UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA
Ingeniería de Telecomunicación

Desarrollo de un sistema para la integración de datos moleculares sobre enfermedades raras

Sara Fernández Novo

MAYO 2016

DESARROLLO DE UN SISTEMA PARA LA INTEGRACIÓN DE DATOS MOLECULARES SOBRE ENFERMEDADES RARAS

AUTORA: Sara Fernández Novo
TUTORA: Mónica Chagoyen Quiles
PONENTE: Ana Mª González Marcos

Computational Systems Biology Group (CNB-CSIC)
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Mayo 2016

Resumen

En esta memoria se presenta ODCs (*Orphan Disease Connections*), una herramienta accesible en http://csbg.cnb.csic.es/odcs que permite explorar potenciales relaciones moleculares entre enfermedades raras. ODCs establece una red de conexiones entre este tipo de enfermedades a través de la integración de genes de susceptibilidad de las enfermedades e interacciones entre sus proteínas. Las afecciones poco comunes requieren un tratamiento especial debido a la baja incidencia en la población y a que la mayoría de ellas son monogénicas. Por tanto, adquiere especial importancia recopilar el máximo de información sobre este tipo de enfermedades y establecer relaciones entre ellas.

ODCs se nutre de los datos proporcionados por Orphadata, la principal base de datos de enfermedades raras, y de HIPPIE, recurso que integra distintas fuentes de datos de interacciones entre proteínas humanas. La herramienta establece la red de relaciones a través de un sistema de puntuación teniendo en cuenta los genes comunes y las interacciones entre las proteínas de las enfermedades. Las relaciones resultantes se pueden consultar de diferentes formas: búsqueda centrada en una enfermedad, búsqueda centrada en la conexión entre dos enfermedades o búsqueda centrada en un gen. La herramienta además proporciona enlaces externos a otros recursos biomédicos on-line que permiten ampliar la información sobre las enfermedades, los genes y las interacciones entre proteínas.

Actualmente, ODCs es una herramienta plenamente operativa, cuya base de datos contiene 54.941 relaciones entre 2.818 enfermedades. De estas relaciones, 5.263 corresponden a relaciones basadas en gen compartido.

ODCs es la primera herramienta que incluye relaciones entre enfermedades raras por medio del interactoma. Eso ha permitido que en la red actual se haya pasado de un 67,8% de enfermedades conectadas al 91,7%, generando un total de 49.678 nuevas conexiones entre las 3.073 enfermedades con gen conocido catalogadas en Orphadata, lo cual es un logro significativo, avalado por el grupo de investigación *Computational Systems Biology Group* (CNB-CSIC).

Palabras clave

Biomedicina, enfermedad rara, enfermedad huérfana, red de enfermedades, gen de susceptibilidad, interactoma humano, relaciones moleculares entre enfermedades, signo clínico, sistema de consultas, interfaz de usuario, base de datos.

Abstract

This M.Sc. Thesis presents ODCs (*Orphan Disease Connections*), a novel resource to explore potential molecular relations between rare diseases, available at http://csbg.cnb.csic.es/odcs. These molecular relations have been established through the integration of disease susceptibility genes and human protein-protein interactions. Due to their low prevalence and the fact that most rare diseases are monogenic, it is important to gather all the information available on these diseases and look for relations between them.

ODCs is built upon two main sources of data: Orphadata, a rare disease reference dataset, and HIPPIE, a resource of human protein interactions compiled from a large number of databases. This novel resource builds a network of rare disease relations through a scoring system based on shared genes and protein interactions. ODCs includes three easy-to-use types of searches (disease, diseases connection and gene) and provides both textual and graphical output in order to explore the relations. Additionally, OCDs contains link-outs to conveniently navigate to other important biomedical resources to amplify the information on diseases, genes and proteins.

The database currently contains 54,941 relations between 2,818 diseases. 5,263 of those relations are based solely on shared genes.

ODCs is the first query tool which includes interactome association between diseases. Due to this, the number of connected diseases has increased from 67.8% to 91.7% in the current network, with a total of 49,678 new connections being established between the 3,073 diseases with associated gene in Orphadata database. That is a significant achievement and it is endorsed by the *Computational Systems Biology Group* (CNB-CSIC).

Keywords

Biomedicine, rare disease, orphan disease, disease network, susceptibility gene, human interactome, molecular relations between diseases, clinical sign, user interface, web server, database.

Agradecimientos

Me gustaría dar las gracias a todas las personas que me han ayudado durante la realización de este proyecto, y por extensión, a lo largo de la etapa que hoy concluye con la presentación de este escrito. Ha sido larga, pero por fin se acabó.

Comienzo por mi tutora. Gracias Mónica por confiar en mí para llevar a cabo este proyecto, por tus directrices y por brindarme la oportunidad de crear una herramienta que confiamos sirva de ayuda en la investigación de las enfermedades raras.

Quiero acordarme también, aunque ahora quede lejos, de todos los profesores y compañeros que estuvieron presentes en la Universidad, especialmente de los que trabajaron mano a mano conmigo en prácticas o compartieron horas de estudio. De todos ellos he aprendido algo.

Gracias a aquéllos que, además de compañeros, fueron y son amigos. Gracias a los que seguís cerca, a quienes cuesta ver e incluso a quienes ya no están, pero estuvieron. Gracias a quien compartió conmigo un café en las escaleras, un refresco al sol, una charla en aquel campo de fútbol, una cerveza en un bar o todas aquellas horas muertas en la “uni”, y a su vez, llenas de vida.

Gracias a las “nenas”, por todos los momentos de alegrías, y de algunas penas, pero siempre compartidas; por los viajes, las quedadas y las conversaciones interminables.

Gracias a mis excompañeros de trabajo, los cuales se han convertido en un gran apoyo. En especial, gracias Borja por tus consejos a la hora de escribir esto, y gracias Luis, por esperarme un escalón más arriba y animarme siempre a mejorar.

Por último, quiero dar las gracias a las personas más importantes de mi vida: mi familia. Gracias a mi hermano Carlos, y sobre todo, gracias a mis padres. Gracias Remedios y Jesús por permitirme estudiar esta carrera y por vuestra ayuda y apoyo constantes a lo largo de todo el camino. Sin vosotros nada de esto tendría sentido. Gracias por vuestra paciencia y por estar ahí siempre. Esto es por y para vosotros. Os quiero.

Gracias,

Sara

Índice

Índice de tablas

Índice de figuras

1. Introducción

1.1. Marco del proyecto - Motivación

Las enfermedades raras son aquellas que afectan a un número relativamente reducido de personas; concretamente en Europa, una enfermedad rara se define como aquella que afecta a menos de 5 por cada 10.000 habitantes.

A pesar de la baja incidencia de cada enfermedad rara en la población, actualmente existen entre 5.000 y 7.000 patologías poco frecuentes conocidas, las cuales en conjunto afectan a un gran número de personas: según la Organización Mundial de la Salud, al 7% de la población mundial. Se estima que en España viven más de 3 millones de personas que padecen alguna enfermedad rara.

Las enfermedades consideradas raras tienen ciertas particularidades que dificultan su estudio. Por su propia definición, al afectar a un pequeño grupo de la población, la información sobre ellas es escasa, se encuentra muy dispersa y en general es poco accesible; lo que determina que algunos sistemas de consulta se muestren poco efectivos.

Otra denominación de este tipo de enfermedades es la de enfermedades huérfanas. Esto se debe al inconveniente que presentan estas patologías en la investigación clínica y experimental, estando así "huérfanas" del interés del mercado y de las políticas de salud pública.

Se pueden encontrar diversas base de datos y herramientas que contienen información sobre enfermedades raras, como por ejemplo Orphanet y Orphadata, portal y base de datos genéticos y clínicos, respectivamente. Estos sistemas son muy valiosos por la cantidad de información que contienen tanto para profesionales de la salud como para pacientes. Sin embargo, en ellos, cada enfermedad se presenta como una entidad independiente sin dar una especial importancia a las relaciones entre enfermedades, aspecto que, como se verá más adelante, puede ser realmente útil.

1.2. Definición del problema

El establecimiento y el estudio de relaciones moleculares entre enfermedades es un área de investigación muy activa en medicina de sistemas[1]. La medicina de sistemas trata de estudiar las enfermedades a través de la integración de una gran variedad de datos biomédicos y de su análisis computacional.

Una de las maneras de abordar el estudio sistémico de las bases moleculares de las enfermedades en general, y de las enfermedades raras en particular, es estableciendo relaciones moleculares entre ellas. En base a estas relaciones se pueden generar redes de enfermedades donde los nodos son afecciones y los enlaces relaciones moleculares compartidas. La búsqueda de relaciones moleculares potenciales entre enfermedades es especialmente importante en el caso de las enfermedades raras por su baja incidencia en la población.

Sin embargo, a la hora de relacionar entre sí enfermedades raras, el primer inconveniente que se encuentra radica en el escaso número de genes asociados a ellas. Frente a las enfermedades frecuentes, causadas por la combinación de un importante número de factores genéticos y ambientales, las enfermedades raras son en su mayoría monogénicas. De esta manera, si se intenta generar una red de enfermedades raras a partir de los genes que comparten, se observa un alto número de afecciones aisladas en pequeños grupos sin conexión con el resto de la red[2]. Por lo tanto, es necesario definir una estrategia que aumente el número de potenciales relaciones moleculares.

Con la intención de permitir el estudio global de las enfermedades con base genética conocida, en la literatura se han llevado a cabo diversas estrategias. Algunas de ellas se han utilizado en sistemas informáticos que permiten la consulta de relaciones entre enfermedades. Sin embargo, ninguno de estos sistemas está específicamente centrado en las enfermedades raras. Por tanto, no existe un sistema que permita la consulta de las relaciones moleculares entre enfermedades raras de manera sencilla y clara.

1.3. Objetivos

El objetivo general que se pretende alcanzar con el presente proyecto es la creación de una red de potenciales relaciones moleculares entre enfermedades raras y de una herramienta que permita la consulta y exploración detallada de estas relaciones. El conjunto ha de facilitar el estudio de las afecciones y la búsqueda de mecanismos moleculares de relación.

Los objetivos específicos para alcanzar el objetivo general son:

1. Definir una estrategia que permita establecer el mayor número posible de relaciones moleculares entre enfermedades raras.
2. Diseñar y desarrollar una herramienta web que permita a los investigadores y personal clínico consultar estas relaciones.

Por tanto, a lo largo de este proyecto se pretende desarrollar un sistema que permita explorar el máximo de información referente a las enfermedades raras, especialmente las relaciones entre ellas, integrando datos genéticos y clínicos de las distintas enfermedades.

1.4. Estructura de la memoria

Una vez expuestos los objetivos, el presente trabajo se estructura de la siguiente manera:

Las características específicas de este proyecto requieren explicar con cierto detalle la parte más biológica, para centrarse después en la parte más tecnológica. De este modo, en el siguiente capítulo de esta memoria se describe el estado del arte. Comienza con una breve definición de las fuentes de información sobre enfermedades raras y sobre proteínas de las cuales se han extraído los datos necesarios para la realización de este proyecto. A continuación se explica cómo se estudian las enfermedades raras mediante redes de relación y cuáles son los sistemas de consulta actuales. Por último, como contribución propia al estado del arte de este proyecto, se incluye un apartado explicando la nueva estrategia para la relación de enfermedades raras mediante interacciones proteína-proteína que sirvió como primera prueba de concepto.

En el capítulo tercero se presenta la arquitectura diseñada y las tecnologías utilizadas para la implementación de la herramienta desarrollada (ODCs). Y a continuación, se describen con detalle los procedimientos y los módulos generados a lo largo de este proyecto.

En el cuarto capítulo se detalla el análisis previo que se utilizó como referencia y que da validez y sentido al resto del trabajo realizado durante el proyecto. Además se incluyen los resultados obtenidos a partir del propio proyecto.

En el último capítulo se exponen las conclusiones a las que se ha llegado tras la realización de este proyecto fin de carrera, tanto a nivel técnico como científico.

2. Recursos científicos y estado del arte

A lo largo de este capítulo se lleva a cabo un análisis del contexto en el que se va a desarrollar la herramienta objeto de este proyecto. En los dos primeros apartados se describen brevemente las diversas bases de datos y recursos de los cuales se ha obtenido información sobre enfermedades raras y proteínas. A continuación, se aborda el estudio de enfermedades mediante redes de relación y se exponen los sistemas de consulta actuales.

El estado del arte se concluye con un apartado que trata de justificar la aproximación al estudio de las relaciones entre enfermedades raras por medio de interacciones entre proteínas. Este apartado se enmarca dentro de una prueba de concepto, realizada como paso previo al presente trabajo, que fue necesario abordar debido a la novedad de la propuesta y la falta de referencias en la literatura. Cabe destacar que esta prueba de concepto es ya por sí misma una contribución propia al estado del arte en la materia.

2.1. Fuentes de información sobre enfermedades

La principal fuente de información sobre enfermedades de la cual se han obtenido los datos necesarios para la realización de este proyecto es **Orphadata**, un extenso conjunto de datos sobre enfermedades raras. Además, con el fin de explorar al máximo la información referente a enfermedades y sus bases moleculares, en la herramienta final se muestran además datos extraídos de ICD-10 y se proporcionan enlaces externos a los recursos de Orphanet, OMIM, MeSH y PubMed.

A continuación se detalla el contenido de cada fuente, explicando también el motivo de su importancia y que, en última instancia, justificará su incorporación a la herramienta desarrollada.

Figura 2.1: Fuentes de información sobre enfermedades.

2.1.1. ICD-10

ICD-10 es la décima revisión de la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud (CIE-10, en castellano), una lista de clasificación médica realizada por la Organización Mundial de la Salud. Contiene códigos de enfermedades, signos y síntomas, hallazgos anormales, denuncias, circunstancias sociales y causas externas de lesiones o enfermedades.

El conjunto de códigos ICD-10 provee más de 14.400 códigos diferentes y permite el seguimiento de multitud de nuevos diagnósticos. Además, los códigos se pueden ampliar a más de 16.000 códigos mediante el uso de sub-clasificaciones opcionales.

La clasificación de ICD-10 es útil en la herramienta para apoyar o sustentar las conexiones entre enfermedades. De esta forma, que dos enfermedades estén relacionadas y compartan la misma clasificación puede ser un buen indicador para "validar" esa conexión.

2.1.2. OMIM

El proyecto *Mendelian Inheritance in Man* es una base de datos que cataloga todas las enfermedades conocidas con un componente genético, y cuando es posible, la asociación a los genes en el genoma humano.

Años atrás la información estaba disponible en formato de libro, cuyo título era *Inheritance in Man* (MIM). Actualmente la base de datos está disponible de forma telemática en la web oficial[5] o en la web del Centro Nacional para la información biotecnológica, NCBI, y se actualiza prácticamente a diario. Esta versión en línea se denomina *Online Mendelian Inheritance in Man* (OMIM).

Se incluirán en la herramienta enlaces externos a las referencias OMIM sobre enfermedades y genes debido a la importancia de la información genética en este campo de estudio.

2.1.3. Orphanet

Orphanet[3] es el portal de información de referencia en enfermedades raras y medicamentos huérfanos, dirigido a todos los públicos. El objetivo de Orphanet es contribuir a la mejora del diagnóstico, cuidado y tratamiento de los pacientes con enfermedades raras.

Orphanet ofrece libre acceso a los siguientes servicios:

- Un listado de enfermedades raras y la clasificación de éstas elaborada a partir de las clasificaciones publicadas por expertos.
- Una enciclopedia de enfermedades raras.
- Un listado de medicamentos huérfanos en todas las etapas de desarrollo.
- Un directorio de recursos especializados que ofrece información sobre: centros expertos, laboratorios clínicos, proyectos de investigación en curso, ensayos clínicos, registros, redes, plataformas tecnológicas y asociaciones de pacientes; en el ámbito de las enfermedades raras y en cada uno de los países del consorcio Orphanet.

- Una herramienta de ayuda al diagnóstico, que permite a los usuarios buscar una enfermedad por los signos y los síntomas asociados.
- Una enciclopedia de recomendaciones y directrices para la atención médica de emergencia y la anestesia.
- Un boletín de noticias bimensual, *OrphaNews*, que ofrece una visión general sobre la actualidad científica y política en el ámbito de las enfermedades raras y los medicamentos huérfanos.
- Una colección de informes temáticos, los *Informes Periódicos de Orphanet*, que tratan temas relevantes y que pueden descargarse directamente de la web.

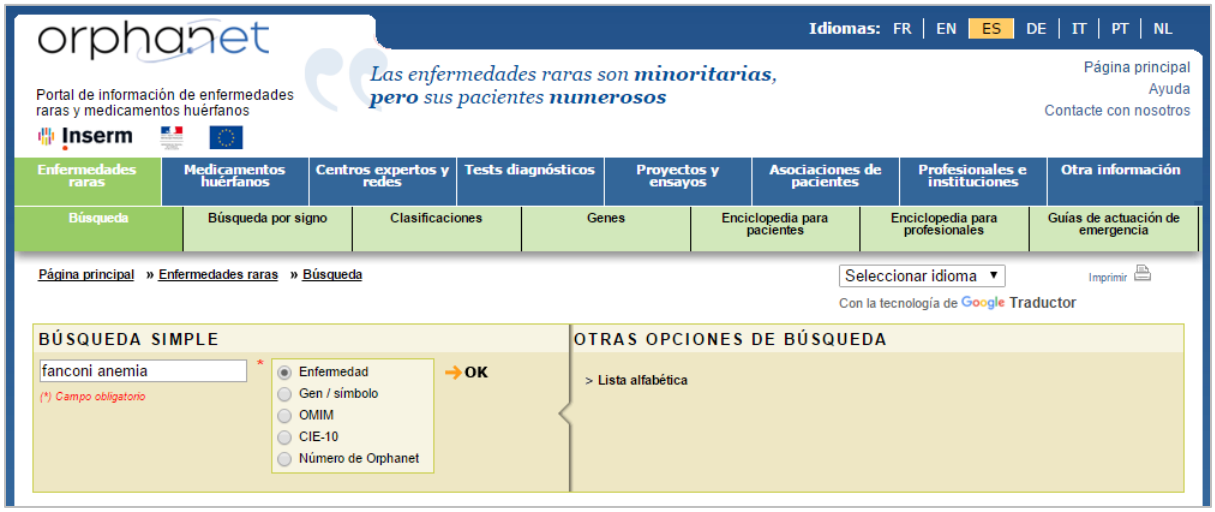

Figura 2.2: Portal de información de enfermedades raras Orphanet.

Orphanet está formado por un consorcio de alrededor de 40 países, coordinado por el equipo francés del INSERM. Los equipos nacionales se encargan de recopilar la información relacionada con las consultas especializadas, laboratorios médicos, investigación en curso y asociaciones de pacientes en su país. Además, Orphanet está dirigido por diferentes comités, que de forma independiente, supervisan el proyecto a fin de garantizar su coherencia, evolución y fiabilidad.

La información de Orphanet resulta imprescindible en cualquier sistema que trate las enfermedades raras. En la herramienta desarrollada se facilitarán por tanto enlaces externos a este recurso para todas las enfermedades.

2.1.4. Orphadata

La misión de *Orphadata*[4] es proporcionar a la comunidad científica de un conjunto de datos o “dataset” de libre acceso, exhaustivo y de alta calidad relacionados con las enfermedades raras y los medicamentos huérfanos, en un formato reutilizable.

El conjunto de datos es una extracción parcial de los datos almacenados en Orphanet, que también es accesible sólo con fines de consulta y de acceso gratuito.

Figura 2.3: Creación del dataset de Orphadata a partir de la base de datos de Orphanet.

