRING (15), GeneMANIA (16), FunCoup (17), ConsensusPathDB (18), I2D (19) and others. Such resources aim to integrate all types of interactions, as defined in their scopes. The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins), for example, is dedicated to finding all types of 'functional associations' between proteins, on a global scale (15).

The goal of the web server presented here (APID, Agile Protein Interactomes DataServer) is different because it does not include either 'predicted' protein interactions or 'functional associations' between proteins that do not reveal physical contacts established between two or more proteins based on specific biomolecular forces. In fact, many genetic studies have provided interesting 'functional associations' between individual pairs of genes that are defined as 'genetic interactions' (20) and are reported in several of the resources cited above (7,15,16). However, APID is focused solely on the generation and delivery of unified compendiums of known and experimentaly proven protein–protein physical interactions (PPIs). The protein interactions are provided, including quality levels associated with the number of experiments, methods and publications that report each interaction, and they are organized in interactomes per organism, mapped to their respective proteomes.

APID: providing proteome-based interactomes at different quality levels

APID (Agile Protein Interactomes DataServer) is a bioinformatics web server developed to provide protein interactomes at different quality levels and allowing their analysis and visualization as networks. This resource is a new, fully redesigned version of the APID web server (21) that provides a comprehensive collection of protein interactions for 448 organisms derived from the integration of known experimentally validated protein–protein physical interactions (22). Construction of the interactomes is done with a methodological approach (detailed below) to report quality levels and coverage over the proteomes for each organism included. Figure 1 presents a view of APID main web page showing an example for the *Escherichia coli (strain K12)* interactome (Figure 1A). In other panels, the figure presents statistics about several interactomes provided (Figure 1B and E), as well as images of the interactions display tool (Figure 1C) and the network tool (Figure 1D) where the colored pie charts on the nodes present user-selected biological functions that are shared by several proteins in the network.

As a whole, APID provides easy access to the interactomes from specific species and includes a global uniform compendium of 90,379 distinct proteins and 678,441 singular interactions. APID unifies PPIs from five major primary databases of molecular interactions (BioGRID (9), DIP (6), HPRD (23), IntAct (7), MINT (8)); from some specific repositories not included in the previous ones (BioPlex, http://wren.hms.harvard.edu/bioplex/) (2) and also from experimentally resolved 3D structures of protein complexes (PDB, http://www.rcsb.org/pdb/home/home.do) (24), where more than two proteins had been identified.

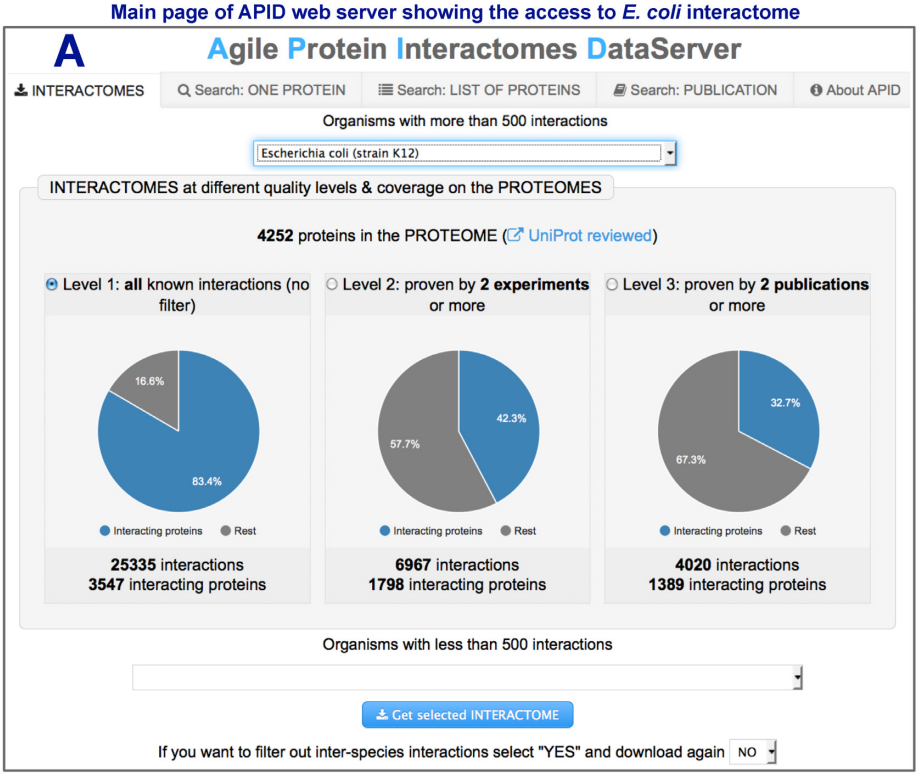
To incorporate the 3D structural information, 45,410 interfaces corresponding to 8,388 structures from the PDB were analyzed, searching for specific PPIs involving two different UniProt IDs (i.e. two distinct proteins). Using the criteria defined in PDBsum for protein–protein contacts (25), all of the interfaces between two protein chains were tested for at least one salt bridge, one disulfide bond or one hydrogen bond inferred from the 3D molecular proximity and atomic configuration (25). Interacting protein pairs found in this manner were registered with the corresponding PDB identifiers (PDB IDs), in order to count the specific number of 3D structures that validate each PPI. This process allowed us to assign 8,215 3D structures to 3,220 interactions. Details of the interfaces within these structures are provided on the web server, as they are considered to be one of the most credible proofs of the existence of a protein interaction.