El dataset de Orphadata, publicado en formato XML, incluye:

- Un inventario de enfermedades raras, con referencias cruzadas de OMIM e ICD-10, y sus genes asociados con referencias de otras fuentes externas.
- Una clasificación de enfermedades poco comunes establecida por Orphanet, basada en clasificaciones publicadas por expertos.
- Los datos epidemiológicos relacionados con las enfermedades raras en Europa (prevalencia, edad media de aparición, edad promedio de defunción) extraídos de la literatura.
- Una lista de los signos y síntomas asociados con cada enfermedad y su frecuencia.
- La lista de los signos y síntomas de Orphanet utilizados para anotar las enfermedades, con referencias cruzadas con otras nomenclaturas.

El conjunto de datos de Orphadata constituye la base a partir de la cual se elabora la herramienta de este proyecto. Entre toda la información disponible, cabe destacar la importancia de las asociaciones de las enfermedades raras con sus genes de susceptibilidad, así como las referencias cruzadas que permiten relacionar toda la información contenida en las diferentes fuentes.

2.1.5. MeSH

MeSH (del inglés, *Medical Subject Headings*, Encabezados de Temas Médicos), es el término empleado para describir un amplio vocabulario terminológico controlado para publicaciones de artículos y libros de ciencia.

El MeSH contiene alrededor de 25.000 títulos de material, también conocidos como descriptores, la mayor parte de los cuales se acompañaban por una breve descripción o definición, enlaces a los descriptores relacionados y una lista de sinónimos o términos muy similares.

Los enlaces externos a MeSH incluidos en la herramienta facilitarán la búsqueda de enfermedades raras en la literatura científica.

2.1.6. PubMed

PubMed es un motor de búsqueda de libre acceso a la base de datos MEDLINE de citaciones y resúmenes de artículos de investigación biomédica. Es ofrecido por la Biblioteca Nacional de Medicina de los Estados Unidos y tiene alrededor de 4.800 revistas publicadas en Estados Unidos y en más de 70 países de todo el mundo desde 1966 hasta la actualidad.

La herramienta permitirá buscar referencias conjuntas de enfermedades y genes en la literatura. Esto ayudará a sustentar o refutar conexiones entre enfermedades raras.

2.2. Fuentes de información sobre proteínas y sus relaciones moleculares

En este apartado se describen los recursos públicos de los cuales se han extraídos los datos relativos a proteínas y relaciones moleculares, HIPPIE y UniProt, y el motivo de su incorporación en el sistema.

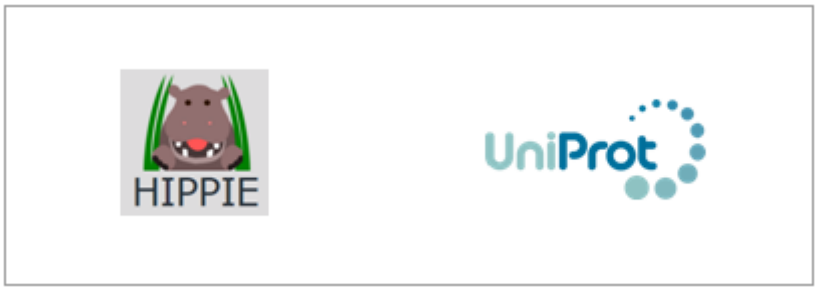

Figura 2.4: Fuentes de información sobre proteínas.

2.2.1. HIPPIE

HIPPIE[5] (del inglés, *Human Integrated Protein-Protein Interaction rEference*) es un conjunto de datos de interacciones entre proteínas humanas con un sistema de puntuación normalizada que integra múltiples conjuntos de datos experimentales de interacciones.

El componente central de HIPPIE es la puntuación de confianza de las interacciones basadas en la cantidad y la fiabilidad de las pruebas correspondientes a cada interacción. Esta puntuación se calcula como una suma ponderada del número de estudios en los que se detectó una interacción, el número y la calidad de las técnicas experimentales utilizadas para medir dicha interacción y el número de organismos no humanos en los que se ha reproducido la interacción. Los parámetros de este sistema de puntuación han sido optimizados conjuntamente por un grupo de expertos y un algoritmo informático con la intención de que dicha puntuación de calidad experimental refleje la fiabilidad y la tasa de error de las técnicas.

El dataset de HIPPIE, presentado en formato tabular, integra datos de interacciones de diez bases de datos y once estudios diferentes, conteniendo de esta manera más de 190.000 interacciones entre proteínas.

interactor - UniProt id	interactor - Entrez gene id	interactor - gene symbol	score (click on a score value to see the evidence)	Interacting proteins
SNF5_HUMAN	6598	SMARCB1	0.92	Show
PR40A_HUMAN	55660	PRPF40A	0.90	Show
SMCA2_HUMAN	6595	SMARCA2	0.89	Show
HDAC1_HUMAN	3065	HDAC1	0.89	Show
SIN3A_HUMAN	25942	SIN3A	0.89	Show
H32_HUMAN	126961, 333932, 653604	HIST2H3C, HIST2H3A, HIST2H3D	0.88	Show
SPI1_HUMAN	6688	SPI1	0.88	Show
NCOR1_HUMAN	9611	NCOR1	0.86	Show
CBX5_HUMAN	23468	CBX5	0.82	Show

Figura 2.5: Resultados de la búsqueda de la proteína MECP2 en HIPPIE.

Como se explicará más adelante, incorporar el conjunto de datos de HIPPIE es fundamental para establecer nuevas conexiones entre enfermedades raras gracias a interacciones proteína-proteína, utilizando sus puntuaciones como referencia.

2.2.2. UniProt

UniProt es el recurso de proteínas universal, un repositorio central de datos sobre proteínas creado por la combinación de otras bases de datos como Swiss-Prot, TrEMBL y PIR.

UniProt es una amplia base de datos, de alta calidad y de libre acceso, líder mundial sobre la secuencia de proteínas e información funcional. Muchas de sus entradas se derivan de los proyectos de secuenciación del genoma y contiene una gran cantidad de información acerca de la función biológica de las proteínas derivadas de la literatura de investigación.

La herramienta proporcionará enlaces externos a este recurso con la intención de ampliar la información sobre proteínas contenidas en el sistema.

2.3. Estudio de enfermedades mediante redes de relación

Existen diferentes estrategias para establecer relaciones entre enfermedades. Algunas de ellas se basan en diferentes tipos de datos tales como genes compartidos[2,7], microRNAs compartidos[8], vínculos funcionales[9], localización de proteínas[10], interacciones proteína-proteína[11], reacciones metabólicas consecutivas[12], fenotipos y síntomas comunes[13,14] o asociaciones de comorbilidad[15,16].

Establecer relaciones entre enfermedades a partir de genes de susceptibilidad compartidos entre ellas es una forma de estudio global de enfermedades, que además permite crear redes de relación cuyos nodos identifican enfermedades y los enlaces entre ellas se construyen en base a genes de susceptibilidad compartidos.

Siguiendo esta estrategia se han construido varias redes de enfermedades humanas que presentan una visión global de la mayoría de las patologías y trastornos conocidos haciendo referencia a sus características genéticas. Un ejemplo de estas redes son *Diseasome*[7] y *Orphan Diseasome*[2], en las cuales se han establecido relaciones por genes compartidos entre enfermedades comunes y raras, respectivamente. Sin embargo, existe una diferencia fundamental al comparar ambas redes: en *Orphan Diseasome* se observa un alto número de enfermedades aisladas en pequeños grupos sin conexión con el resto de la red (Figura 2.6). Esto se debe a que las enfermedades raras tienen menos genes asociados que las comunes, la mayoría de ellas solamente uno, lo que dificulta el establecimiento de relaciones entre ellas mediante genes compartidos.

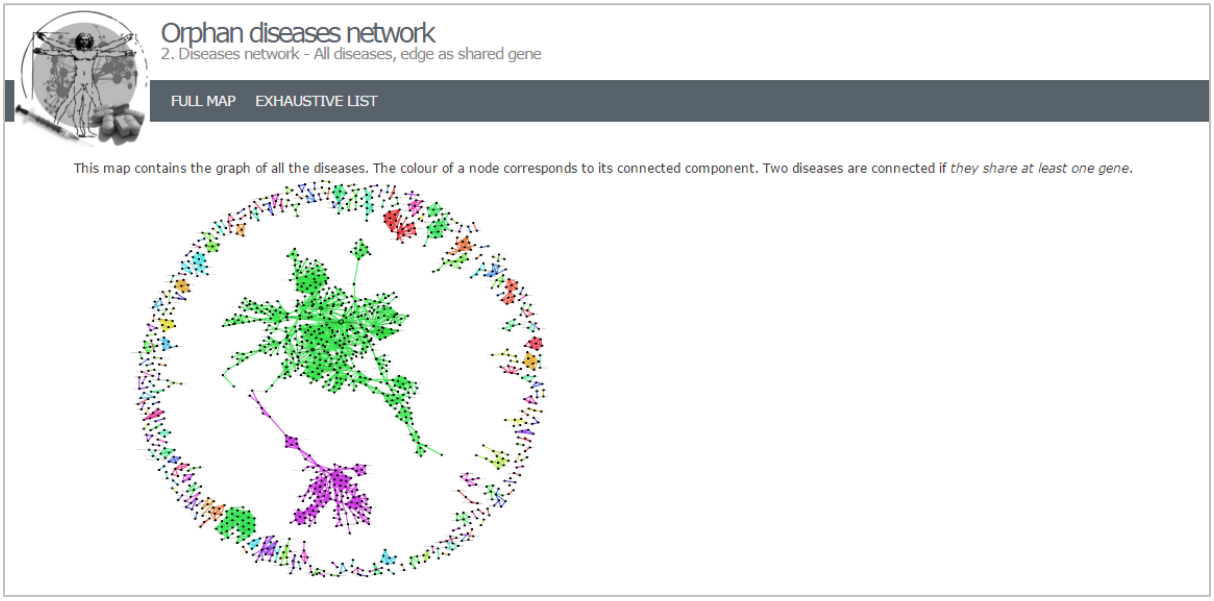

Figura 2.6: *Orphan Diseasome*, Red de Enfermedades Huérfanas. En esta imagen se puede observar cómo multitud de enfermedades se encuentran aisladas alrededor de la red principal sin conexión con ella.

Con la intención de estudiar las enfermedades huérfanas de un manera global Zhang et al.[2] construyeron y estudiaron varias redes de asociación de enfermedades raras. En la red más poblada, dos enfermedades estaban conectados si compartían al menos un gen de susceptibilidad, creando y analizando en ese momento 2.259 conexiones entre 1.170 enfermedades. A pesar del valor de esta red para el estudio global de las enfermedades raras, sugerían que muchas relaciones entre las enfermedades raras no pueden ser descubiertas basándose únicamente en genes compartidos. Para superar esta limitación, construyeron y analizaron además otras redes para el subgrupo de enfermedades con al menos cuatro genes de susceptibilidad en base a las anotaciones enriquecidas que compartían (procesos biológicos, componentes celulares, fenotipos y rutas). No obstante, estas redes representaban sólo una pequeña fracción de las enfermedades de la red basada en genes (aproximadamente un 15%), ya que la mayoría de las enfermedades raras son monogénicas.

Teniendo en cuenta esta limitación, surge la necesidad de buscar una estrategia alternativa que permita establecer un mayor número de relaciones entre enfermedades raras.

2.4. Sistemas públicos de consulta sobre relaciones entre enfermedades

A pesar de su valor para una amplia gama de investigadores y clínicos biomédicos, hay muy pocos recursos públicos disponibles que presenten relaciones potenciales entre enfermedades para ser consultadas por los usuarios a través de interfaces gráficas “amigables”. Entre las pocas excepciones se encuentran MalaCards, que construye relaciones basadas en una variedad de información (como genes, vías, fenotipos, compuestos y términos de *Gene Ontology*), y DiseaseConnect, que establece conexiones por genes compartidos y datos de expresión diferencial.

A continuación se muestran en detalle los recursos públicos mencionados, los cuales presentan la línea de trabajo en la cual se desarrolla la herramienta objeto de este proyecto.

2.4.1. MalaCards: La base de datos de enfermedades humanas

MalaCards[17] es una base de datos integrada de las enfermedades humanas y sus anotaciones, modelada a partir de la arquitectura y el contenido de la popular base de datos *GeneCards* sobre los genes humanos.

Esta base de datos de enfermedades y trastornos se organiza mediante "tarjetas de enfermedad", las cuales integran información priorizada y una lista de numerosos sinónimos o alias conocidos para cada enfermedad, junto con una variedad de anotaciones, así como las conexiones entre enfermedades, fundamentadas por la base de datos relacional *GeneCards* y el análisis de los sets de *GeneAnalytics*. Las anotaciones incluyen: síntomas, medicamentos, artículos, genes, ensayos clínicos, enfermedades o trastornos relacionados y mucho más. Un motor automático de recuperación de información rellena las fichas de enfermedades, a partir de datos remotos, así como la información obtenida mediante la plataforma GeneCards para compilar la base de datos de la enfermedad. La base de datos de MalaCards integra ambas listas de enfermedades, general y especializada, incluyendo enfermedades raras, enfermedades genéticas, trastornos complejos y más.

Las secciones de MalaCards se construyen gracias a:

- La consulta directa a los recursos de la enfermedad, para establecer nombres de enfermedades, sus sinónimos, resúmenes, medicamentos, terapias, tratamientos, características clínicas, pruebas genéticas y contexto anatómico (Figura 2.7).
- La búsqueda en GeneCards para publicaciones relacionadas y genes asociados.
- El análisis del conjunto de genes asociados a enfermedades en GeneAnalytics para producir vías asociadas, fenotipos y compuestos.

- La búsqueda dentro de MalaCards en sí, por ejemplo, para enfermedades o trastornos relacionados adicionales.

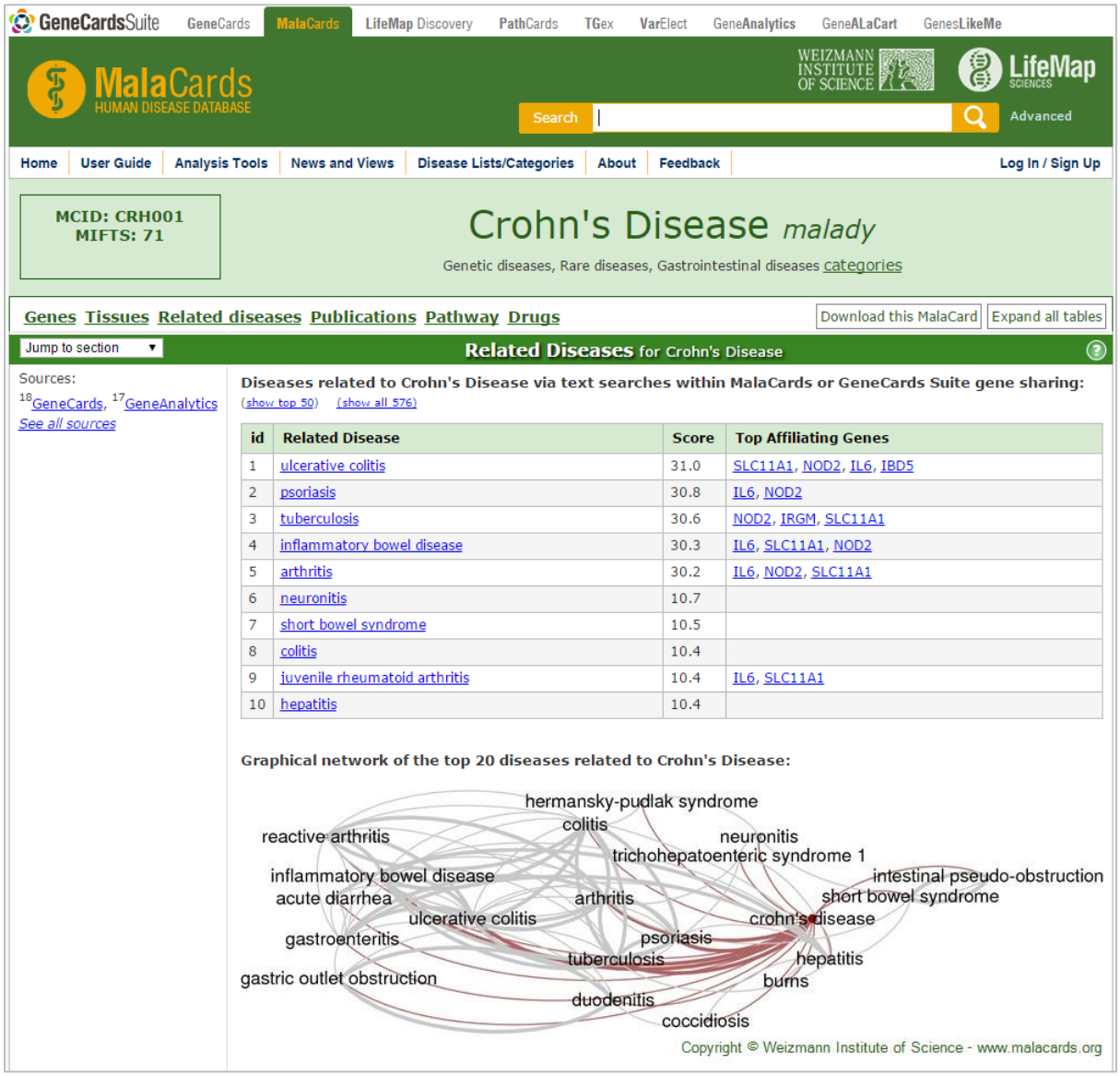

Figura 2.7: Resultados de la búsqueda de la enfermedad de Crohn en MalaCards.

En la actualidad, la base de datos contiene 18.864 entradas de enfermedades, consolidadas a partir de 64 fuentes.

2.4.2. DiseaseConnect

DiseaseConnect[18] es un servidor web para el análisis de enfermedades y visualización de redes basadas en mecanismos moleculares.

La interfaz web de DiseaseConnect (Figura 2.8) incluye diferentes características para hacer la exploración fácil para el usuario:

- Construye redes integrales que describen la conectividad enfermedad-enfermedad, las asociaciones enfermedad-gen y los tratamientos.
- Integra listas detalladas y representaciones de las relaciones enfermedad-gen derivadas de diversas fuentes externas.
- Proporciona herramientas de visualización de red con diferentes opciones de diseño de la red, así como nodos y enlaces personalizables; además permite al usuario hacer zoom y arrastrar el diagrama de red.
- Autocompleta el campo de búsqueda con los nombres completos de las enfermedades y los símbolos de los genes.
- Exporta la red como un archivo png, svg o xgmml.

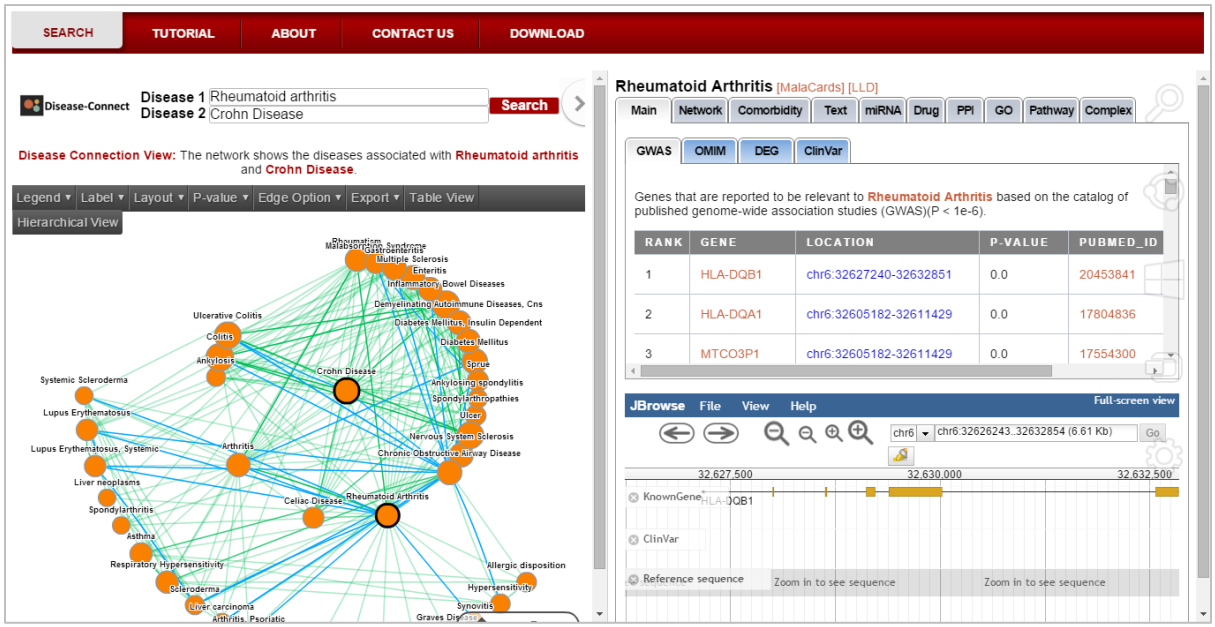

Figura 2.8: Resultados de la búsqueda de la conexión entre la enfermedad de Crohn y la artritis reumatoide en DiseaseConnect.