Network viewer to explore and analyze protein interactions

In order to facilitate the construction and interactive exploration of PPI networks, APID includes a new network visualization tool that is a native, web-based app which follows *HTML/CSS/JavaScript* standards and does not require third-party software, such as *Java* or *Flash*, on the client side. The app is based on Cytoscape.js, a *JavaScript* library for programmers (http://js.cytoscape.org/) (26) that was used to build the core of the viewer and which gave us the framework for developing new graphcentric utilities, such as: a drag-and-drop color tagging system for the nodes to highlight multiple functional annotations; real-time filters for interactions based on the number of experiments that demonstrate them or on the presence of 3D structural data; different graph layouts (Figure 1D). In this way, the APID network viewer enables the construction of sub-interactomes using query lists of proteins of interest and visual exploration of the corresponding networks, including an interactive selection of the value of the interactions (i.e. the reliability of the 'edges' in the network) as well as an interactive mapping of the protein functional environments (i.e. the functional annotations of the 'nodes' in the network), that allows identification of nodes with shared functions (Figure 1D). This mapping uses annotation to four biological functional spaces: Gene Ontology (BP, MF and CC) (27); InterPro (28); Pfam (29) and Reactome (30). All of the networks produced with this tool can be exported either as figures (.jpg, .png) or as flat files (.txt) to be read using other software for graph visualization or analysis. In particular, flat files from the networks and all the interactomes provided by APID can be loaded

Main page of APID web server showing the access to *E. coli* interactome

A

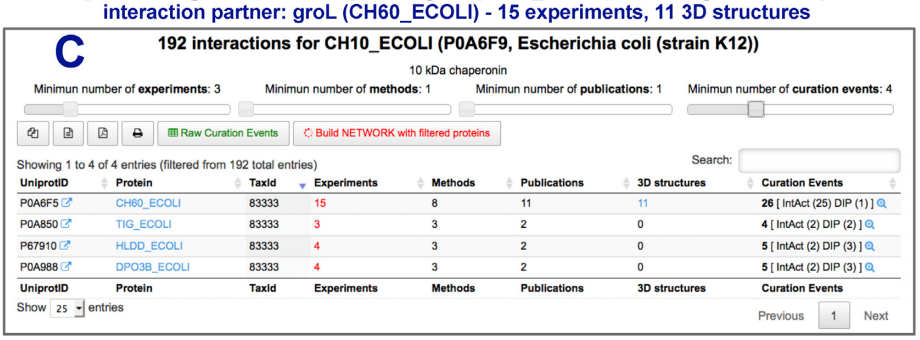


B **Examples of *E. coli* interactomes selected at 3 levels of experimental validation, filtering out inter-species interactions**

Networks	Number of distinct experiments ≥ 2	≥ 3	≥ 4
Proteins (nodes)	1770	1122	531
Interactions (edges)	6936	2870	888
average connectivity	**3.92**	**2.56**	**1.67**

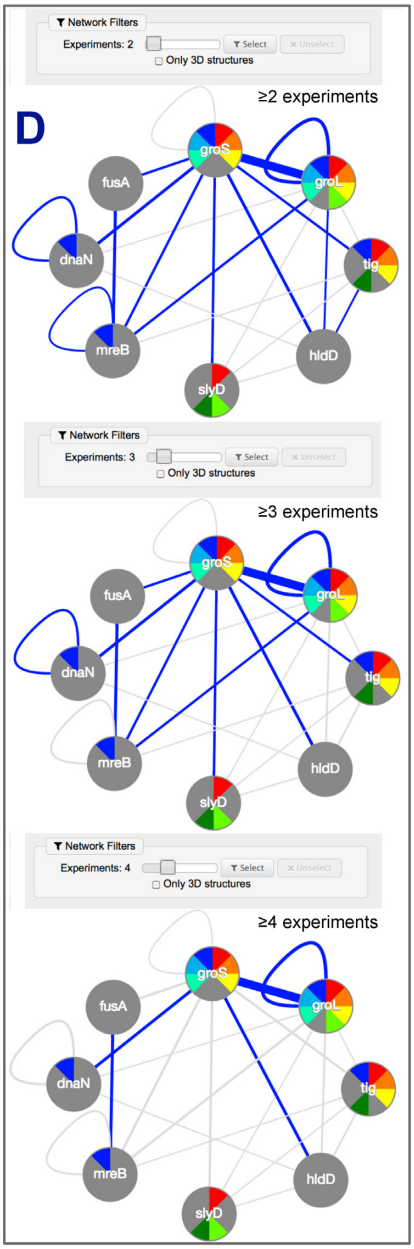
Example showing interactions of protein groS (CH10_ECOLI), including its most reported interaction partner: groL (CH60_ECOLI) - 15 experiments, 11 3D structures

C



View of groS protein closest interactors done with APID network tool

D



E **Table with statistics about the interactomes of 8 organims included in APID**

PROTEOMES and INTERACTOMES (protein-protein physical interactions)	PROTEOME	Data in APID_interactomes (v30Jan2016) Proteins in INTERACTOME	Interactions in INTERACTOME	Proteome Coverage (%)
Homo sapiens	69,986	**28,749**	**334,817**	41.08
Mus musculus	50,189	**17,478**	**53,676**	34.82
Arabidopsis thaliana	31,477	**10,394**	**40,018**	33.02
Rattus norvegicus	29,885	**5,556**	**8,797**	18.59
Caenorhabditis elegans	26,596	**5,569**	**14,762**	20.94
Drosophila melanogaster	22,005	**11,692**	**59,977**	53.13
S. cerevisiae (strain ATCC 204508 / S288c)	6,721	**6,823**	**128,740**	101.52
Escherichia coli (strain K12)	4,252	**3,547**	**25,335**	83.42
PROTEOMES Coverage (mean %)				**48.32**

Figure 1. Presentation of APID (Agile Protein Interactomes DataServer) including five panels with different views and information. (**A**) Panel showing the main web page of the APID web server after selecting '*E. coli (strain K12)*' to present its interactomes. (**B**) Table presenting the size that the interactome datasets from *E. coli* produced at three levels of experimental validation and filtering out inter-species interactions. (**C**) Example showing the interactions of protein groS (CH10_ECOLI), including its most reported interaction partner: groL (CH60_ECOLI), that has been validated in 15 experiments and in 11 3D structures. (**D**) View of the closest protein interactors of the groS protein using the APID network tool which allows coloring the nodes according to specific functional terms annotated to the proteins and also allows selection of the edges (in blue) according to the number of experiments that validate each interaction. (**E**) Table with the numbers included in APID about interactome sizes and proteome coverage for eight model organisms.

into Cytoscape (http://www.cytoscape.org/) (31), which is a very useful open source software platform for complex network analysis and visualization that allows the integration of multiple layers of biological information on the networks (31,32).

Comparison to other related tools

As indicated in the introduction, there are multiple bioinformatics tools or platforms that provide information about functional associations and interactions between proteins. A compendium of these can be found in Pathguide (http://www.pathguide.org/) (14). However, as far as we know, there are no servers identical to the new APID described here, focused on the integration of only experimentally validated protein–protein physical interactions. There are multiple applications or servers that took similar approaches to the first version of APID (21) and built integrated compendiums of PPIs for different organisms. Some of the most remarkable and complete ones are: iRefWeb (http://wodaklab.org/iRefWeb/) published in 2010 (33); HitPredict (http://hintdb.hgc.jp/htp/) published in 2011 (34); PINA (http://cbg.garvan.unsw.edu.au/pina/) published in 2012 (35); and Mentha (http://mentha.uniroma2.it) published in 2013 (36). These tools are currently accessible, but only two of them have been updated in 2016. We present a comparison of the PPI data corresponding to eight model organisms included in these two servers, the ones currently updated, *versus* APID (see Supplementary Data 1). This comparison indicates that APID provides interactomes with a 48.3% average coverage of the proteomes of these eight species; while iRefWeb shows a coverage of 39.3% and Mentha, 41.9%. These numbers correspond to the versions of these resources downloaded in January 2016. The increase in coverage that APID achieves with respect to the other resources may be due to several reasons: **(i)** the compared datasets may not correspond to updates of the same versions of the primary databases (despite the fact that in all cases the comparisons are of data available in January 2016); **(ii)** APID includes some new sources that are not included in iRefWeb or in Mentha (such as BioPlex and HPRD in the case of Mentha) (2); **(iii)** the different resources may not analyze the same raw files from the primary public databases. In fact, this last reason is probably the most important because, for example, Mentha integrates protein-interaction data curated by experts in compliance with IMEx curation policies, using the PSICQUIC protocol to implement an automatic procedure that, every week and without human intervention, aligns the integrated database with data regularly annotated by the primary databases (36). Therefore, anything that is not in PSICQUIC (11,12) will not be in Mentha. Another important difference is that APID uses the XML files (i.e. PSI-MI XML files) drawn from primary databases, but most of the meta-databases and servers that integrate multiple data from molecular interactions use a simpler format called MITAB (i.e. PSI-MI TAB, which is a common tab delimited format for MI data interchange: https://code.google.com/archive/p/psimi/wikis/PsimiTabFormat.wiki). More details about the procedures and methodology that APID employs to achieve an efficient integration and unification of PPI data are explained below. Finally, other differences observed are that these tools do not offer the same validation procedures with quality levels for the PPIs used in APID and do not integrate any extra information derived from the analyses of interactions in 3D structures of protein complexes.