El servidor web DiseaseConnect está implementado utilizando JSP, MySQL, JavaScript y una tecnología avanzada de visualización interactiva de redes denominada Cytoscape Web[19].

Tras analizar las redes de relación entre enfermedades raras y los sistemas públicos de consulta, se comprueba que en la actualidad no existen herramientas públicas que permitan la búsqueda de relaciones entre enfermedades raras de una manera especializada.

2.5. Nueva estrategia para la relación de enfermedades raras: interacciones proteína-proteína

Como se ha expuesto anteriormente, en el estudio de las enfermedades raras se han seguido diferentes estrategias para establecer relaciones entre ellas[2,7,8,9,10,11,12,13,14,15,16]. El más sencillo conceptualmente, y con el que se han relacionado hasta ahora el mayor número de afecciones, consiste en establecer relaciones a partir de los genes de susceptibilidad que comparten.

Sin embargo, siguiendo el flujo de la información biológica, pueden establecerse además otro tipo de relaciones. Las mutaciones en el genoma asociadas a una determinada enfermedad genética hereditaria se manifiestan en el organismo según el flujo de la información genética que codifican. Así las mutaciones en regiones codificantes de los genes pueden manifestarse como alteraciones de la secuencia de aminoácidos de las proteínas que codifican (a través del flujo de información desde la secuencia de DNA que se transcribe a RNA mensajero y finalmente se traduce a secuencia proteica) (Figura 2.9).

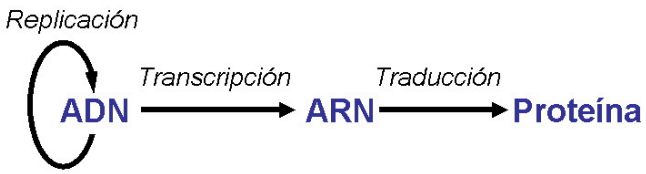

Figura 2.9: Dogma central de la biología molecular. Los genes en el DNA codifican la secuencia de las proteínas que son las que llevan a cabo la función celular.

A su vez son las proteínas, a través de sus acciones coordinadas entre ellas y otras entidades moleculares (como los metabolitos o los ácidos nucléicos como el DNA o el RNA), las que realizan y coordinan las funciones celulares que dan sustento a la vida. Entre estas acciones coordinadas cabe destacar las interacciones físicas que se establecen entre proteínas, ya sea de manera permanente estableciendo complejos macromoleculares estables (como el proteosoma o el ribosoma) o de manera transitoria, como en la mayor parte de las rutas de señalización (como las interacciones entre quinasas y sus proteínas sustrato). El conjunto de todas la interacciones entre proteínas humanas constituye el interactoma humano.

Por lo tanto, una aproximación alternativa para establecer relaciones potenciales a nivel molecular entre un mayor número de enfermedades raras sería uniendo dos enfermedades no sólo porque comparten genes, sino porque los productos de sus genes asociados (las proteínas que codifican) interaccionan también. Esta estrategia se ha utilizado anteriormente para estudiar la topología y la función de la red global de enfermedades

humanas[11]. Está sustentada además por estudios previos que han reportado una mayor similitud tanto de síntomas como de comorbilidades entre las enfermedades asociadas a proteínas que interaccionan, que entre aquellas asociadas a proteínas que no interaccionan[14,20]. De esta manera, aunque el interactoma humano está aún incompleto, su cobertura actual permite el estudio de los mecanismos subyacentes a las enfermedades y las relaciones entre enfermedades a nivel sistémico[21,22].

3. Diseño y desarrollo

En este capítulo se describe el proceso que se ha seguido para la implementación de la solución desarrollada durante el proyecto. En primer lugar se realiza una descripción de alto nivel del diseño del sistema incluyendo la arquitectura propuesta. En este apartado se define la partición en módulos del sistema y se establece su función específica. Posteriormente se profundiza en los detalles del desarrollo de la herramienta y los módulos generados.

3.1. Diseño

Antes de comenzar con la implementación de la herramienta, en la etapa de diseño, es importante definir los siguientes puntos:

- Arquitectura del sistema y módulos en los que se divide.
- Requisitos o especificaciones de la herramienta a nivel funcional.
- Tecnologías seleccionadas para cumplir con las especificaciones en cada uno de los módulos.

En este apartado se detalla cada uno de ellos.

3.1.1. Arquitectura del sistema

Para el diseño de la herramienta se ha definido una arquitectura de tres niveles. Partiendo de la arquitectura cliente-servidor, el sistema se ha dividido en tres capas o niveles con un reparto claro de funciones: una capa para la presentación (interfaz de usuario), otra para el cálculo (donde se encuentra modelado el negocio) y otra para el almacenamiento (persistencia) (Figura 3.1).

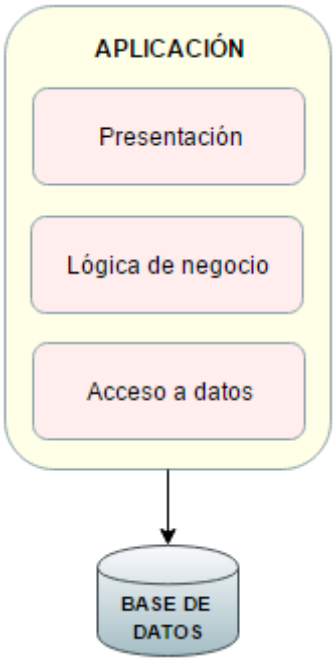

Figura 3.1: Diseño de la arquitectura de la herramienta.

La arquitectura se ha definido de forma que cada capa sólo tenga relación con sus contiguas. Esto permite que cambios en una sección no afecten a todo el sistema, y además, dota al sistema de mayor flexibilidad, favorece la reutilización y disminuye la complejidad.

La capa de **presentación**, también conocida como interfaz gráfica o UI (*User Interface*), presenta el sistema al usuario final e interactúa con él: le comunica la información y captura la información de éste proporcionando parámetros de entrada (realizando un filtrado previo para comprobar que no hay errores de formato) y recibiendo datos como respuesta. Esta capa se comunica únicamente con la capa de negocio y debe tener la característica de ser amigable para el usuario, es decir, entendible y fácil de usar.

La **lógica de negocio** es la parte del sistema que se encarga de codificar las reglas de negocio del mundo real (operaciones, definiciones y restricciones) que determinan cómo la información puede ser creada, mostrada y cambiada, recibiendo las peticiones del usuario y enviando las respuestas tras el proceso. Es aquí donde se establecen todas las reglas que deben cumplirse, recibiendo las solicitudes y validando que las condiciones establecidas se cumplen antes de realizar acciones o de hacer la respectiva solicitud a la capa de acceso a datos. Esta capa se comunica con la capa de presentación, para recibir las solicitudes y presentar los resultados, y con la capa de datos, para solicitar al gestor de la base de datos almacenamiento o recuperación de información. La lógica de negocio debe ser fácil de implementar y testear reduciendo al máximo la complejidad de las operaciones.

Por último, el **acceso a datos** es la capa encargada de la comunicación con la base de datos y en ella descansan las acciones CRUD (del inglés, *Create Read Update Delete*). Está formada por un gestor de bases de datos que recibe solicitudes de almacenamiento o recuperación de información desde la capa de negocio. El acceso a datos debe ser rápido y eficiente para garantizar la calidad de la herramienta.

3.1.2. Requisitos

La arquitectura propuesta define los módulos de persistencia (acceso a datos), negocio y presentación. En este punto se asignan los diferentes requisitos de cada uno de ellos estableciendo así las especificaciones de cada módulo.

- **Persistencia**: Una base de datos relacional que debe integrar la información sobre enfermedades raras e interacciones entre proteínas extraída de diferentes fuentes y recursos. Todas las conexiones entre las enfermedades raras, que han de ser calculadas previamente, han de estar contenidas en la base de datos también. Además, se debe garantizar la exactitud e integridad de la información reduciendo los datos redundantes o innecesarios.

- **Negocio**: La lógica de negocio y cálculos necesarios para acceder a la información presente en el sistema y crear la red resultante de enfermedades y genes. Enfocado a garantizar la interoperabilidad, este módulo debe ser capaz de acceder a la base

de datos, extraer la información, establecer las relaciones oportunas y generar los resultados de manera eficiente. El acceso a base de datos y los cálculos se realizarán cada vez que una petición de búsqueda de enfermedad, conexión o gen se realice desde la interfaz de usuario.

- **Presentación**: Una herramienta pública, y por tanto con interfaz web, que permita realizar consultas y muestre la información y la red resultantes de manera clara y sencilla (Figura 3.2). Para ello el diseño debe ser minimalista e intuitivo, enfocado a resaltar la información más relevante. La página principal deberá permitir la búsqueda tanto de enfermedades y conexiones entre ellas como de genes, y la página de resultados de la búsqueda deberá contener los siguientes bloques:

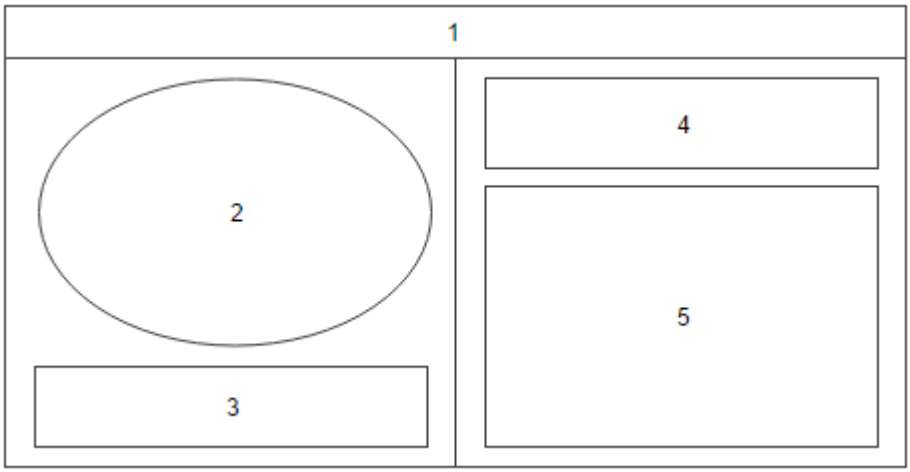

Figura 3.2: Esquema (*wireframe*) de la página de resultados de la interfaz web.

(1) barra de navegación que permita el acceso a otras páginas del sitio,
(2) visualización de la red resultante,
(3) resultados de la interacción del usuario con la red,
(4) panel resumen del ítem consultado,
(5) tablas o informes de resultados clasificados en pestañas.

El sistema está orientado a la comunidad científica internacional, por lo que los datos y la presentación de los mismos se realiza en inglés.

Además, teniendo en cuenta que la herramienta está enfocada a mostrar información sobre enfermedades raras y establecer nuevas relaciones entre ellas, se decide el nombre de la aplicación ***Orphan Disease Connections*** (Conexiones de Enfermedades Huérfanas), o en su abreviatura, **ODCs**.

3.1.3. Tecnologías utilizadas

Una vez definida la arquitectura y los módulos que se van a desarrollar, se procede a hacer una selección de tecnologías que permitan el almacenamiento, el tratamiento y la consulta de los datos presentes en la herramienta (Tabla 3.1).

Tabla 3.1: Tecnologías elegidas para la implementación de la herramienta agrupadas por módulo de desarrollo.

Persistencia	Lógica de negocio y tratamiento de datos	Interfaz de usuario
MySQL	**PHP** **Python**	**HTML** **JavaScript** **jQuery** **CSS** **Bootstrap** **Cytoscape Web**

Un requisito común a todas las tecnologías, y que se debe remarcar, es que se exigen dos características fundamentales para la elección de lenguajes de programación y herramientas: que se trate de software libre y que sean compatibles con el sistema operativo Linux[1] utilizado lo largo de todo el proyecto.

3.1.3.1. Persistencia

Teniendo en cuenta el volumen y el tipo de información que se desea almacenar y consultar en la herramienta, se elige una base de datos relacional SQL. Entre las alternativas más populares se encuentran PostgreSQL y MySQL. Se decide que la gestión de la base se realiza con **MySQL** por la rapidez en lectura, el buen rendimiento y la facilidad de instalación y configuración.

La extracción de la información necesaria y relevante de las diferentes fuentes de datos se realiza mediante el lenguaje de programación **Python**. Se elige esta tecnología por su sencilla sintaxis y la rapidez de desarrollo. Otros lenguajes como Java requieren una configuración del entorno, la curva de aprendizaje es mayor, y en este caso, resultan más complejos que el problema que deben resolver.

[1] Se toma la distribución de referencia Ubuntu.

3.1.3.2. Lógica de negocio

La utilización de MySQL como gestor de la base de datos conlleva la elección de **PHP** como lenguaje de programación para el modelado de negocio. Su facilidad y rapidez en el acceso a la base datos y su baja curva de aprendizaje lo hacen más apropiado en este caso para el desarrollo de la aplicación que lenguajes como Java, Python o Ruby.

3.1.3.3. Interfaz de usuario

Para el desarrollo de la interfaz de usuario se utilizan como base las tecnologías estándares para que una página web funcione en los navegadores actuales:

- **HTML**, lenguaje de programación basado en etiquetas para la estructura básica del contenido de la página.
- **CSS**, hojas de estilo para la presentación y el diseño.
- **JavaScript**, lenguaje orientado a objetos que ayuda a mejorar la lógica y el dinamismo de la página.

Además, para trabajar con JavaScript se escoge **jQuery**, librería que simplifica la escritura de código mediante funciones predefinidas.

Por otra parte, para facilitar la generación de estilos se elige **Bootstrap**[23], un popular y potente conjunto de herramientas (*framework*) que contiene plantillas de diseño fáciles de utilizar y con una estética actual. Permite además generar formularios, botones, tablas, menús y demás elementos ajustando dinámicamente el diseño gráfico de la página según el tipo de dispositivo usado y su tamaño.

Por último, como tecnología de visualización de redes en la interfaz se utiliza **Cytoscape Web**[19], software que, gracias a un código JavaScript relativamente fácil de implementar, permite la visualización y el análisis de la red. Cytoscape Web es un *plugin* de Cytoscape[24], programa bioinformático usado también en este proyecto para la construcción de las redes resultantes de enfermedades raras y genes.

3.2. Desarrollo

El desarrollo de la herramienta comienza con el diseño e implementación de la base de datos que contiene toda la información sobre enfermedades raras y proteínas disponible en las citadas fuentes. A continuación, se lleva a cabo la construcción de diferentes redes de enfermedades raras y genes, gracias a los datos extraídos. Para concluir este proyecto se desarrolla la aplicación web que permite la consulta y visualización de las enfermedades y sus relaciones de manera gráfica, amigable y pública.

3.2.1. Modelo de datos

El almacenamiento de información requiere un primer paso de diseño del modelo de datos. El diagrama Entidad-Relación representado en la Figura 3.3 refleja la estructura de la información en la base de datos de ODCs. En el diagrama se pueden observar las siguientes entidades:

- **GEN**
  Un gen tiene dos atributos: el símbolo, que lo define unívocamente, y las referencias externas, códigos que permiten enlazar cada gen con información presente en OMIM (catálogo de genes humanos y trastornos genéticos) y UniProt (recurso de proteínas universal).

- **ENFERMEDAD**
  Cada enfermedad tiene como atributos dos números identificativos: el número de Orphanet (identificador unívoco del trastorno) y otro id necesario para enlazar con su clasificación. Además, se añaden otros atributos como el nombre de la enfermedad, sinónimos, referencias externas de OMIM y MeSH (vocabulario médico) y clasificación de ICD10.

- **SÍNTOMA**[2]
  Un síntoma tiene sólo un campo de texto que contiene el nombre completo del síntoma clínico.

- **DATOS EPIDEMIOLÓGICOS Y HEREDITARIOS**
  Este conjunto de datos contiene información relativa a la prevalencia de la enfermedad, periodos de edad de aparición de la enfermedad y de defunción de los enfermos afectados y el tipo de herencia.

[2] No se añaden enlaces externos en síntomas para evitar que la herramienta se convierta en una fuente para el diagnóstico de enfermedades.

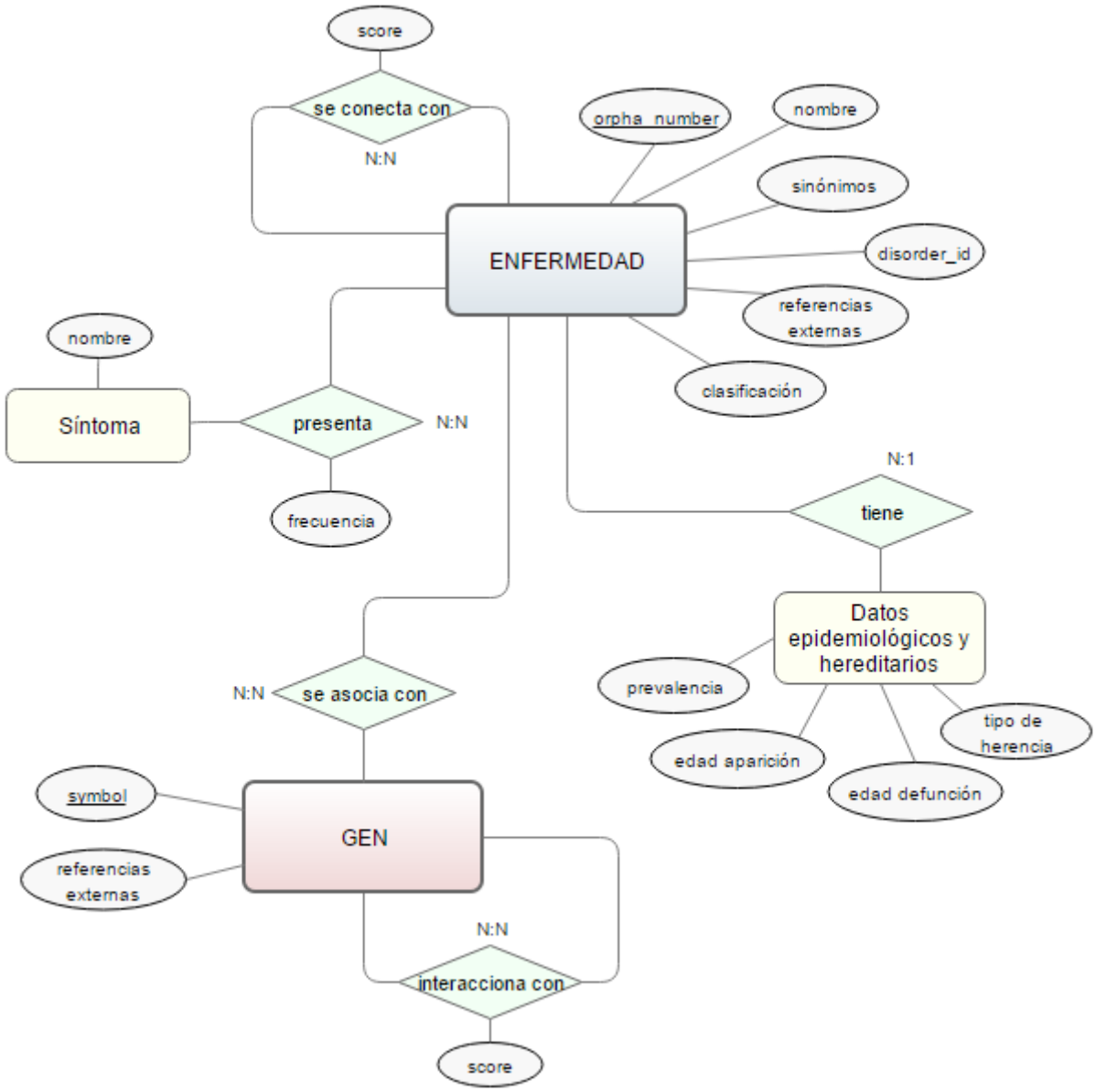

Figura 3.3: Diagrama entidad-relación de la base de datos de ODCs.