Experimentally proven protein–protein physical interactions, unified and weighted

The APID server presents a way to evaluate and qualify PPIs based on identification of the distinct 'experiments' from the literature (i.e. from specific scientific articles reported in PubMed) that prove a given protein pair interaction. In other words, APID counts the number of 'experiments' as the number of times that the interaction between two proteins has been tested and demonstrated in a research lab with one specific method and reported in a published article. This is a different approach to the procedure followed by other PPI resources that count 'evidences' defined as the 'aggregated experimental evidences retrieved from the different databases' (8). Moreover, often these PPI resources build and provide a 'score' calculated for each interaction that is based on such counts of 'evidences' (8,11).

In APID, 'evidences' correspond to 'curation events' and they provide larger numbers than the 'experiments' because several primary databases can curate the same published articles and, when they do, it does not mean that a new experiment was done to test and validate the interaction. In fact, we performed an analysis to show that counting 'curation events' produces a clear overestimation of the interactions and, therefore, an overestimation of the size of the interactomes. Supplementary Data 2 presents a graphic comparison of the number of interactions included in the human interactome considering several numbers of 'experiments' or 'curation events'. This analysis shows that an interactome validated with 3 or more 'curation events' per interaction will be 48.5% larger than an interactome validated with 3 or more 'experiments'; thus demonstrating that producing scores based on curation events may not be very accurate.

Another fact is that in APID, counting the experiments is a simple and transparent process, since it does not attempt to calculate a 'score' derived from a rational combination of factors. The procedures to calculate such scores need to reach a compromise between every variable that describes an interaction and, therefore, are usually quite arbitrary and can sometimes be difficult to understand or confusing for the users. In fact, to illustrate the problems associated with the definition of an integrated score, we compared the results in APID for two well-known interactions that are validated by very different experimental approaches: HRAS (P01112) interaction with RAF1 (P04049) was validated in 36 singular experiments and HRAS interaction with SOS1 (Q07889) was validated in only seven singular experiments; by contrast we found 18 distinct PDB 3D structures that validate HRAS interaction with SOS1 but only three PDB structures that validate HRAS interaction with RAF1. It is very difficult to make a fair decision to rank and give a higher score to RASH–RAF1 interaction or to RASH–SOS1 interaction based on these numbers: which is better, 36 singular reported experiments or 18 distinct PDB structures?. For these reasons, we prefer to leave this discussion

open, providing all the experimental results for each singular interaction in APID and allowing the users to employ their own criteria to sort or rank the interactions. This ranking may even follow different approaches appropriate to different types of interactomic studies.

According to the strategy described, for each PPI pair the APID server provides a combination of four counts that measure the level of experimental validation: (**i**) the number of 'experiments' (calculated as described above); (**ii**) the number of 'methods' that validate such interaction (following PSI-MI ontology for the identification of different 'interaction detection methods') (37,38); (**iii**) the number of 'publications' that have reported such interaction (including specific PMIDs from PubMed); (**iv**) the number of '3D structures' from the PDB that include two proteins interacting in a specific way at molecular level (i.e. with H-bonds or other types of specific bonding inferred from the PDB) (24,25).

Architecture of the web server and procedures for integrating and unifying PPI data

The APID server was built with a protein and proteome-centered strategy, using the UniProt database (http://www.uniprot.org) as the main guide to identify and handle all of the proteins and map them into the reference proteomes of each species (based on the new proteome identifiers that UniProt recently developed: http://www.uniprot.org/proteomes/) (39). In this way UniProt, including both Swiss-Prot and TrEMBL, was used as the main reference database and we used protein or gene identifier recursive mapping to UniProtKB AC/ID as the key way to integrate and unify data, thus avoiding duplications or incorrect identifications.

To provide a global view of the methodology and procedures followed to build APID, a graphic scheme presents the main workflow with the pipelines and steps applied to integrate the PPI data. This scheme is included as Supplementary Data 3 and also as a figure on the APID website.

With coverage as one of the main objectives, the procedure begins with an exhaustive parsing of the complete raw PSI-MI XML files from the five major public databases of molecular interactions: BioGRID (9), DIP (6), HPRD (23), IntAct (7), MINT (8). A TSV file with the data from BioPlex project (2) is also downloaded, parsed and integrated. For this part of the workflow, we designed a protocol based on JAMI (Java Framework for Molecular Interactions) (40) that processed all of the XML entries contained in the downloaded files. This approach allowed us to acquire all of the information contained in the source databases, and design a pipeline to discard any dataset that was incomplete or not appropriate, such as: (**i**) any participant of an interaction that is not a protein; (**ii**) any apparent participant with an ID that could not be matched to a UniProt ID; (**iii**) any Uniprot ID that was obsolete and deprecated and could not be replaced by a current UniProt ID. This procedure guaranteed that every participant in an interaction was registered as a protein and mapped to the UniProt database (SwissProt or TrEMBL). Gene names (i.e. official gene symbols such as KRAS for RASK_HUMAN or Tp53 for P53_MOUSE) were added as an annotation after the ID mapping to facilitate identification of participants in each interaction and the use of the PPI data in other resources that employ gene identifiers. At the beginning of the workflow (Supplementary Data 3), for the records reporting protein interactions that include more than two proteins (i.e. records with multiple proteins) we applied the *spoke model* to expand the data and generate binary interactions from these co-complex data (22).

Once the ID mapping was completed, a unification pipeline was followed to merge data. For example: (**i**) curation events from different sources that reported the same interactions after protein ID matching, and (**ii**) isoforms of the same protein reported as different interactors. This unification allowed the identification of singular interaction pairs and eliminated many duplications. Unification of the interaction protein pairs was always performed following HUPO Proteomics Standards (PSI-MI) (37,38) including the ontology of terms with its hierarchy (as shown in http://www.ebi.ac.uk/ols/beta/ontologies/mi).

APID: download protein interactomes and visualize networks of specific PPI sets

The APID web server is fully functional, free and open to all users at URL: http://apid.dep.usal.es. The server's first page allows downloading of protein interactions for more than 400 organisms at three different quality levels: level **1**) all known interactions; level **2**) interactions proven by 2 or more experiments; level **3**) interactions reported in two or more research publications. Data for organisms with more than 500 known interactions are presented in an alphabetically ordered drop-down list to allow rapid access. The rest of the organisms are included in a second similar drop-down list. For each organism the interactomes can be downloaded, including interactions with proteins from other species (inter-species interactions) or by simply filtering out such interactions. The server also includes pages with search engines for single proteins ('Search: ONE PROTEIN') or lists of proteins ('Search: LIST OF PROTEINS'), using either UniProt AC/ID identifiers or standard gene/protein Symbols. On another page the server includes a search tool ('Search: PUBLICATION') to query by published articles (i.e. a PubMed ID number, PMID) in order to find all of the PPIs that have been reported in a given publication, including all of the information about such interactions that is currently integrated in the server. APID includes examples in all of these search pages.

Search results deliver PPIs in a tabular format, showing all the interactor pairs with protein names (UniProt IDs) and taxonomy IDs (http://www.ncbi.nlm.nih.gov/taxonomy) to identify the species, plus all experimental evidences counted and presented in five different columns: (**i**) number of experiments, (**ii**) number of methods, (**iii**) number of publications, (**iv**) number of 3D structures (PDBs) and (**v**) number of curation events (including source databases) (see Figure 1C). The data can be sorted by any of the columns and filtered to select a minimum number of experiments, methods, publications or curation events. Once a set or subset of interactions is displayed, the web allows one to build a network with the network viewer app for the proteins and interactions selected.

This website also contains a section called 'About_APID' with useful information including a 'HELP' page with a brief tutorial presenting some simple cases that illustrate how to use the server. It also includes a page named 'METHODOLOGY' that provides a global view of the procedures followed to build APID with a figure presenting the main workflow with the pipelines and steps applied to integrate the PPI data (this figure is also included here as Supplementary Data 3). Another page named 'DOWNLOADS' allows downloading (in MITAB format) of all of the raw curation events from PPIs that are integrated in APID resulting from unification of the primary public databases, grouped into single files by organism. Two other pages ('STATISTICS' and 'ACKNOWLEDGEMENTS') provide more information about source databases, versions, updates, references and technologies. The site also includes a 'Show_HELP' button on all pages, which presents captions with brief descriptions of each one of the elements viewed on a given page. Throughout the site, server links to the corresponding source databases are included, such as: UniProt for proteins; UniProt-Proteomes for proteomes; PubMed for publications; PDB for 3D structures; and the corresponding primary molecular interaction databases for all the singular curation events reported.

Finally, the web server presented here is a fully redesigned PPI resource providing agile access to protein interactomes, but it maintains the value and credit of the first APID version (published in 2006 in *Nucleic Acids Research, Web Server* issue) (15) keeping the same acronym for its name. We feel that this will allow it to be of better service to the research community and facilitate a broader use.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Spanish Government, Ministerio de Economía y Competitividad, Instituto de Salud Carlos III [PI12/00624 and PI15/00328 to Dr J. De Las Rivas group]; and EU Joint Programme, JPND [AC14/00024 to Dr J. De Las Rivas group]. Regional Government, Junta de Castilla y León [BIO/SA68/13 to Dr J. De Las Rivas group].
Conflict of interest statement. None declared.

REFERENCES

1. Rolland,T., Tasan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
2. Huttlin,E.L., Ting,L., Bruckner,R.J., Gebreab,F., Gygi,M.P., Szpyt,J., Tam,S., Zarraga,G., Colby,G., Baltier,K. *et al.* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, **162**, 425–440.
3. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
4. Vidal,M., Cusick,M.E. and Barabasi,A.L. (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.
5. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S., Cesareni,G. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
6. 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.
7. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
8. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
9. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
10. Orchard,S., Salwinski,L., Kerrien,S., Montecchi-Palazzi,L., Oesterheld,M., Stumpflen,V., Ceol,A., Chatr-aryamontri,A., Armstrong,J., Woollard,P. *et al.* (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **25**, 894–898.
11. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
12. del-Toro,N., Dumousseau,M., Orchard,S., Jimenez,R.C., Galeota,E., Launay,G., Goll,J., Breuer,K., Ono,K., Salwinski,L. *et al.* (2013) A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res.*, **41**, W601–W606.
13. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
14. Bader,G.D., Cary,M.P. and Sander,C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.
15. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
16. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res.*, **41**, W115–W122.
17. Schmitt,T., Ogris,C. and Sonnhammer,E.L. (2014) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res.*, **42**, D380–D388.
18. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
19. Niu,Y., Otasek,D. and Jurisica,I. (2010) Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, **26**, 111–119.
20. Mani,R., St Onge,R.P., Hartman,J.L. 4th, Giaever,G. and Roth,F.P. (2008) Defining genetic interaction. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 3461–3466.
21. Prieto,C. and De Las Rivas,J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.*, **34**, W298–W302.
22. De Las Rivas,J. and Fontanillo,C. (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, **6**, e1000807.
23. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
24. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
25. de Beer,T.A., Berka,K., Thornton,J.M. and Laskowski,R.A. (2014) PDBsum additions. *Nucleic Acids Res.*, **42**, D292–D296.
26. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.