Esas cuatro entidades se relacionan entre ellas mediante cinco relaciones:

- La **asociación** entre enfermedad y gen se establece mediante una relación N:N, puesto que una enfermedad puede tener varios genes asociados, y a su vez, un gen puede estar asociado con más de una enfermedad.
  Esta relación es fundamental, ya que a partir de ella se establecerán conexiones entre enfermedades gracias a genes comunes entre ellas.

- Una enfermedad puede **presentar** varios síntomas, los cuales pueden ser comunes en otras enfermedades.

- Cada enfermedad **tiene** asociado además un único conjunto de datos epidemiológicos y hereditarios.

- Un gen puede **interaccionar** con uno o varios genes mediante interacciones proteína-proteína. Esta relación tiene como atributo el *score* o puntuación establecido por HIPPIE (base de datos de interacciones integrada en la herramienta), el cual determina la fiabilidad de la interacción.
  Esta relación permitirá establecer relaciones de segundo orden entre enfermedades a partir de los genes asociados a ellas.

- Por último, toda enfermedad incluida en la base de datos puede estar **conectada** con otra enfermedad, pudiendo estarlo con muchas de ellas. Se establece así mismo una puntuación que permitirá ordenar las enfermedades según la conexión sea más fuerte y fiable.
  Esta relación es muy importante, ya que constituye la base de este proyecto.

3.2.1.1. Interoperabilidad

Uno de los objetivos de este proyecto es garantizar el acceso a los diferentes recursos disponibles sobre enfermedades raras y genes. La persistencia de los datos, es decir, la permanencia de ellos en memoria, es fundamental para el desarrollo de la herramienta.

La información extraída de Orphadata y HIPPIE constituye el grueso de la base de datos. Además, en ella también se integran los códigos necesarios para que la herramienta proporcione enlaces externos a ICD-10, Orphanet, OMIM, UniProt y MeSH (Figura 3.4). Se establecen por tanto las relaciones en la base de datos entre los códigos de las diferentes fuentes y los identificadores de las enfermedades y los genes.

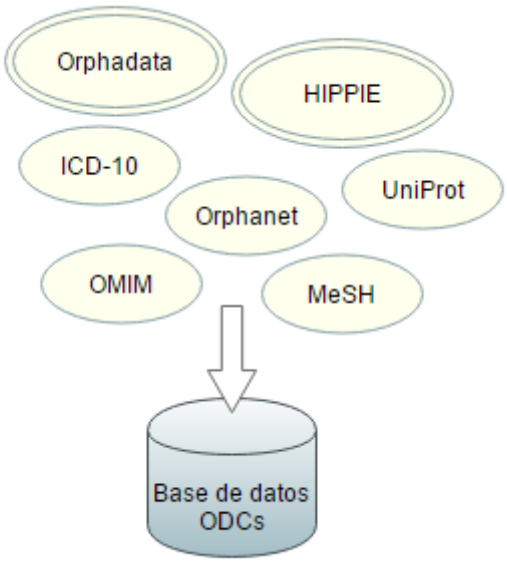

Figura 3.4: Integración de información de las distintas fuentes en la base de datos de la herramienta.

Por su parte, el acceso a las publicaciones de PubMed no se materializa en la base de datos, sino que se realizará bajo demanda en la capa de presentación mediante la construcción dinámica de búsquedas remotas en la interfaz de usuario. Para ello, la aplicación permitirá la selección de enfermedades y genes mediante casillas de verificación (*checkboxes*) en los resultados de conexión entre dos enfermedades.

3.2.1.2. Limpieza e integración de los datos

Para crear la base de datos de la herramienta es necesario disponer de los datos correctos provenientes de las diferentes bases de datos y fuentes de información que se van a integrar en ella. Hay que tener en cuenta además que algunos conjuntos de datos contienen información redundante o innecesaria para los objetivos de este trabajo.

Atendiendo especialmente a la validez, la consistencia, la uniformidad y la unicidad de los datos, se procede a un proceso de análisis, limpieza de información y eliminación de duplicados. Para ello, y teniendo en cuenta los diferentes formatos de los datos, se crean varios programas sencillos que generan ficheros listos para importar en ODCs.

Orphadata contiene mucha información en diferentes archivos. Partiendo del modelo de datos diseñado, e intentando a su vez respetar los ficheros y la estructura del dataset original, se divide la información en tablas basadas en temas, como las representadas en la Figura 3.5. Esta práctica ayuda a clarificar el contenido y reducir los datos redundantes. Este dataset incluye además la información de otras fuentes como Orphanet, OMIM y MeSH.

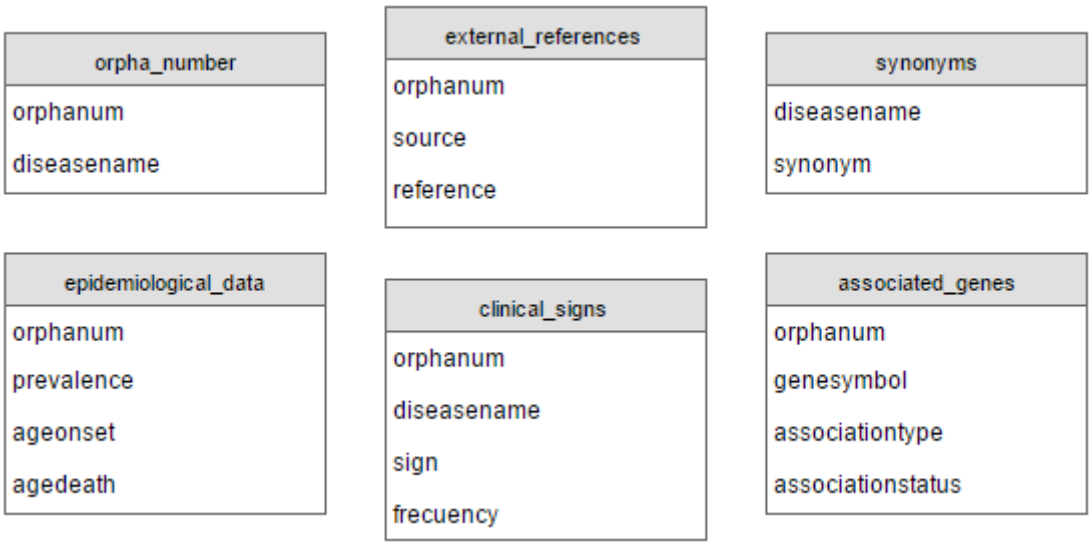

Figura 3.5: Ejemplo real de algunas tablas de ODCs para el dataset de Orphadata.

Por otra parte, se crea una tabla relativa a la clasificación de las enfermedades raras según la OMS (ICD-10). Esta tabla junto a las anteriores contienen la totalidad de datos relativos a enfermedades y sus asociaciones con genes.

Los datos relativos a interacciones entre proteínas contenidos en HIPPIE se almacenan en una tabla diferente. Además, se añade una tabla con el "parámetro q" asociado a cada interacción: valor necesario para la creación de URLs en los enlaces externos a las interacciones dentro de la herramienta. Por otra parte, surge la necesidad de crear una tabla que permita convertir los identificadores de gen de UniProt para compatibilizar la información de gen en Orphadata y en HIPPIE. Esta conversión de identificadores es esencial para poder establecer aquellas relaciones entre enfermedades (Orphadata) basadas en interacciones de proteínas (HIPPIE).

Se genera por último una tabla derivada de los datos de otras tablas, pero necesaria por eficiencia de la herramienta al ser de acceso frecuente, en la cual se incluyen los pares de enfermedades conectadas con el número de genes comunes entre ellas.

3.2.2. Redes de enfermedades y genes

A partir de la información materializada en base de datos se generan varias redes estáticas de enfermedades y genes, las cuales sirven como referencia para el desarrollo de la herramienta y permiten analizar el comportamiento de los datos así como visualizar las primeras conexiones entre enfermedades raras. Para este análisis se utiliza *Cytoscape*, programa que ayuda a la visualización de las redes y la interacción con nodos y aristas.

En una primera aproximación las redes creadas son las siguientes:

- Red de enfermedades y genes, con conexiones enfermedad-gen por asociación.
- Red de genes, relacionados por interacciones proteína-proteína.
- Red de enfermedades, relacionadas por genes comunes.

En esta última red se comprueba un elevado número de enfermedades conectadas, pero al mismo tiempo, se aprecia un gran número de enfermedades aisladas sin conexión con la red principal. Se procede por tanto a establecer nuevas conexiones entre enfermedades basándose en interacciones entre proteínas.

3.2.2.1. Sistema de puntuación de relaciones

Las relaciones entre enfermedades se establecen utilizando dos criterios:

1. **Genes comunes**: dos enfermedades están relacionadas si comparten al menos un gen.

2. **Interacciones proteína-proteína**: dos enfermedades están relacionadas si existe al menos una interacción entre las proteínas codificadas por sus genes.

Ambos tipos de relaciones tendrán una puntuación asociada, que permitirá el ordenamiento de las relaciones en la capa de presentación. La puntuación de las relaciones por genes comunes indica el número de genes que comparten. La puntuación de las relaciones por interacciones proteína-proteína deberá resumir en un único valor el número de interacciones y la fiabilidad de las mismas (*score*). Cuando la relación entre dos enfermedades se haya establecido por genes comunes y por interacciones proteína-proteína, el primer criterio será considerado como más relevante en la capa de presentación.

El cálculo de la puntuación asociada a un par de enfermedades a y b se realiza según la fórmula:

$$s_{ab} = \sum_{i \subset a, j \subset b} score_{ij}$$

donde $score_{ij}$ es el valor *score* asociado a la interacción entre las proteínas i y j asociadas a las enfermedades a y b.

Con esta nueva información se generan otras dos redes:

- Red de enfermedades relacionadas únicamente por interacción entre proteínas.
- Red global de enfermedades, relacionadas por genes comunes o por interacciones.

Esta última red requiere ser persistida para el establecimiento de la red global *offline* en el sistema de consultas. Para ello se añade una tabla con el total de conexiones entre enfermedades y se incluye en la base de datos de ODCs.

3.2.3. Sistema de consultas

Para hacer accesible toda la información sobre enfermedades raras y genes contenida en la base de datos de ODCs, así como la red para la visualización de las conexiones, se implementa un sistema de consultas.

Tal como se explica en el apartado de Diseño, el sistema cuenta con un frontal web que muestra los resultados en tablas o listados, y a su vez de forma gráfica, en una vista amigable, interactiva y accesible desde cualquier tipo de dispositivo.

Adquiere especial importancia la representación de los resultados de manera gráfica, ya que esto permite la visualización rápida y sencilla de las conexiones entre enfermedades y genes. Las conexiones que se pueden establecer son:

- ➔ enfermedad-gen: genes asociados a una enfermedad,

- ➔ gen-gen: por interacciones proteína-proteína,
- ➔ enfermedad-enfermedad: enfermedades conectadas por genes comunes o por interacciones entre sus genes.

A partir de ellas se construye una red *on-the-fly* (en tiempo real) dependiendo de la consulta hecha a la herramienta y a partir de la información contenida en la base de datos. Se genera un archivo en formato JSON que contiene los nodos de la red y las aristas que los unen, así como las etiquetas y los enlaces externos ligados a ellos. Estos datos serán posteriormente interpretados por el *plugin* de visualización de redes contenido en la herramienta.

El sistema de consultas distingue tres tipos de búsqueda: enfermedad, conexión entre dos enfermedades y gen, las cuales se explican en detalle a continuación.

3.2.3.1. Enfermedad

A partir de la consulta de una enfermedad rara (por nombre) se accede a una tabla donde se obtiene el identificador de Orphanet. Este número identificativo permitirá consultar todos los datos de interés de la enfermedad, sus genes asociados y las enfermedades conectadas con ella, tal como muestra el esquema representado en la Figura 3.6.

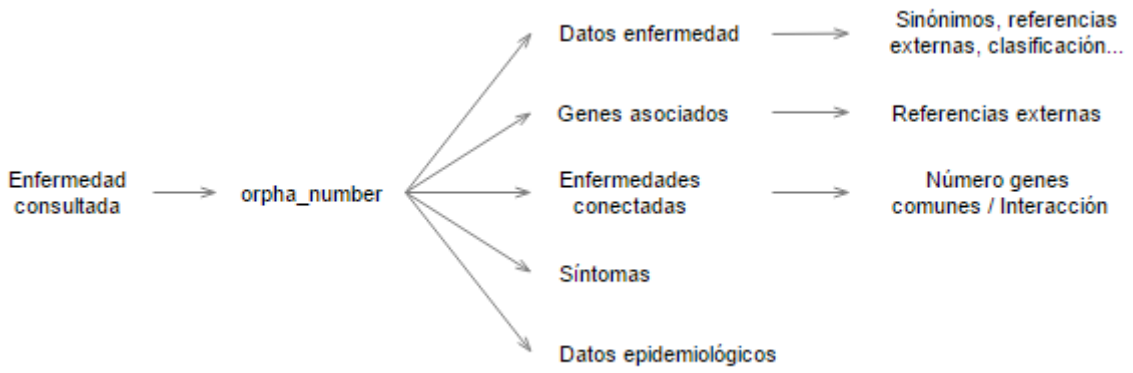

Figura 3.6: Búsqueda de enfermedad en la herramienta.

Con los códigos de referencias y clasificación se construyen los enlaces externos necesarios para la consulta en las otras fuentes de información.

Por otra parte, se ordena siempre la información según su relevancia. Por ejemplo, las enfermedades conectadas se listan ordenadas por mayor número de genes comunes, y después, las conectadas por interacciones entre proteínas. Los síntomas por su parte, aparecen ordenados por frecuencia. Las colecciones de datos que no tienen una relevancia asociada se muestran siempre por orden alfabético. Este criterio se mantiene en todas las vistas de la herramienta.

Por último, la red de enfermedades que se visualiza en la aplicación aparece centrada en la enfermedad consultada y a su alrededor las enfermedades conectadas con ella, diferenciando el tipo de conexión: genes comunes o interacción de proteínas.

3.2.3.2. Conexión entre dos enfermedades

En la herramienta se puede realizar también una consulta a partir de dos enfermedades. Esto permite comprobar si existe una conexión entre ellas, saber cómo se establece dicha conexión y elaborar comparaciones entre la información de ambas enfermedades. Para cada enfermedad se realizan las consultas necesarias siguiendo el esquema de la Figura 3.7.

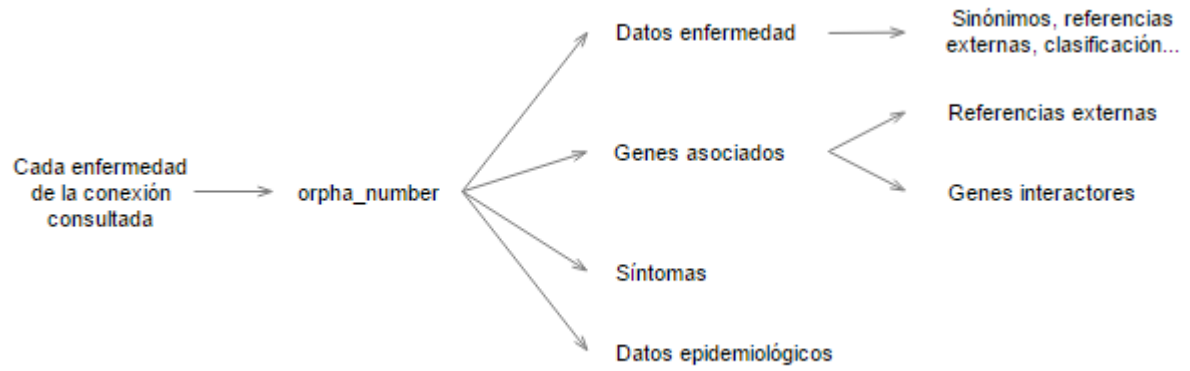

Figura 3.7: Búsqueda de conexión en la herramienta.

La consulta de cada una de las enfermedades de la conexión es muy similar a la búsqueda por enfermedad rara. En ésta sin embargo, se añade un paso más: para cada gen asociado a la enfermedad se buscan los genes que presentan interacciones proteína-proteína.

Las asociaciones entre genes por interacción proteína-proteína se listan ordenadas según sea mayor su puntuación. Además, la red que muestra la conexión incluye las dos enfermedades y todos los genes asociados a cada una de ellas, destacando si son genes comunes o las posibles interacciones entre ellos.

Cabe destacar que en la implementación de este tipo de consulta se encuentra un problema: relaciones entre genes repetidas debido a interacciones duplicadas en los datos extraídos de HIPPIE. Por tanto, se requiere realizar un paso extra tras la consulta que consiste en la eliminación de estas relaciones repetidas.

3.2.3.3. Gen

La tercera consulta disponible se realiza mediante el símbolo de un gen. A partir de éste se obtienen las referencias externas, las enfermedades asociadas y los genes relacionados con él mediante interacciones proteína-proteína (Figura 3.8).

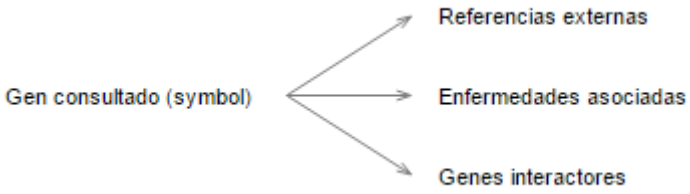

Figura 3.8: Búsqueda de gen en la herramienta.

La red que se representa en la vista muestra todos los genes y enfermedades relacionadas centrándose en el gen de consulta.

3.2.4. Monitorización de uso

Con la intención de poder medir el comportamiento de los usuarios de la aplicación web creada se añade una sección de código JavaScript que permite monitorizar el uso de la aplicación y extraer métricas de él.

Para ello se elige *Google Analytics* (Figura 3.9): una herramienta de analítica web gratuita que ofrece información agrupada del tráfico que llega a un sitio web según la audiencia, la adquisición, el comportamiento y las conversiones que se llevan a cabo en el sitio. Es un servicio con herramientas estadísticas y de análisis y una interfaz muy completa con gráficos e informes predeterminados y personalizables.

Algunos de los datos que se pueden obtener son: número de visitas, duración de las sesiones, tasa de rebote, datos sociodemográficos de los usuarios (lenguaje, ubicación, proveedor de Internet), registro de su comportamiento dentro del sitio web (fuentes de tráfico, páginas visitadas, secciones preferidas, desplazamientos entre ellas, palabras clave usadas), informes en tiempo real, registro del contenido más popular, detalles técnicos de los dispositivos de los visitantes (navegador, sistema operativo, referencia del móvil utilizado para acceder) y múltiples gráficos estadísticos entre otros.

Cuando un visitante llega a cualquiera de las páginas que contienen el código de seguimiento, éste se carga de manera simultánea a los demás elementos de la página y genera una *cookie*, un archivo de datos que se guarda en el ordenador o dispositivo móvil a través del navegador, el cual va registrando las variables antes mencionadas hasta que termina la visita. Mientras esto sucede, se van cargando a los servidores de Google todos

los datos capturados y luego se generan en el panel de Google Analytics los informes correspondientes, incluso en tiempo real.

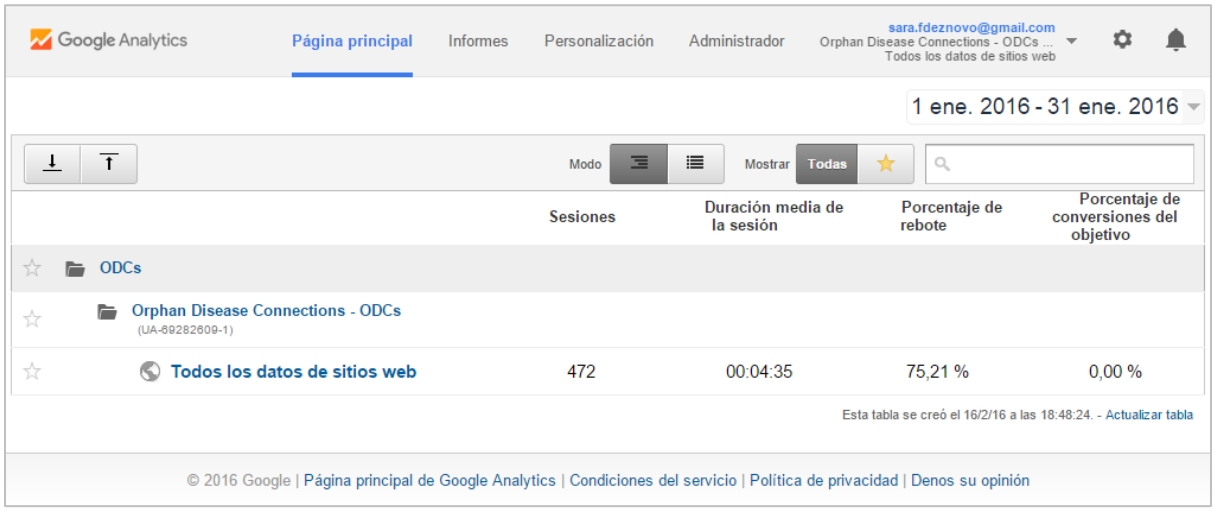

Figura 3.9: Portal de Google Analytics para el sitio web ODCs. Se muestra el número de sesiones, la duración media de la sesión, el porcentaje de rebote y las conversiones en el mes de enero de 2016.

La incorporación de este sistema de analíticas al sitio web ODCs permitirá evaluar el comportamiento de los usuarios con el fin de conocer el volumen de uso de la aplicación así como actuar en consecuencia en caso de ser necesario, por ejemplo, realizando mejoras en el servicio o en la presentación de la herramienta.

4. Resultados

En este capítulo se presentan los resultados alcanzados tras el desarrollo del presente trabajo divididos en cinco secciones. Primero se muestran las redes de enfermedades raras y genes construidas a partir de la información integrada en la base de datos y se presentan los resultados de la validación de dichas redes. A continuación se expone la herramienta pública de consulta desarrollada y además se demuestra su utilidad presentando nuevas conexiones entre enfermedades que se establecen gracias a ella. Se finalizan los resultados con una muestra del uso de la aplicación calculado mediante el sistema de analíticas integrado en el sistema.

4.1. Redes de enfermedades raras y genes

A continuación se describen las redes de enfermedades raras y genes construidas durante la realización del proyecto.

4.1.1. Red de enfermedades raras relacionadas por genes comunes

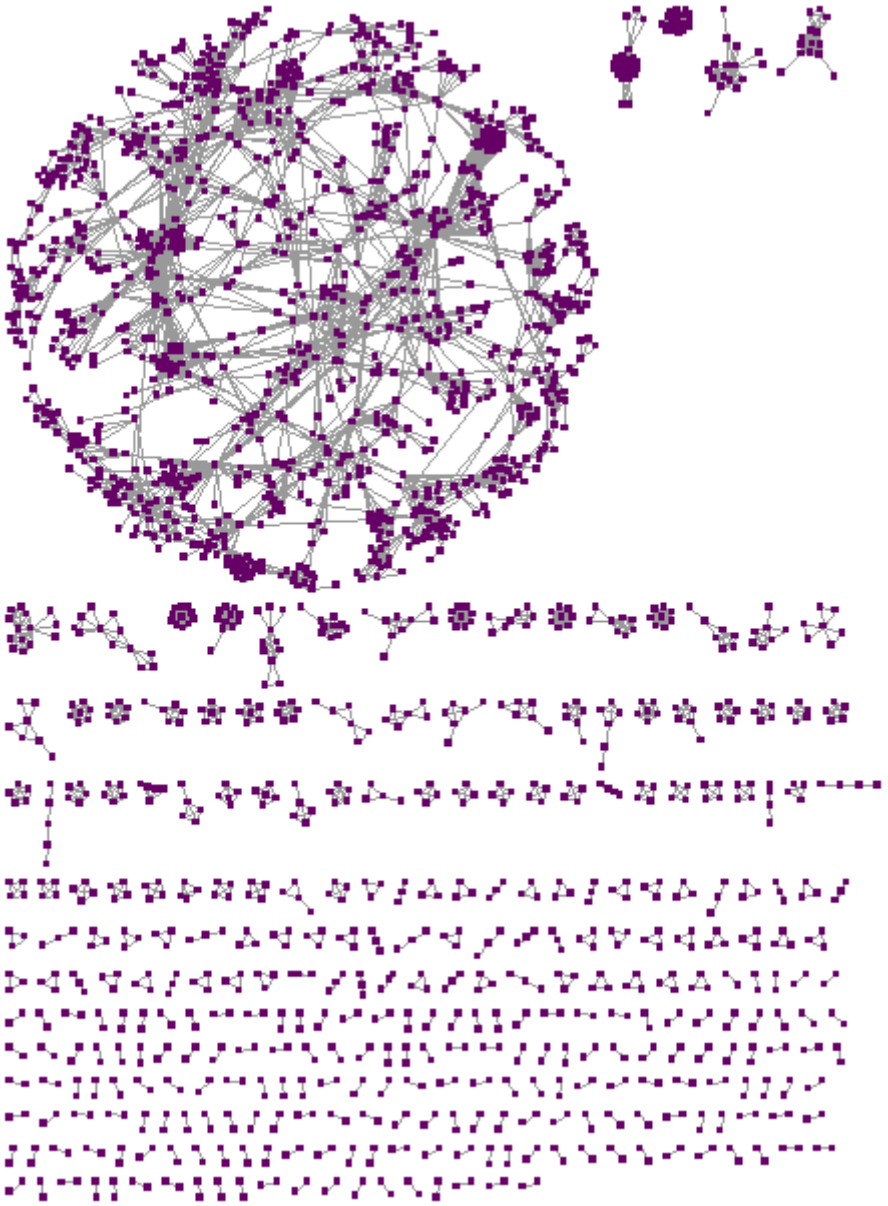

Figura 4.1: Red de enfermedades raras relacionadas por genes de susceptibilidad comunes.

A partir de la información integrada en la base de datos de ODCs se construye una red de enfermedades con el software de Cytoscape[20]. Para ello, se integran en el programa las conexiones entre enfermedades por genes de susceptibilidad compartidos, extrayéndose éstas mediante una *query* a la tabla que contiene los pares de enfermedades relacionados.

La Figura 4.1 muestra la red de enfermedades (representadas mediante puntos morados) relacionadas por genes de susceptibilidad comunes (enlaces grises). En la parte superior se aprecia una red principal (componente conexo) más grande y en la parte inferior multitud de enfermedades formando subredes aisladas de la principal. La red presenta 2.083 enfermedades con un total de 5.263 conexiones entre ellas. En la red se observa un elevado número de enfermedades conectadas en el mayor componente conexo, concretamente 1.066, quedando un gran número de ellas aisladas sin conexión con la red principal.

4.1.2. Red de enfermedades raras relacionadas por genes e interacciones

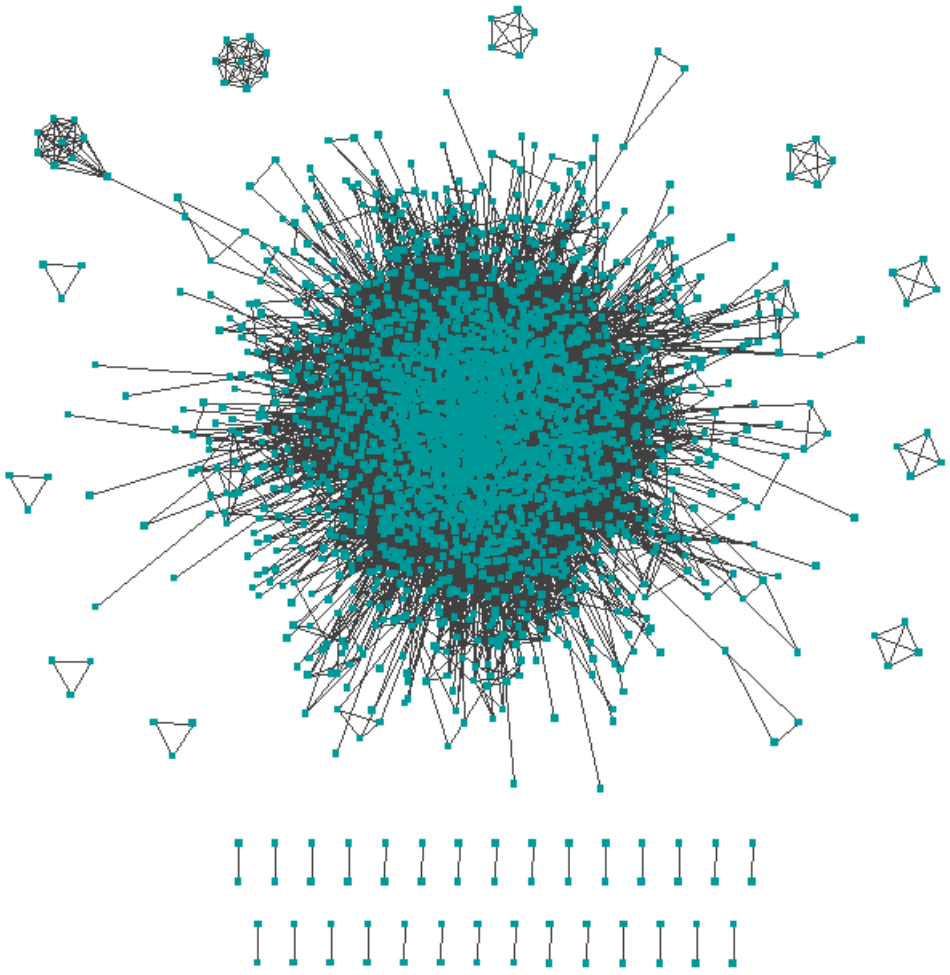

Figura 4.2: Red de enfermedades raras relacionadas por genes de susceptibilidad comunes e interacciones proteína-proteína.

Se incorporan las interacciones proteína-proteína a la red anterior con la intención de establecer un número mayor de conexiones entre enfermedades raras. Se genera por tanto una red global de enfermedades cuyas relaciones se constituyen a través de la integración de genes de susceptibilidad de las enfermedades así como las interacciones entre proteínas de dichos genes. Para ello, se utiliza el sistema de puntuación de relaciones desarrollado durante este proyecto, materializando en base de datos los pares de enfermedades con el *score* asociado y extrayendo posteriormente esa información para integrarla en el programa Cytoscape.

La red (Figura 4.2) presenta 2.818 enfermedades (representadas mediante puntos verdes), de las cuales 2.718 están en red principal, relacionadas mediante 54.941 conexiones (aristas negras). Se observan menos enfermedades aisladas gracias al mayor número de conexiones establecidas. Las relaciones que figuran en esta red se materializan en base de datos para el sistema de consultas.

4.1.3. Otras redes de enfermedades y genes

Además de las redes de enfermedades raras expuestas, se construyen otras redes para analizar el comportamiento de los datos así como visualizar las conexiones entre enfermedades raras y sus genes de susceptibilidad.

Las redes creadas, junto con sus resultados, son las siguientes:

- Red de enfermedades raras y genes con conexiones enfermedad-gen por asociación: 3.234 nodos (2.083 enfermedades y 1.151 genes) con 3.723 enlaces.

- Red de genes relacionados por interacciones proteína-proteína: 15.001 genes con un total de 179.899 interacciones entre ellos.

- Red de enfermedades raras relacionadas únicamente por interacción entre proteínas: 2.583 enfermedades con 48.847 conexiones.

Para concluir el análisis de las redes construidas la tabla 4.1 muestra una comparativa de los datos de cada red. Se destaca la red con mayor número de enfermedades raras relacionadas y el elevado número de conexiones entre ellas.

Tabla 4.1: Comparativa de las redes construidas. Se muestra el número de enfermedades, genes y conexiones presentes en las diferentes redes.

Red	Número de enfermedades	Número de genes	Número de conexiones
Red de enfermedades raras relacionadas por genes comunes	2083	-	5263

Red de enfermedades raras relacionadas por genes e interacciones	**2818**	-	**54941**
Red de enfermedades raras y genes asociados	2083	1151	3723
Red de genes relacionados por interacciones proteína-proteína	-	15001	179899
Red de enfermedades raras relacionadas sólo por interacción entre proteínas	2583	-	48847

4.2. Validación de las redes de enfermedades raras

Con el objetivo de validar las relaciones entre enfermedades obtenidas se calculó la similitud fenotípica de todos los pares de enfermedades raras para el subconjunto de enfermedades con información fenotípica y genes de susceptibilidad disponibles. La similitud fenotípica se calculó de la siguiente manera:

En primer lugar, los fenotipos asociados a las enfermedades raras fueron elaborados a partir de la Ontología de Fenotipo Humano (*Human Phenotype Ontology*, HPO)[25] y las asociaciones directas enfermedad-fenotipo se ampliaron para incluir los términos padre de un determinado fenotipo en la jerarquía HPO. En segundo lugar, se calculó la probabilidad de cada término fenotípico en la ontología p(c), como el número de enfermedades asociadas a ella, dividido por el número total de enfermedades. Finalmente, la similitud fenotípica entre dos enfermedades se definió basándose en la probabilidad del más específico fenotipo común:

$$sim(d1, d2) = \max_{c \in S\,(c1,c2)} [-\log p(c)]$$

Fórmula 4.1: Cálculo de similitud fenotípica para cada par de enfermedades.

donde c1 y c2 son todos los términos fenotipo asociados a las enfermedades d1 y d2, respectivamente.

A continuación, se clasificaron los pares de enfermedades en cinco grupos: aquellos que compartían

(i) genes,
(ii) interacciones entre proteínas,
(iii) rutas (*pathways*) moleculares,
(iv) complejos de proteicos, y
(v) los pares restantes, que no compartían nada de lo anterior.

Las interacciones entre proteínas humanas fueron compilados a partir de los datos de HIPPIE[6], las rutas de Reactome[26] y los complejos proteicos de CORUM[27].

La Figura 4.3 muestra la distribución de similitud fenotípica de cada grupo. Como se esperaba, las enfermedades que comparten genes tienen más fenotipos similares (valor medio 1,38), seguidas de aquellas que comparten interacciones (0,93), rutas (0.84), complejos (0,64) y el resto de pares (0.39).

De acuerdo con estos resultados, la construcción de conexiones entre enfermedades basándose en interacciones entre proteínas es el enfoque que producirá la mayor proporción de asociaciones significativas entre enfermedades con el fin para aumentar aquellas construidas mediante genes comunes.

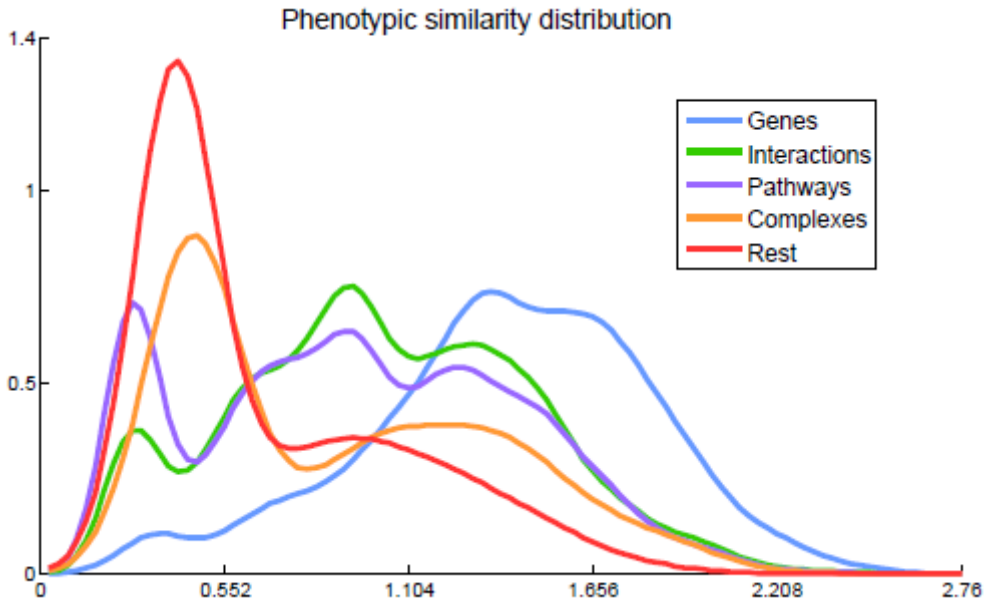

Figura 4.3: Distribución de similitud fenotípica de pares de enfermedades raras correspondiente a aquellos que comparten genes, interacciones, vías, complejos y pares restantes. El eje de abscisas representa la similitud entre dos enfermedades sim(d1,d2), calculada según la Fórmula 4.1, y el eje de ordenadas, la estimación de la función densidad de probabilidad (fdp).

Este análisis justifica por tanto la inclusión de las interacciones entre proteínas en el sistema con la intención de establecer mayor número de relaciones entre enfermedades raras.

4.3. ODCs: La herramienta de consulta

La herramienta resultante de este proyecto es **ODCs** (**Orphan Disease Connections**), disponible en http://csbg.cnb.csic.es/odcs, un recurso que contiene información sobre enfermedades raras, prestando especial atención a las relaciones entre las afecciones por genes de susceptibilidad comunes así como por interacciones entre proteínas de los correspondientes productos génicos. Esta herramienta incluye una interfaz web que permite la búsqueda de las enfermedades relacionadas con una enfermedad de interés (y sus genes/proteínas asociados), la exploración en detalle de las conexiones entre dos enfermedades raras o la búsqueda de enfermedades raras asociadas a un gen concreto.

4.3.1. Interfaz web

La interfaz web (Figura 4.4) está diseñada pensando en la simplicidad, de modo que rellenando un sencillo cuadro de texto se acceda directamente a los datos y se pueda navegar cómodamente por ellos. Los resultados se presentan tanto en forma textual como en un gráfico interactivo (visualización de la red) e incluyen enlaces a diversos recursos externos. También hay una página de ayuda tutorial y se proporcionan ejemplos en el formulario de entrada.