27. Huntley,R.P., Sawford,T., Mutowo-Meullenet,P., Shypitsyna,A., Bonilla,C., Martin,M.J. and O'Donovan,C. (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057–D1063.
28. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
29. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
30. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
31. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
32. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocols*, **2**, 2366–2382.
33. Turner,B., Razick,S., Turinsky,A.L., Vlasblom,J., Crowdy,E.K., Cho,E., Morrison,K., Donaldson,I.M. and Wodak,S.J. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, **2010**, baq023.
34. Patil,A., Nakai,K. and Nakamura,H. (2011) HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.*, **39**, D744–D749.
35. Cowley,M.J., Pinese,M., Kassahn,K.S., Waddell,N., Pearson,J.V., Grimmond,S.M., Biankin,A.V., Hautaniemi,S. and Wu,J. (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res.*, **40**, D862–D865.
36. Calderone,A., Castagnoli,L. and Cesareni,G. (2013) mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods*, **10**, 690–691.
37. 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.
38. Kerrien,S., Orchard,S., Montecchi-Palazzi,L., Aranda,B., Quinn,A.F., Vinod,N., Bader,G.D., Xenarios,I., Wojcik,J., Sherman,D. *et al.* (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
39. UniProt_Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
40. Orchard,S., Albar,J.P., Binz,P.A., Kettner,C., Jones,A.R., Salek,R.M., Vizcaino,J.A., Deutsch,E.W. and Hermjakob,H. (2014) Meeting new challenges: The 2014 HUPO-PSI/COSMOS Workshop: 13-15 April 2014, Frankfurt, Germany. *Proteomics*, **14**, 2363–2368.