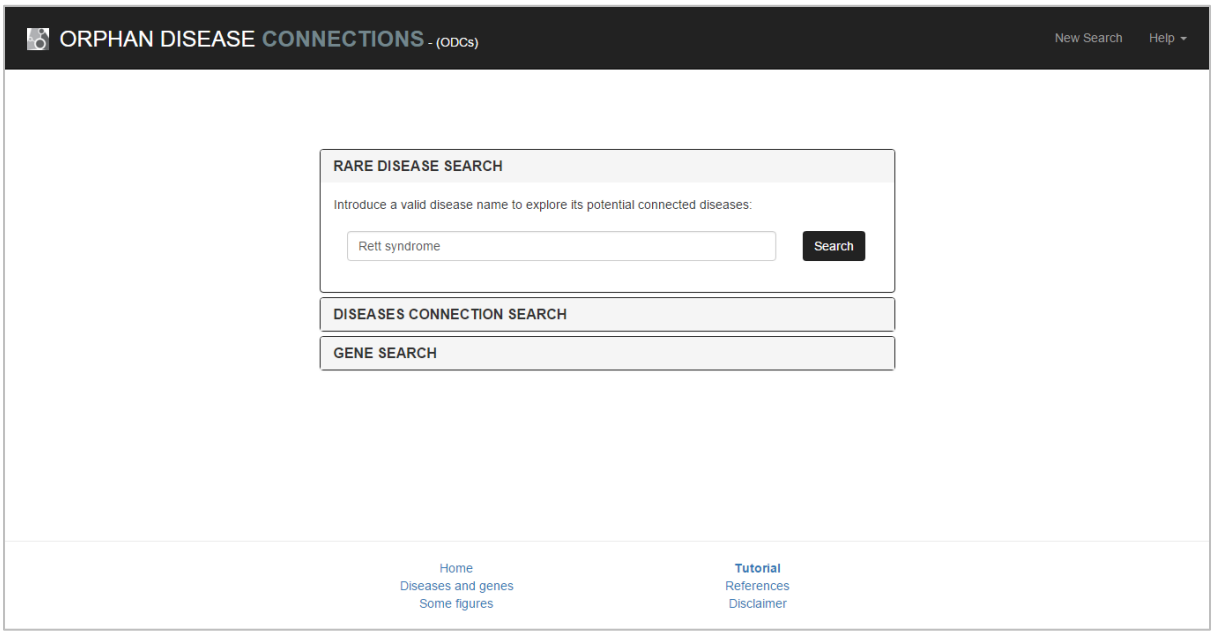

Figura 4.4: Portal web *Orphan Disease Connections* (ODCs). La página principal del sitio muestra unos cuadros de texto habilitados para la búsqueda de enfermedades, conexiones y genes. En el pie de página se facilitan enlaces a las páginas de ayuda, entre ellas el tutorial, las referencias y algunos datos sobre ODCs.

4.3.2. Datos en ODCs

ODCs tiene dos fuentes principales de datos: información sobre enfermedades raras y genes de susceptibilidad extraídos de Orphadata e interacciones entre proteínas humanas tomadas de HIPPIE.

Dos enfermedades raras están conectados en ODCs si comparten un gen de susceptibilidad o si existe al menos una interacción entre las proteínas codificadas por sus genes de susceptibilidad. Con este enfoque integrador se puede establecer un número mucho mayor de conexiones que las basadas únicamente en los genes compartidos, ya que la mayoría de las enfermedades raras (73%) están asociados con un único gen.

La versión actual contiene 54.941 relaciones entre 2.818 enfermedades, de las cuales 5.263 corresponden a relaciones basadas en gen compartido.

Además ODCs contiene enlaces externos para navegar cómodamente por otros recursos biomédicos importantes como la Clasificación Internacional de Enfermedades (ICD-10), el catálogo de genes humanos y desórdenes genéticos OMIM, el recurso de proteínas universal UniProt y el vocabulario terminológico controlado MeSH.

4.3.2.1. ODCs en números

Todas las enfermedades raras incluidas en la aplicación de ODCs tienen al menos un gen asociado. Del mismo modo, los genes incorporados en la herramienta están asociados con una o varias enfermedades. Además, con el propósito de poder establecer relaciones basadas en el interactoma, los genes incluidos en la base de datos de ODCs tienen referencia UniProt.

Tabla 4.2: ODCs en números.

Número total de enfermedades raras en la base de datos Orphadata	6.838
Número total de interacciones proteína-proteína en la base de datos HIPPIE	179.899
Enfermedades con al menos un gen asociado	3.032
Nombres de enfermedades y sinónimos	8.432
Genes asociados con al menos una enfermedad	3.061
Número total de asociaciones enfermedad-gen	5.718
Número total de asociaciones enfermedad-enfermedad	54.941
Número total de asociaciones enfermedad-enfermedad basadas en genes comunes	5.263

Enfermedades con al menos un gen común con otra	2.083
Número total de síntomas incluidos en ODCs	1.179
Enfermedades con información sintomática	1.120

4.3.3. Búsqueda en la herramienta

Cada uno de los tres tipos diferentes de búsqueda inicial que se puede realizar en ODCs proporciona una visión alternativa de las conexiones, centrándose en una enfermedad, la conexión entre dos enfermedades o un gen. Todos ellos ofrecen una visualización gráfica de las asociaciones encontradas (enfermedad-gen, enfermedad-enfermedad y gen-gen) en forma de red, así como información textual detallada y enlaces a recursos externos, junto con la opción de descargar los resultados (en formato CSV) (Figura 4.5).

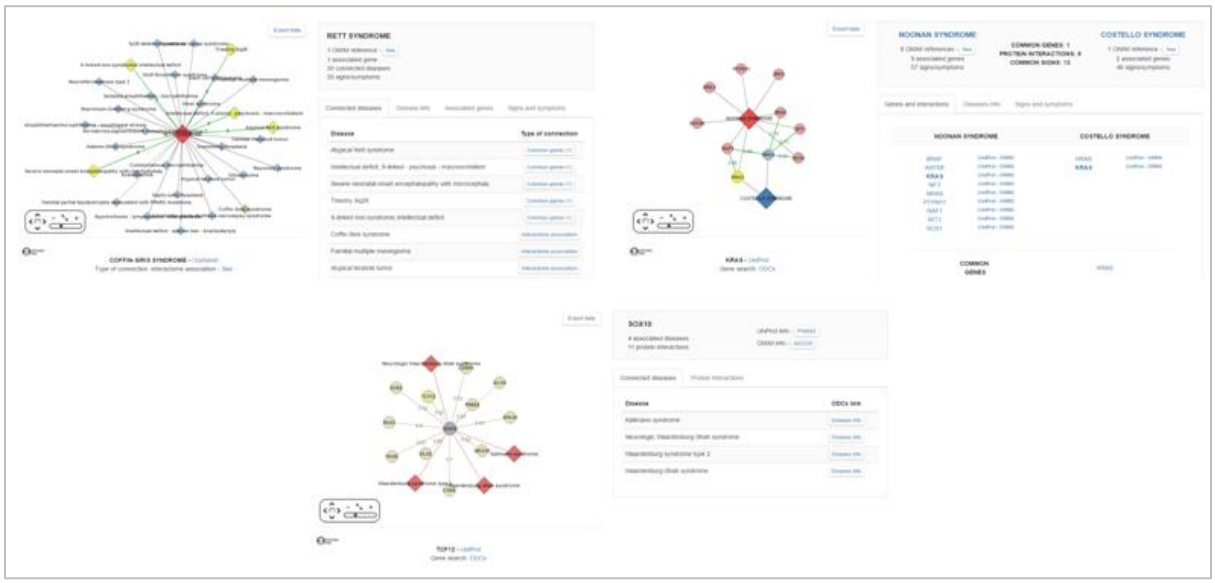

Figura 4.5: Visualización gráfica e información textual en los resultados de la búsqueda por enfermedad, conexión y gen.

Los cuadros de texto de búsqueda inicial para enfermedades y genes (Figura 4.6) incluyen una lista desplegable que muestra los elementos que coinciden con el criterio de búsqueda cuando el usuario introduce un número predeterminado de caracteres (en la configuración actual, 3 para nombre de enfermedad y 1 para símbolo de gen).

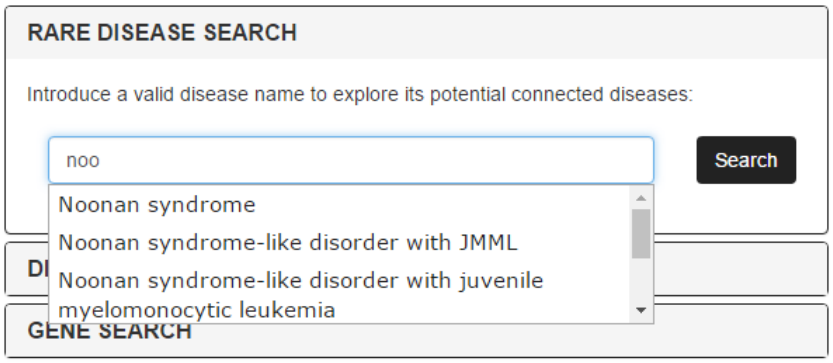

Figura 4.6: Cuadro de texto para la búsqueda de enfermedad rara. En el desplegable se muestran las enfermedades que coinciden textualmente con las letras introducidas.

Una vez introducido el campo de búsqueda en la capa de presentación, la herramienta realiza una *query* SQL desde la capa de negocio a la base de datos ODCs, gestionando entre otras las siguientes excepciones e informando de ello al usuario:

- La enfermedad o gen no se encuentra en la base de datos.
- La enfermedad no posee información asociada referente a sinónimos, datos epidemiológicos, síntomas o referencias externas.
- El gen no presenta interacciones proteína-proteína o referencias externas asociadas.

4.3.4. Resultados en ODCs

La búsqueda por enfermedad rara conduce a una página donde los resultados aparecen en una vista centrada en la enfermedad (Figura 4.6). Todas las enfermedades relacionadas con la enfermedad de consulta se muestran tanto gráficamente, en una red interactiva centrada en la enfermedad de consulta, como en una lista. En esta lista se incluyen en primer lugar las enfermedades relacionadas por genes comunes, ordenadas por mayor número, y a continuación las de interacciones de proteínas, ordenadas por su puntuación calculada. El tipo de conexión se diferencia también en el gráfico. Se proporciona además la clasificación de la enfermedad (Orphanet e ICD-10), los datos epidemiológicos, los síntomas y los genes asociados (con enlaces a UniProt y OMIM). Desde esta página, el usuario puede navegar a la página de vista de conexión pinchando en las enfermedades de la lista o a la de vista de gen pinchando en los genes asociados.

La búsqueda de conexión entre dos enfermedades conduce a una representación donde se muestran todos los genes asociados a las dos enfermedades, así como las interacciones establecidas entre sus productos, si existen (Figura 4.8). Los genes compartidos se remarcan tanto en la red como en el texto y las interacciones de HIPPIE se acompañan con la puntuación correspondiente. Para ambas enfermedades se muestran también sinónimos, clasificación y datos epidemiológicos, así como síntomas comunes y no comunes.

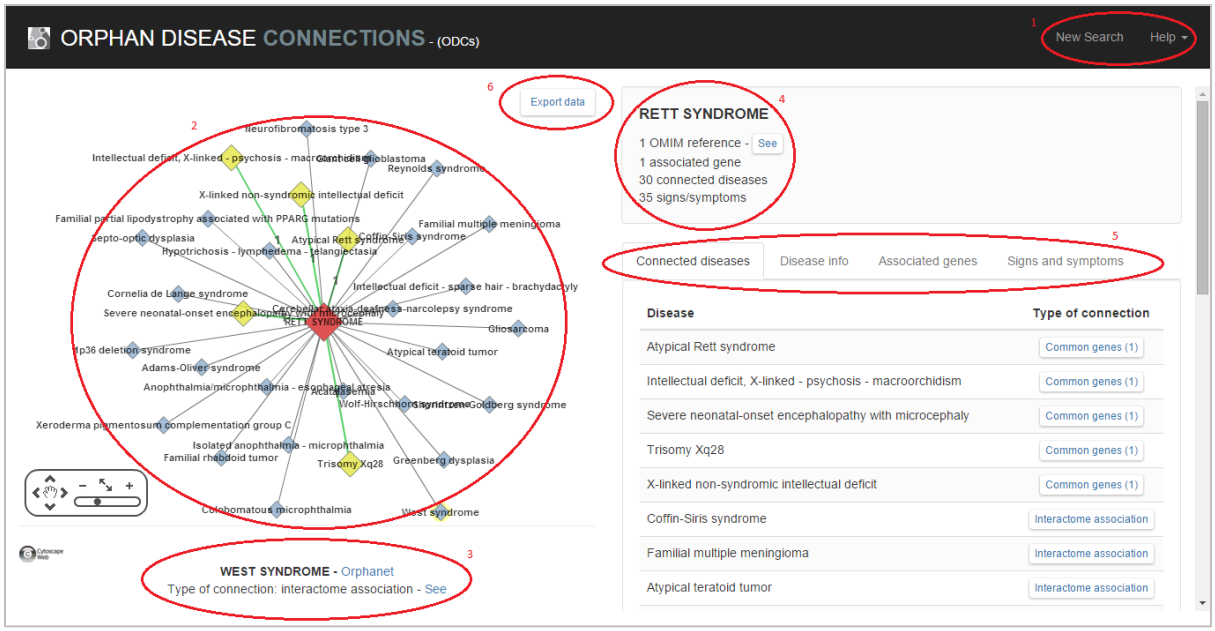

Figura 4.7: Resultados de la búsqueda de enfermedad rara. Se resaltan las secciones más relevantes de la vista: (1) acceso a la página de inicio y páginas de ayuda desde la barra de navegación, (2) visualización de la red resultante, (3) resultados de la interacción del usuario con la red, (4) panel resumen de la enfermedad consultada, (5) pestañas que contienen las tablas e informes de resultados (enfermedades conectadas, información de la enfermedad, genes asociados y síntomas clínicos) y (6) botón que permite la descarga de los resultados de la búsqueda.

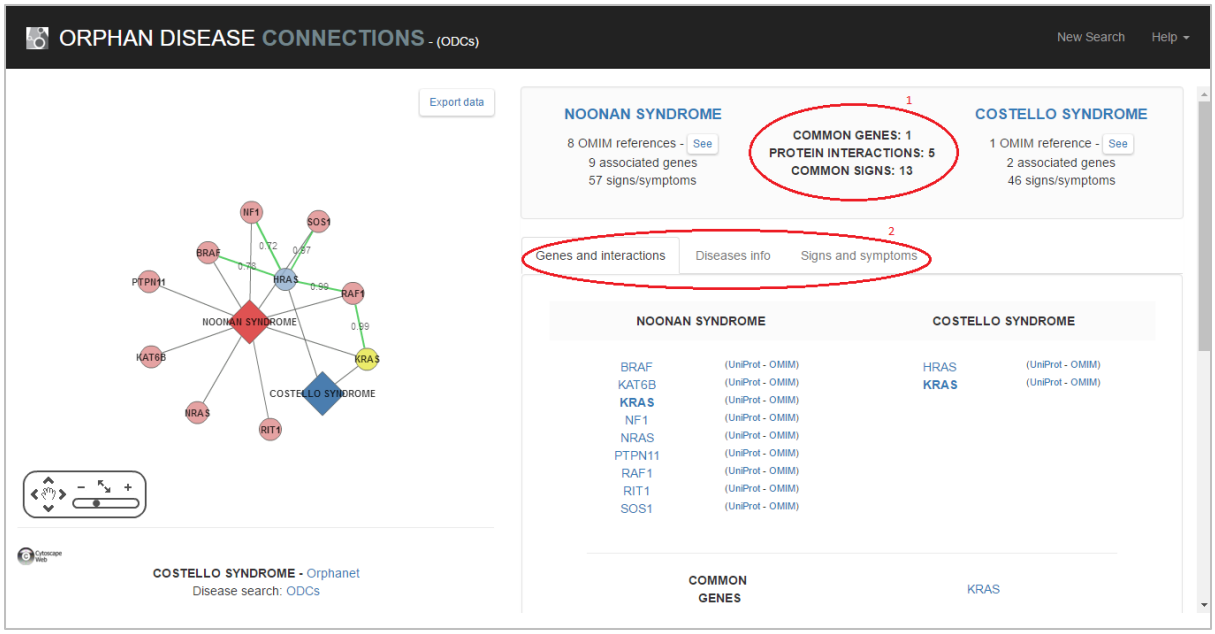

Figura 4.8: Resultados de la búsqueda de conexión entre dos enfermedades raras. Se destaca la comparativa entre los datos de las dos enfermedades en el cuadro resumen (1) y las pestañas que contienen las tablas e informes de resultados comparativos (genes e interacciones, información de ambas enfermedades y sus síntomas clínicos) (2).

Para ayudar a los usuarios a juzgar la relevancia de la conexión de las enfermedades se proporcionan dos mecanismos (ilustrados en la Figura 4.9):

- Enlaces externos al sitio web HIPPIE con las evidencias experimentales que apoyan cada interacción proteína-proteína.
- Búsqueda de forma interactiva en PubMed (mediante la selección de las enfermedades y sus genes), para verificar si diferentes asociaciones entre genes que interactúan y enfermedades han sido reportados en la literatura.

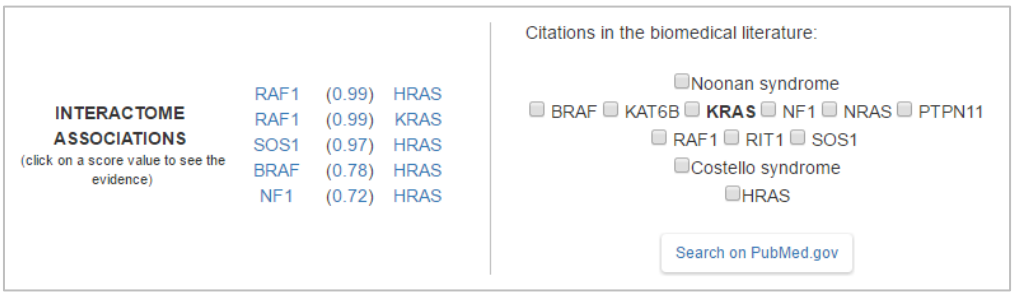

Figura 4.9: Mecanismos para comprobar la conexión entre enfermedades raras: enlaces externos en cada interacción proteína-proteína y búsqueda interactiva en PubMed mediante la selección de enfermedades y genes.

Por último, está disponible la búsqueda centrada en el gen (Figura 4.10). En este caso, se muestran todas las enfermedades asociadas al gen, así como los genes cuyos productos proteicos se sabe que interactúan con los del gen consultado. Desde aquí, el usuario puede navegar a una enfermedad determinada o a otro gen de los interactores.

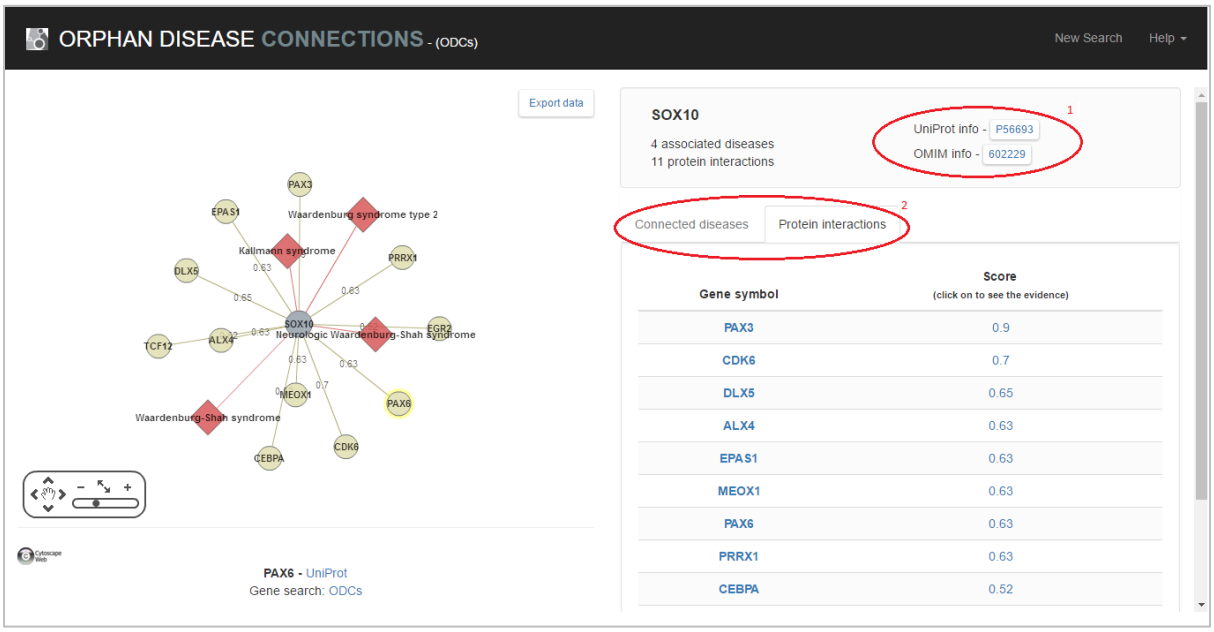

Figura 4.10: Resultados de la búsqueda de gen. Se resaltan las referencias externas en el cuadro resumen (1) y las pestañas que contienen las tablas de resultados (enfermedades conectadas e interacciones entre proteínas) (2).