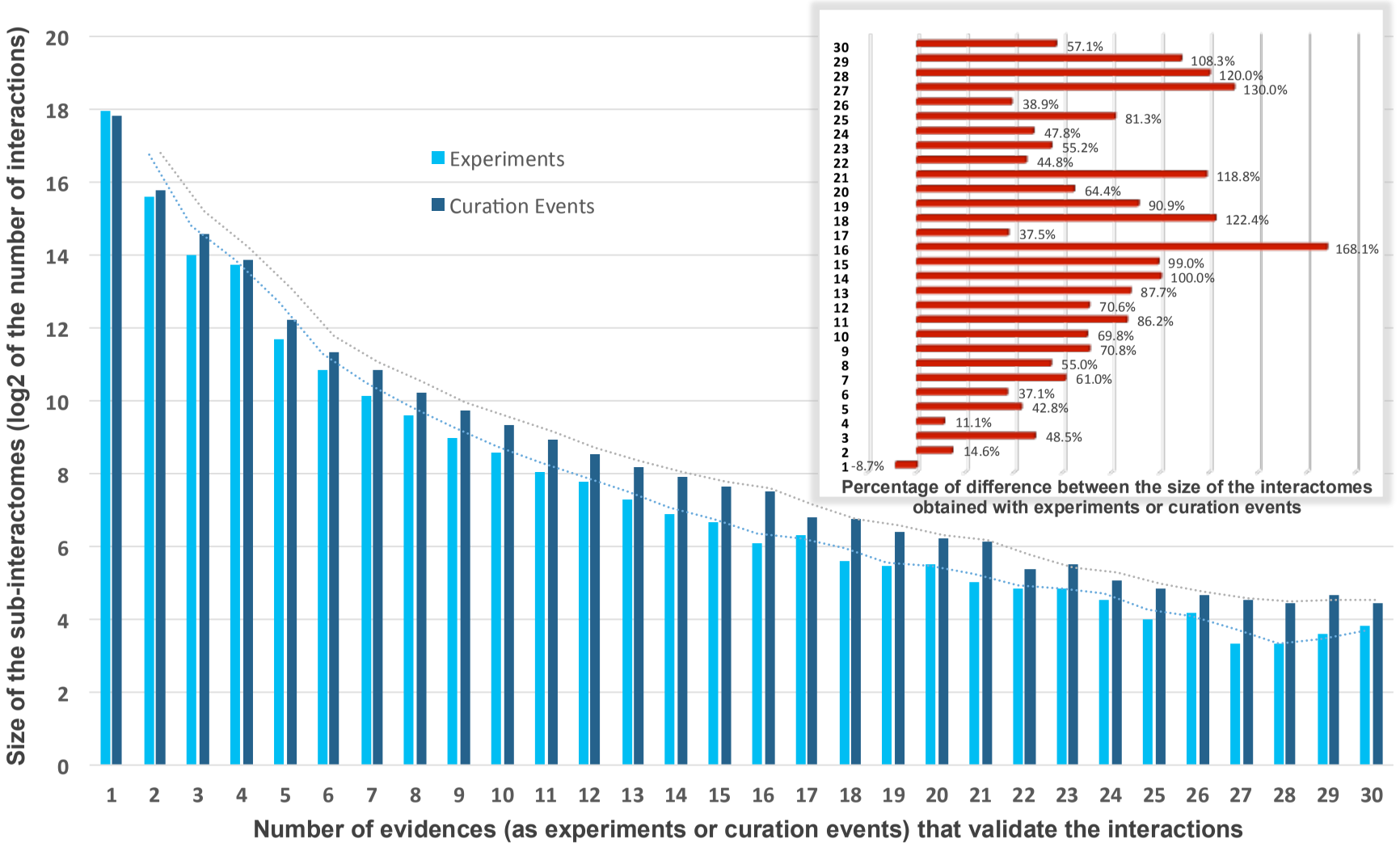
SUPPLEMENTARY DATA

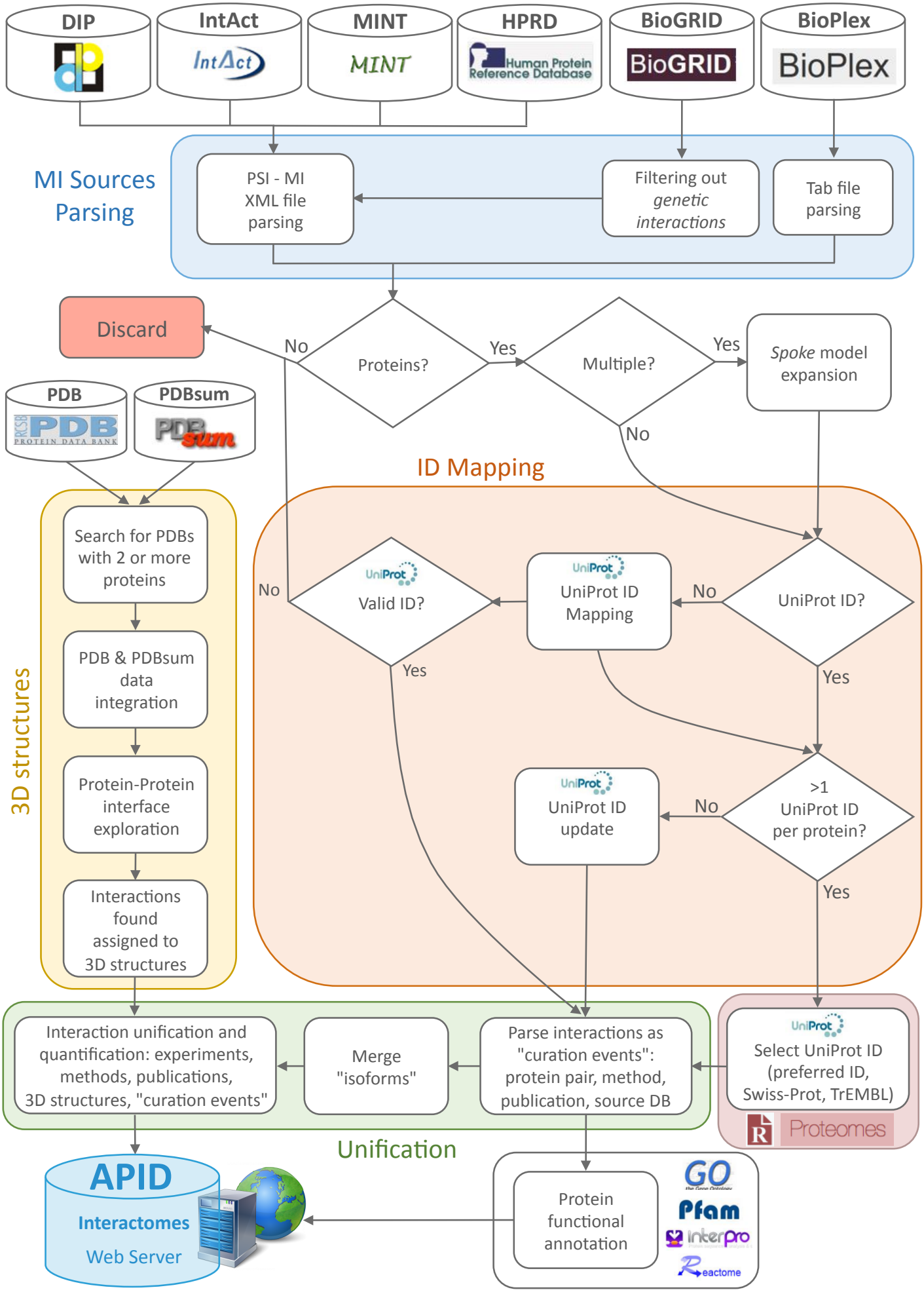
Supplementary Data 1. Comparative table showing the size of 8 model organism interactomes reported by three protein interaction resources: APID, iRefWeb and MENTHA. The data correspond only to experimentally validated protein-protein physical interactions integrated in these platforms for the proteins of these organisms. Coverage of the interacting proteins across the complete proteomes of each organism is also calculated. (NOTE: The proteome coverage for *S. cerevisiae* (strain ATCC 204508 / S288c) is above 100% because it includes TrEMBL proteins that exceed the UniProt reference PROTEOME: http://www.uniprot.org/proteomes/UP000002311).

Supplementary Data 2. The main graph presents a comparison of the number of interactions included in the human interactome considering several number of "experiments" (light blue) or "curation events" (dark blue). The number of evidences (as experiments or curation events) that validate the interactions are represented versus the size of the corresponding sub-interactomes (in a log2 scale). The small graph inside (red bars) indicates the percentage (%) of size change in the interactomes measured using "experiments" or "curation events", taking into consideration sub-sets of the interactome corresponding to interactions at each given number of validations (from 1 to 30).

Supplementary Data 3. Figure that provides a global view of the methodology and procedures followed to build APID. The graphic scheme presents the main workflow with the pipelines and steps applied to integrate the protein-protein interaction data.

		Data in APID_interactomes (v30Jan2016)			Data in iRefWeb (v13.0)			Data in MENTHA (v17Jan2016)		
PROTEOMES and INTERACTOMES (protein-protein physical interactions)	**PROTEOME**	**Proteins in INTERACTOME**	**Interactions in INTERACTOME**	Proteome Coverage (%)	**Proteins in INTERACTOME**	**Interactions in INTERACTOME**	Proteome Coverage (%)	**Proteins in INTERACTOME**	**Interactions in INTERACTOME**	Proteome Coverage (%)
Homo sapiens	69,986	**28,749**	**334,817**	41.08	18,841	222,069	26.92	17,390	223,978	24.85
Mus musculus	50,189	**17,478**	**53,676**	34.82	9,118	30,117	18.17	8,777	24,878	17.49
Arabidopsis thaliana	31,477	**10,394**	**40,018**	33.02	7,879	21,454	25.03	8,346	27,722	26.51
Rattus norvegicus	29,885	**5,556**	**8,797**	18.59	3,330	8,286	11.14	2,388	4,057	7.99
Caenorhabditis elegans	26,596	**5,569**	**14,762**	20.94	5,506	14,102	20.70	5,502	13,965	20.69
Drosophila melanogaster	22,005	**11,692**	**59,977**	53.13	9,509	44,996	43.21	10,509	43,794	47.76
S. cerevisiae (strain ATCC 204508 / S288c)	6,721	**6,823**	**128,740**	101.52	6,083	117,029	90.51	6,159	100,948	91.64
Escherichia coli (strain K12)	4,252	**3,547**	**25,335**	83.42	3,351	15,269	78.81	4,181	26,033	98.33
PROTEOMES Coverage (mean %)				**48.32**			**39.31**			**41.91**


Size of the sub-interactomes (log2 of the number of interactions)
Number of evidences (as experiments or curation events) that validate the interactions
Experiments
Curation Events
Percentage of difference between the size of the interactomes obtained with experiments or curation events
30 57.1%
29 108.3%
28 120.0%
27 130.0%
26 38.9%
25 81.3%
24 47.8%
23 55.2%
22 44.8%
21 118.8%
20 64.4%
19 90.9%
18 122.4%
17 37.5%
16 168.1%
15 99.0%
14 100.0%
13 87.7%
12 70.6%
11 86.2%
10 69.8%
9 70.8%
8 55.0%
7 61.0%
6 37.1%
5 42.8%
4 11.1%
3 48.5%
2 14.6%
1 -8.7%


DIP
IntAct
MINT
HPRD
Human Protein Reference Database
BioGRID
BioPlex
MI Sources Parsing
PSI - MI XML file parsing
Filtering out genetic interactions
Tab file parsing
Discard
No
Proteins?
Yes
Multiple?
Yes
Spoke model expansion
No
PDB
PDBsum
ID Mapping
Search for PDBs with 2 or more proteins
PDB & PDBsum data integration
Protein-Protein interface exploration
Interactions found assigned to 3D structures
3D structures
No
Valid ID?
Yes
UniProt ID Mapping
No
UniProt ID?
Yes
UniProt ID update
No
>1 UniProt ID per protein?
Yes
Interaction unification and quantification: experiments, methods, publications, 3D structures, "curation events"
Merge "isoforms"
Parse interactions as "curation events": protein pair, method, publication, source DB
Select UniProt ID (preferred ID, Swiss-Prot, TrEMBL)
Proteomes
Unification
APID
Interactomes
Web Server
Protein functional annotation
GO
Pfam
InterPro
Reactome