4.4. Nuevas relaciones entre enfermedades raras

Muchas de las conexiones entre enfermedades raras presentes en la herramienta de consulta se sustentan por genes de susceptibilidad comunes, y otras por la combinación de éstos y un conjunto de interacciones proteína-proteína. Sin embargo, hay otras asociaciones que sólo pueden ser establecidas mediante interacciones proteína-proteína.

La Figura 4.11 muestra un ejemplo de conexión entre dos enfermedades raras generada por la herramienta siguiendo el criterio anterior. En ella se pueden observar las asociaciones moleculares entre el síndrome de Noonan y el síndrome de microdeleción 22q11.2 distal (diamantes rojo y azul, respectivamente). Esta conexión se constituye en base a las interacciones entre las proteínas de sus genes de susceptibilidad (representados mediante círculos). En efecto, ambos síndromes pertenecen al grupo de enfermedades del desarrollo producidas por mutaciones que afectan a las proteínas de la ruta de señalización de la ERK MAP quinasa[28].

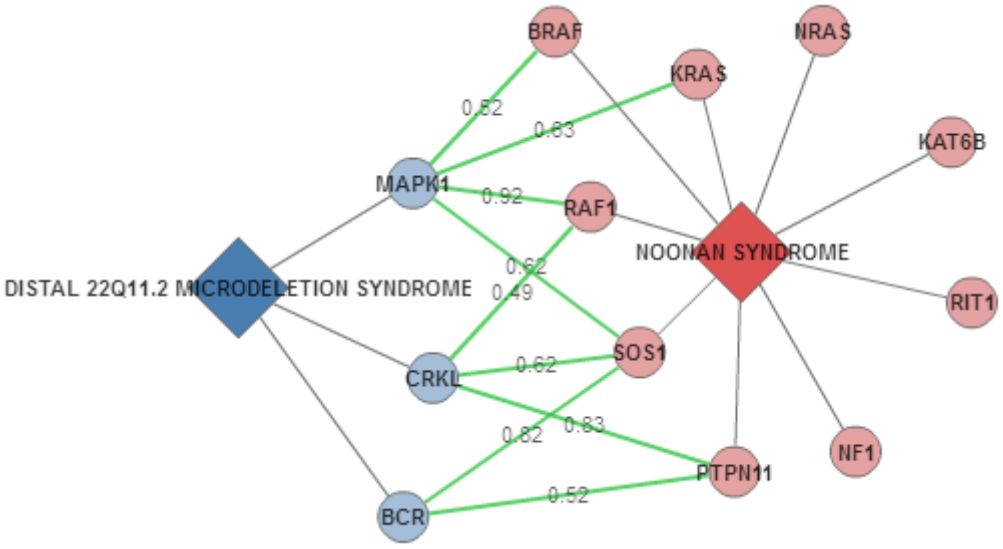

Figura 4.11: Conexión de dos enfermedades raras generada por ODCs establecida en base a las interacciones entre las proteínas de sus genes de susceptibilidad.

De igual manera, gracias a la incorporación de las interacciones proteína-proteína en el sistema desarrollado se pueden valorar nuevas relaciones entre enfermedades raras que hasta ahora no habían sido planteadas en ningún otro sistema. La herramienta desarrollada posibilita por tanto a los investigadores biomédicos una nueva forma de estudio de estas afecciones mediante conexiones con otras enfermedades raras y no como entidades independientes, tal como se venía realizando hasta el momento.

4.5. Analíticas de uso

ODCs es una herramienta plenamente funcional, operativa desde noviembre de 2015. Este hecho permite presentar este módulo de analítica con resultados de usuarios reales. De esta forma, además de validar el correcto funcionamiento de este módulo, permite demostrar su utilidad.

El sistema de analíticas integrado en la herramienta de consulta permite obtener métricas del uso de la aplicación. Dentro de los múltiples resultados que se proporcionan se presta especial atención al número de usuarios y a las sesiones realizadas por ellos.

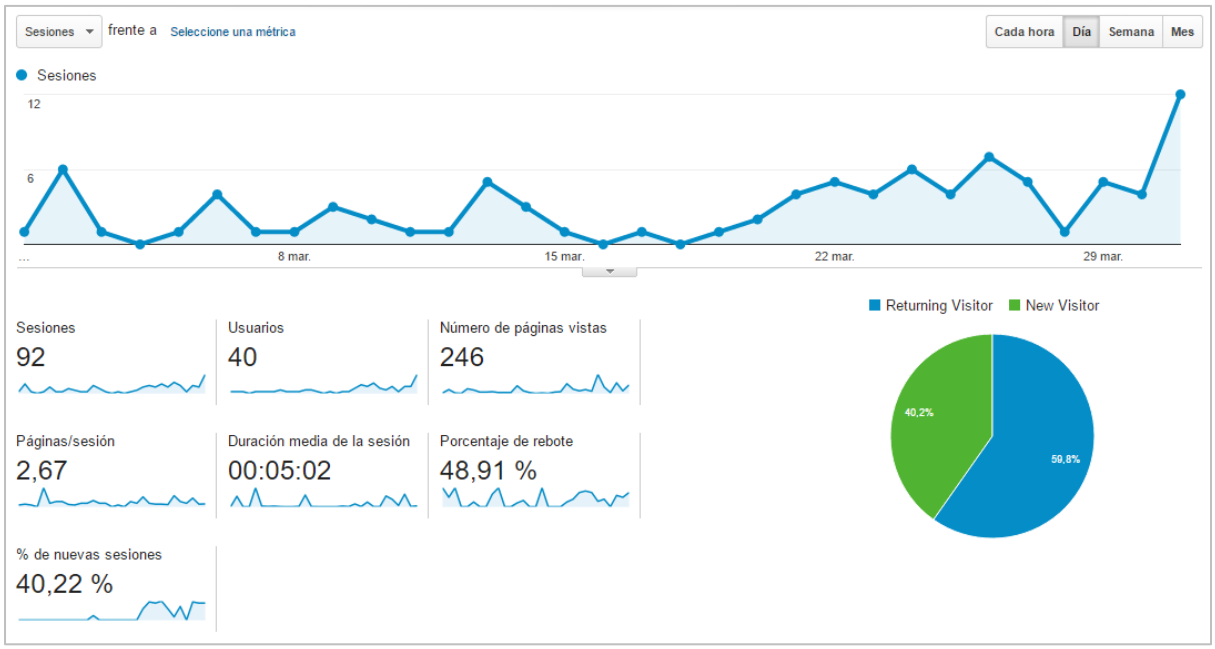

Figura 4.12: Resultado de sesiones en el portal ODCs entre el 1 y el 31 de marzo de 2016.

En la Figura 4.12 y la Figura 4.13 se pueden observar dos presentaciones diferentes de las métricas correspondientes a los meses de marzo y febrero, respectivamente. En la primera de ellas se muestra un cuadro resumen que contiene el número de sesiones, el número de usuarios, el total de páginas visitadas, las páginas por sesión y la duración media de ésta. Además se diferencia entre visitantes nuevos (nuevas sesiones) y visitantes conocidos que vuelven a la aplicación. Se pueden apreciar también las sesiones por día en una gráfica y el tipo de visitante en un diagrama de sectores.

En la Figura 4.13 se analizan resultados similares a los que aparecen en la Figura 4.12. Sin embargo, aporta mayor información, ya que todos los resultados de usuarios y sesiones

aparecen en gráficas con datos diferenciados por día, y además, se añaden la ubicación de las sesiones y el navegador elegido para realizarlas.

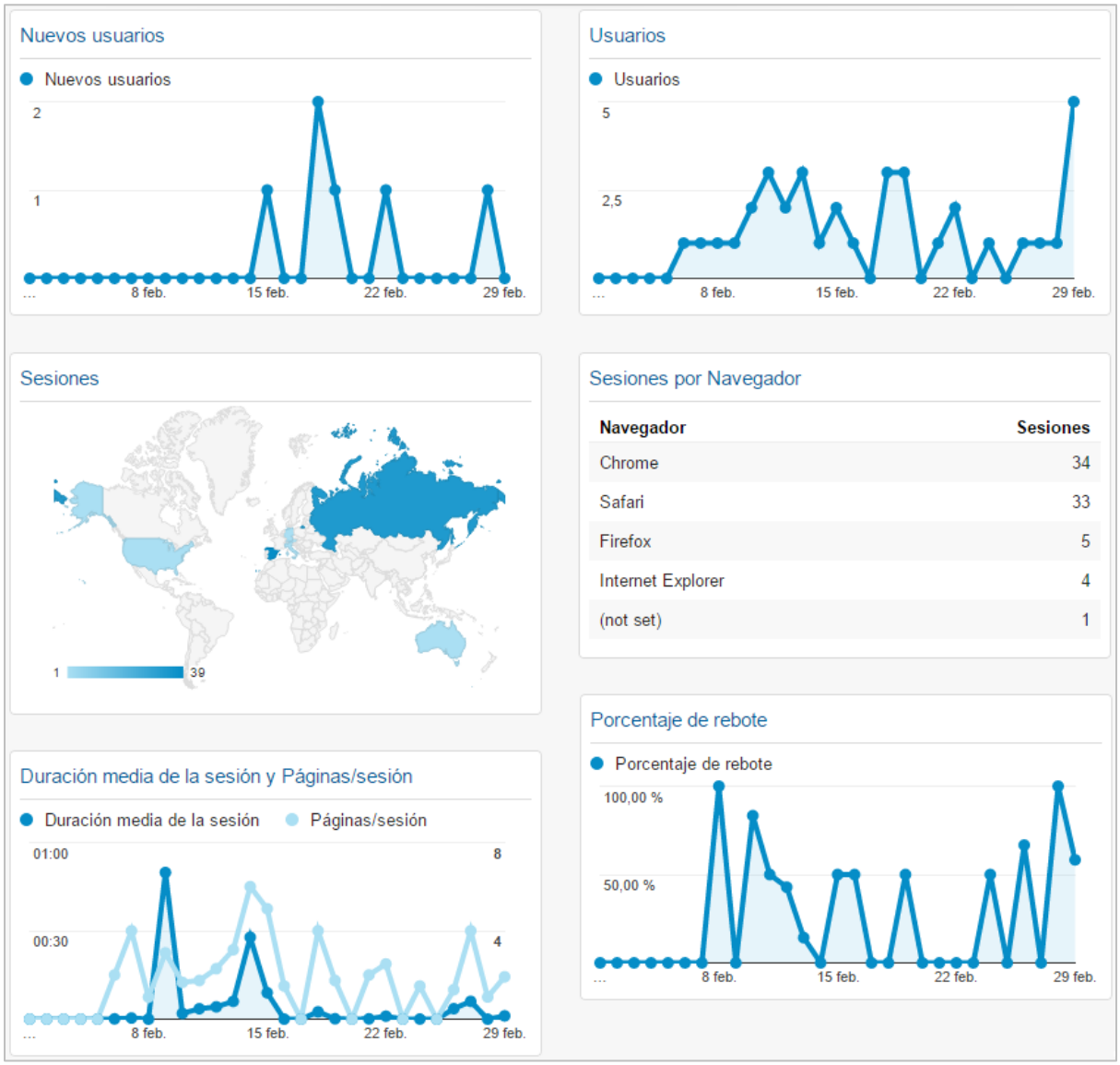

Figura 4.13: Resultado de usuarios, localización y sesiones del sitio ODCs entre el 1 y el 29 de febrero de 2016.

5. Conclusiones

A lo largo de este proyecto se ha desarrollado una nueva estrategia para establecer relaciones moleculares entre enfermedades raras y una herramienta informática para la consulta de dichas relaciones (ODCs). Se trata de una solución completa y plenamente funcional disponible para ser consultada de forma pública. Esta aplicación permite explorar el máximo de información referente a las enfermedades raras, especialmente las relaciones entre ellas, integrando datos de redes de interacción de proteínas y genes de susceptibilidad de las distintas afecciones. La herramienta proporciona además la búsqueda sencilla de enfermedades y genes en recursos externos, tales como catálogos de enfermedades y literatura médica.

Se ha generado una red global de relaciones entre enfermedades raras facilitando el estudio general de las afecciones y la búsqueda de mecanismos moleculares de relación. Las conexiones entre enfermedades se han establecido a través de dos mecanismos: genes de susceptibilidad comunes y relaciones indirectas definidas por el interactoma. Se ha demostrado además que las interacciones entre proteínas codificadas por los genes tienen validez en el establecimiento de conexiones entre enfermedades raras y que abren una puerta al estudio de nuevas relaciones entre ellas.

La herramienta se enmarca dentro de la línea de trabajo sobre el estudio funcional de redes biológicas llevado a cabo por el grupo de investigación *Computational Systems Biology Group* del Centro Nacional de Biotecnología del Consejo Superior de Investigaciones Científicas (CNB-CSIC). Se encuentra en pleno servicio en el servidor del departamento y las analíticas demuestran que está en uso.

El sistema desarrollado durante este proyecto ha dado lugar a una publicación en una revista científica (*peer review*): Fernández-Novo S, Pazos F y Chagoyen M, “Rare disease relations through common genes and protein interactions”, Molecular and Cellular Probes (2016) (incluida como Anexo A - Publicación).

Confiamos en que el sistema desarrollado permita a los investigadores biomédicos y clínicos la transferencia de conocimientos entre distintas enfermedades y establecer sinergias entre líneas de investigación aisladas, especialmente importantes en el caso de las enfermedades raras debido a la especialización y la dispersión de los recursos dedicados a ellas.

Glosario

Checkbox Casilla de verificación, elemento de la interfaz gráfica de usuario que permite hacer selecciones múltiples de un conjunto de opciones.
Comorbilidad Presencia de dos o más enfermedades en un mismo paciente.
CSS Cascading Style Sheets.
CSV Comma-Separated Values, formato abierto sencillo para representar datos en forma de tabla.
Dataset Conjunto de datos.
DNA Ácido DesoxirriboNucleico.
Fenotipo Expresión del genotipo (información genética de un organismo en forma de DNA) en función de un determinado ambiente.
Framework Estructura conceptual y tecnológica de soporte definido, con módulos de software, que sirven de base para la organización y el desarrollo de software.
Gene Ontology Vocabulario controlado que describe diversos aspectos funcionales de los productos génicos (actividades moleculares, procesos biológicos y componentes celulares).
HTML HyperText Markup Language.
Interactoma Identificación sistemática de interacciones entre proteínas dentro de un organismo.
Interfaz de usuario Medio con que el usuario se comunica con un dispositivo o máquina.
JSON JavaScript Object Notation, formato ligero para el intercambio de datos.
MySQL Sistema de gestión de bases de datos relacional, multihilo y multiusuario.
Offline Fuera de línea.
OMS Organización Mundial de la Salud.
On-the-fly En tiempo real, en el momento.
PHP Hypertext Pre-processor.
Plugin Aplicación que se relaciona con otra para agregarle una función nueva específica.
RNA Ácido RiboNucleico.
SQL Structured Query Language.
UI User Interface, interfaz de usuario.
URL Uniform Resource Locator.
Wireframe Guía visual que representa el esqueleto o estructura visual de un sitio web.

Referencias

[1] A. L. Barabási, N. Gulbahce and J. Loscalzo, Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.* 12(1) (2011) 56-68.
[2] M. Zhang, C. Zhu, A. Jacomy, L.J. Lu, A.G. Jegga, The orphan disease networks, *Am. J. Hum. Genet.* 88 (2011) 755e766.
[3] Orphanet©: http://www.orpha.net/ (visitada en abril de 2016).
[4] Orphadata: Free access data from Orphanet© INSERM 1997. Available on http://www.orphadata.org (visitada en abril de 2016).
[5] J.S. Amberger, C.A. Bocchini, F. Schiettecatte, A.F. Scott, A. Hamosh, OMIM.org: online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders, *Nucleic Acids Res*. 43 (2015) D789eD798.
[6] M.H. Schaefer, et al., HIPPIE: integrating protein interaction networks with experiment based quality scores, *PLoS One* 7 (2012) e31826.
[7] K.I. Goh, et al., The human disease network, *Proc. Natl. Acad. Sci. U. S. A.* 104 (2007) 8685e8690.
[8] M. Lu, et al., An analysis of human microRNA and disease associations, *PLoS One* 3 (2008) e3420.
[9] B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. Delisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, *Genome Biol*. 10 (2009) R91.
[10] S. Park, et al., Protein localization as a principal feature of the etiology and comorbidity of genetic diseases, *Mol. Syst. Biol.* 7 (2011) 494.
[11] X. Zhang, et al., The expanded human disease network combining protein protein interaction information, *Eur. J. Hum. Genet*. 19 (2011) 783-788.
[12] D.S. Lee, et al., The implications of human metabolic network topology for disease comorbidity, *Proc. Natl. Acad. Sci. U. S. A*. 105 (2008) 9880-9885.
[13] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner and J. A. Leunissen, *Eur. J. Hum. Genet.* 14 (2006) 535-542.
[14] X. Zhou, J. Menche, A. L. Barabasi and A. Sharma, *Nature comm*. 5 (2014) 4212.
[15] A. Rzhetsky, D. Wajngurt, N. Park and T. Zheng, *Proc. Natl. Acad. Sci. U. S. A.* 104 (2007) 11694-11699.
[16] C. A. Hidalgo, N. Blumm, A. L. Barabasi and N. A. Christakis, *PLoS Comput. Biol*. 5 (2009) e1000353.
[17] N. Rappaport, et al., MalaCards: an integrated compendium for diseases and their annotation, *Database* (Oxford) 2013 (2013) bat018.
[18] C.C. Liu, et al., DiseaseConnect: a comprehensive web server for mechanism based disease-disease connections, *Nucleic Acids Res*. 42 (2014) W137eW146.
[19] C.T. Lopes, et al., Cytoscape web: an interactive web-based network browser, *Bioinformatics* 26 (2010) 2347e2348.
[20] J. Park, D.S. Lee, N.A. Christakis, A.L. Barabasi, The impact of cellular networks on disease comorbidity, *Mol. Syst. Biol*. 5 (2009) 262.
[21] J. Menche, et al., Disease networks. uncovering disease-disease relationships through the incomplete interactome, *Science* 347 (2015) 1257601.
[22] M. Chagoyen, F. Pazos, Characterization of clinical signs in the human interactome, *Bioinformatics* (2016), http://dx.doi.org/10.1093/bioinformatics/btw054.
[23] Bootstrap: http://getbootstrap.com/ (visitada en abril de 2016).

[24] Shannon P, Markiel A, Ozier O; et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res*. 13 (11) (2003), 2498–504.
[25] S. Kohler, et al., The human phenotype ontology project: linking molecular biology and disease through phenotype data, *Nucleic Acids Res*. 42 (2014) D966-D974.
[26] D. Croft, et al., The Reactome pathway knowledgebase, *Nucleic Acids Res*. 42 (2014) D472-D477.
[27] A. Ruepp, et al., CORUM: the comprehensive resource of mammalian protein complexese-2009, *Nucleic Acids Res*. 38 (2010) D497-D501.
[28] I.S. Samuels, S.C. Saitta, G.E. Landreth, MAP'ing CNS development and cognition: an ERKsome process, *Neuron* 61 (2009) 160e167.

Anexos

A – Publicación

ARTICLE IN PRESS

Molecular and Cellular Probes xxx (2016) 1–4

Contents lists available at ScienceDirect

Molecular and Cellular Probes

journal homepage: www.elsevier.com/locate/ymcpr

Short communication

Rare disease relations through common genes and protein interactions

Sara Fernandez-Novo [a], Florencio Pazos [b], Monica Chagoyen [b, *]

[a] Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid 28049, Spain
[b] Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Madrid 28049, Spain

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form
15 March 2016
Accepted 15 March 2016
Available online xxx

Keywords:
Rare diseases
Biomedical databases
Disease relations
Protein interactions

ABSTRACT

ODCs (Orphan Disease Connections), available at http://csbg.cnb.csic.es/odcs, is a novel resource to explore potential molecular relations between rare diseases. These molecular relations have been established through the integration of disease susceptibility genes and human protein-protein interactions. The database currently contains 54,941 relations between 3032 diseases.

Rare diseases are those diseases that affect a relatively limited number of individuals (no more than 5 out of 10,000 in the European Union). Due to their low prevalence, it is important to gather all the information available on these diseases and to look for relations between them. This could open new opportunities for transferring knowledge from one disease to the other (markers, targets, drugs, etc.) and for exploiting synergies between independent research lines.

Previous studies have used different approaches to establish molecular relations between diseases: common genes [1,2], microRNAs [3], functional linkages [4], protein localization [5], protein-protein interactions [6] or consecutive metabolic reactions [7]. In spite of the value of these molecular relations for a wide range of biomedical researchers and clinicians, only a few resources offer friendly graphical interfaces for users to search, retrieve or inspect a given disease-disease relation in detail. This is the case of MalaCards [8], a disease database that allows users to navigate from a disease to related diseases based on a variety of annotations (like genes, pathways, phenotypes, compounds and Gene Ontology terms). And DiseaseConnect [9], a resource focused on disease relations built by shared susceptibility genes and differential expression data.

Focusing on rare diseases, Zhang et al. constructed and studied global properties of several disease networks [2]. In the most populated network, two rare diseases were connected if they shared at least one susceptibility gene thus analysing at that time 2259 relations among 1170 diseases. But by using this shared-gene formalism they recognised that many rare disease relations could not be discovered. To overcome this limitation, they additionally constructed networks on the basis of enriched annotations: biological processes, cellular components, phenotypes and pathways. These networks were constructed for the subset of diseases with at least four known susceptibility genes. Thus they contained only a small fraction of the diseases in the gene-based network (up to 15%), as most rare diseases are monogenic.

In order to find additional potential molecular relations for the largest number of rare diseases, an alternative strategy is to link two rare diseases not only because they share genes, but because their associated gene products are interacting as well. This strategy was applied to study the topology and function of a global human disease network [6]. It is also sustained by previous studies that have reported a higher similarity of both symptoms [10] and comorbidities [11] among those diseases associated to interacting proteins than among those associated to proteins that do not interact.

In this work we present Orphan Disease Connections (ODCs), a resource of potential relations between rare diseases established by

* Corresponding author. Centro Nacional de Biotecnologia (CNB-CSIC), Darwin 3, 28049 Madrid, Spain.
E-mail address: monica.chagoyen@cnb.csic.es (M. Chagoyen).

http://dx.doi.org/10.1016/j.mcp.2016.03.004
0890-8508/

Please cite this article in press as: S. Fernandez-Novo, et al., Rare disease relations through common genes and protein interactions, Molecular and Cellular Probes (2016), http://dx.doi.org/10.1016/j.mcp.2016.03.004

shared susceptibility genes and protein interactions of the corresponding gene products. To the best of our knowledge there are no searchable public resources providing potential rare disease connections on the basis of protein-protein interactions. ODCs has a web interface, available at http://csbg.cnb.csic.es/odcs, to search for the diseases connected to one of interest, to explore in the detail the connections between two rare diseases, or to search for rare diseases associated with a given gene. This interface was designed with simplicity in mind, in a Google-like style so that filling a single text-box is enough to start navigating. Search results are presented both in text and in a graphical interactive way (network visualization), allow users to explore in detail the molecular connections and include links to external resources.

The first version of OCDs is built upon two main sources of data: rare disease susceptibility genes from Orphadata version 1.0.20. (Orphadata: Free access data from Orphanet© INSERM 1997. Available on http://www.orphadata.org), and human protein-protein interactions from HIPPIE [12]. HIPPIE is a resource of human protein interactions compiled from a large number of databases. In HIPPIE, each interaction is associated to a numeric score that reflects its reliability. Therefore, a HIPPIE score is a good estimator of the amount and quality of the experimental evidence for a given interaction.

Additionally OCDs contains link-outs to conveniently navigate to other important biomedical resources like the International Classification of Diseases (ICD-10), OMIM human genes and reported clinical cases [13], UniProt protein sequences and functions [14] and MeSH controlled vocabulary (www.nlm.nih.gov/mesh).

Two rare diseases are connected in OCDs if they share a susceptibility gene or if there is at least an interaction between the proteins encoded by their susceptibility genes. With this approach, a much larger number of connections than those based only in shared genes could be established, as most rare diseases (73%) are associated with only one gene. The current version contains 54,941 relations among 3032 diseases, of which 5263 relations correspond to shared-genes.

To assess the value of our approach, we have calculated the phenotypic similarity of all pairs of rare diseases for the subset of diseases where phenotypic information as well as susceptibility genes are available. The phenotypic similarity between two diseases was calculated as follows. First, phenotypes associated to rare diseases were compiled from the Human Phenotype Ontology (HPO) [15] and direct disease-phenotype associations were expanded to include the parent terms of a given phenotype in the HPO hierarchy. Second, we calculated the probability of each phenotype term in the ontology $p(c)$, as the number of diseases associated with it, divided by the total number of diseases. Finally the phenotypic similarity between two diseases was defined based on the probability of the most specific common phenotype [16]:

$$sim(d1, d2) = \max_{c \in S(c1, c2)} [-\log p(c)]$$

where $c1$ and $c2$ are all the phenotype terms associated to diseases $d1$ and $d2$ respectively.

Then we classified disease pairs in five groups, those that share: (i) genes, (ii) protein interactions, (iii) molecular pathways, (iv) protein complexes, and (v) the remaining pairs (not sharing any of the previous). In order to build this classification human protein interactions were compiled from HIPPIE [12], pathways from Reactome [17] and protein complexes from CORUM [18].

Fig. 1 shows the distribution of phenotypic similarity of each of the five groups. As expected, diseases sharing genes have more similar phenotypes (mean value 1.38), followed by those sharing interactions (0.93), pathways (0.84), complexes (0.64) and the remaining pairs (0.39).

According to these results, apart from common genes, the approach that yields the highest proportion of significant disease relations for the largest set of diseases is the one based on protein interactions. This justifies the development of resources providing potential rare disease associations through protein interactions such as this one.

Three types of searches can be performed in ODCs: retrieving all the relations of a rare disease (Rare disease search), exploring in detail the potential relation between two rare diseases (Diseases connection search), or retrieving the rare diseases associated with a given gene (Gene search). To facilitate searches, the input text-boxes include pull down menus that show a list of matching diseases or genes as the user types 3 characters or more. Each of the three searches provides an alternative view on the relations: disease-centric, disease-disease connection centric and gene-centric. They all offer a graphical visualization of the relations found (disease-gene, disease-disease and gene-gene) in the form of networks by means of Cytoscape Web [19]. Views provide as well detailed textual information and links to external resources for navigation. Finally, users can download the relations found with an export utility.

A rare disease search takes to a disease-centric view. In this view, the query disease is shown together with all the related diseases, both graphically in an interactive network, as well as in a list. In this list ODCs first ranks relations based on shared genes (by number), followed by those based on protein interactions (by the sum of HIPPIE scores). The type of relation (by common genes or by interacting proteins) is also differently represented in the graph. Additional disease information is also provided: disease classification (according to Orphanet and ICD-10), epidemiological data, signs and symptoms, and associated genes (with links to UniProt and OMIM databases). From this page, the user can explore the details of a particular disease-disease relation.

When a user searches for a potential relation between two diseases ODCs shows a graph with all the genes associated to the two diseases, as well as their protein product interactions (if any). Shared genes are marked (both in the network and in the text) and HIPPIE interactions scores are reported. This view also shows, in textual format, other information for both diseases: their nomenclature, classification and epidemiological data, as well as their common and distinct signs and symptoms. ODCs offers two additional utilities to help users in judging the relevance of a disease-disease relation. The first consists of link-outs to the HIPPIE web site that show the experimental evidences supporting each protein-protein interaction. The second is a PubMed search builder that allows users to verify whether different associations among diseases and interacting gene products have been reported in the literature (by combining disease and gene terms through the provided checkboxes).

As an example, a detailed graphical view on the connections among Noonan and Distal 22q11.2 microdeletion syndromes is shown in Fig. 2. The relation between these two syndromes is supported by the fact that the ERK MAP kinase signalling pathway is disrupted in both developmental syndromes [20]. This relation, established in ODCs through protein interactions, could not be found based solely on shared genes.

Finally, a gene search takes to a gene-centric view. In this case, all the diseases associated to this gene are shown, together with the genes whose protein products are known to interact with its product. From here, the user can navigate to a given disease ('rare disease view') or to another gene view.

Establishing and studying molecular relations between diseases is a very active area of research in systems medicine. Looking for potential relations between diseases is especially important in the

Please cite this article in press as: S. Fernandez-Novo, et al., Rare disease relations through common genes and protein interactions, Molecular and Cellular Probes (2016), http://dx.doi.org/10.1016/j.mcp.2016.03.004

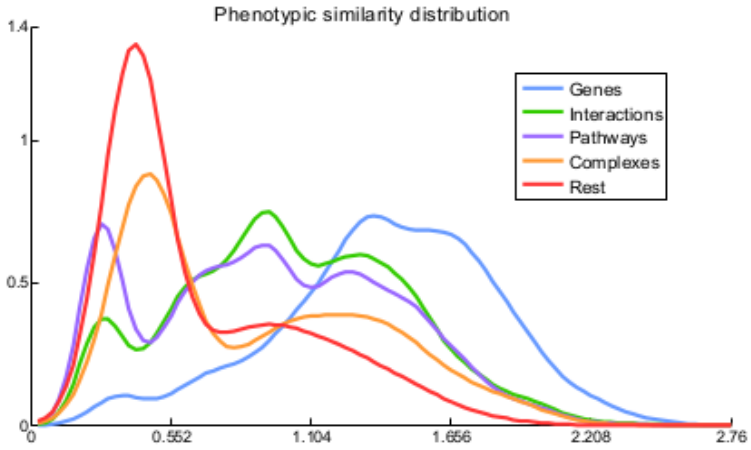

Fig. 1. Phenotypic similarity distribution of rare disease pairs corresponding to those sharing genes, interactions, pathways, complexes and remaining pairs.

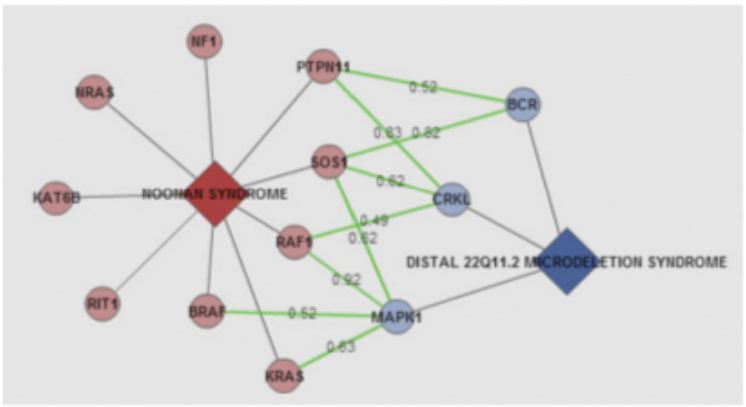

Fig. 2. A disease connection generated by ODCs showing the molecular associations between Noonan syndrome and Distal 22q11.2 microdeletion syndrome (red and blue diamonds). This connection is established based on the protein interactions among their susceptibility genes (circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

case of rare diseases due to their low prevalence. ODCs is a new resource which allows biomedical researchers and clinicians to easily search and explore potential molecular relations between rare diseases. ODCs relations are based on common susceptibility genes and protein-protein interactions. Although the human interactome is still incomplete, its current coverage enables the study of disease mechanisms and disease relations at a system level [21,22]. Moreover, it is desirable to increase the coverage of disease associations beyond those based on shared genes, as most rare diseases are monogenic, even at the expenses of losing some accuracy. According to our analysis on phenotypic similarities, rare disease associations built on protein interactions provide the best trade-off compared to those built on common pathways or complexes. The ODCs interface allows the user to interactively inspect the reported relations and the protein interactions they are based on. In this way, the expected physiological and clinical relevance of the disease connections can always be judged based on the experimental evidences supporting them as well as on expert knowledge.

ODCs includes three easy-to-use types of searches (disease, diseases connection and gene), provides both textual and graphical output, and numerous link-outs to related biomedical resources. This new resource on rare disease relations can help biomedical investigators to build novel hypothesis and start new collaborations joining otherwise separated research efforts.

We thank the members of the Computational Systems Biology Group (CNB-CSIC), and David San León (CNB-CSIC) for useful comments and suggestions. Special thanks to Martin Schaefer (CRG) for providing support in building link-outs to HIPPIE web site.

References

[1] K.I. Goh, et al., The human disease network, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8685–8690.
[2] M. Zhang, C. Zhu, A. Jacomy, L.J. Lu, A.G. Jegga, The orphan disease networks, Am. J. Hum. Genet. 88 (2011) 755–766.
[3] M. Lu, et al., An analysis of human microRNA and disease associations, PLoS

Please cite this article in press as: S. Fernandez-Novo, et al., Rare disease relations through common genes and protein interactions, Molecular and Cellular Probes (2016), http://dx.doi.org/10.1016/j.mcp.2016.03.004

One 3 (2008) e3420.
[4] B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. Delisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biol. 10 (2009) R91.
[5] S. Park, et al., Protein localization as a principal feature of the etiology and comorbidity of genetic diseases, Mol. Syst. Biol. 7 (2011) 494.
[6] X. Zhang, et al., The expanded human disease network combining protein-protein interaction information, Eur. J. Hum. Genet. 19 (2011) 783–788.
[7] D.S. Lee, et al., The implications of human metabolic network topology for disease comorbidity, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 9880–9885.
[8] N. Rappaport, et al., MalaCards: an integrated compendium for diseases and their annotation, Database (Oxford) 2013 (2013) bat018.
[9] C.C. Liu, et al., DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections, Nucleic Acids Res. 42 (2014) W137–W146.
[10] X. Zhou, J. Menche, A.L. Barabasi, A. Sharma, Human symptoms-disease network, Nat. Commun. 5 (2014) 4212.
[11] J. Park, D.S. Lee, N.A. Christakis, A.L. Barabasi, The impact of cellular networks on disease comorbidity, Mol. Syst. Biol. 5 (2009) 262.
[12] M.H. Schaefer, et al., HIPPIE: integrating protein interaction networks with experiment based quality scores, PLoS One 7 (2012) e31826.
[13] J.S. Amberger, C.A. Bocchini, F. Schiettecatte, A.F. Scott, A. Hamosh, OMIM.org: online mendelian inheritance in man (OMIM(R)), an online catalog of human genes and genetic disorders, Nucleic Acids Res. 43 (2015) D789–D798.
[14] C. UniProt, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.
[15] S. Kohler, et al., The human phenotype ontology project: linking molecular biology and disease through phenotype data, Nucleic Acids Res. 42 (2014) D966–D974.
[16] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1999) 95–130.
[17] D. Croft, et al., The Reactome pathway knowledgebase, Nucleic Acids Res. 42 (2014) D472–D477.
[18] A. Ruepp, et al., CORUM: the comprehensive resource of mammalian protein complexes–2009, Nucleic Acids Res. 38 (2010) D497–D501.
[19] C.T. Lopes, et al., Cytoscape web: an interactive web-based network browser, Bioinformatics 26 (2010) 2347–2348.
[20] I.S. Samuels, S.C. Saitta, G.E. Landreth, MAP'ing CNS development and cognition: an ERKsome process, Neuron 61 (2009) 160–167.
[21] J. Menche, et al., Disease networks. uncovering disease-disease relationships through the incomplete interactome, Science 347 (2015) 1257601.
[22] **M. Chagoyen, F. Pazos, Characterization of clinical signs in the human interactome, Bioinformatics (2016),** http://dx.doi.org/10.1093/bioinformatics/btw054.

Please cite this article in press as: S. Fernandez-Novo, et al., Rare disease relations through common genes and protein interactions, Molecular and Cellular Probes (2016), http://dx.doi.org/10.1016/j.mcp.2016.03.004

B – Carta del director del Programa de Biología de Sistemas (CNB-CSIC)

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD

CSIC
Centro Nacional
de Biotecnología

Don Víctor de Lorenzo Prieto, como Director del Programa de Biología de Sistemas del Centro Nacional de Biotecnología,

CERTIFICO que Dña. Sara Fernández Novo ha realizado el Proyecto Fin de Carrera en mi departamento bajo la supervisión de Dña. Mónica Chagoyen Quiles, Doctora en Ingeniería Informática y de Telecomunicación.

El principal resultado de este proyecto ha sido el desarrollo de una novedosa aplicación informática, Orphan Disease Connections (ODCs), que permite la consulta de potenciales relaciones moleculares entre enfermedades raras. ODCs está disponible públicamente a través de uno de los servidores de nuestro centro (en la dirección http://csbg.cnb.csic.es/odcs). La aplicación es de interés, no sólo para los investigadores de nuestro departamento, sino para todos aquellos que aborden el estudio de la enfermedad desde la incipiente perspectiva sistémica.

En Madrid a 3 de mayo de 2016

Víctor de Lorenzo Prieto

V. DE LORENZO (CSIC)
CENTRO NACIONAL DE BIOTECNOLOGIA
CAMPUS DE CANTOBLANCO
28049 MADRID (ESPAÑA)

c/ Darwin, 3
Campus Universidad
Autónoma de Madrid
Cantoblanco
28049 MADRID
TEL.: 91 5854500
FAX: 91 5854506

vdlorenzo@cnb.csic.es

C – Presupuesto

1. Ejecución material

Compra de ordenador personal (software incluido)	1500 €
Material de oficina	50 €
Total de ejecución material	1550 €

2. Gastos generales

21% sobre Ejecución material	325,50 €

3. Beneficio industrial

6% sobre Ejecución material	93€

4. Honorarios proyecto

1000 horas a 15 €/hora	15000 €

5. Material fungible

Gastos de impresión	100 €
Encuadernación	50 €

6. Subtotal del presupuesto

Subtotal presupuesto	17118,50 €

7. I.V.A. aplicable

21% sobre Subtotal del presupuesto	3594,90 €

8. Total del presupuesto

Total del presupuesto	20713,40 €

Madrid, mayo de 2016
La Ingeniera Jefe de Proyecto

Fdo.: Sara Fernández Novo
Ingeniera de Telecomunicación

D – Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, del desarrollo de un sistema para la integración de datos moleculares sobre enfermedades raras. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados.

Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata, pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa, y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere, y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras, así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra se girarán visitas de inspección por personal facultativo de la empresa cliente para hacer las comprobaciones que se crean oportunas. Es obligación del contratista la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa por retraso de la ejecución, siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él deberá ser notificada al Ingeniero Director del Proyecto, y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.