Enrich_Gen:

Plataforma web para el enriquecimiento clínico y farmacológico de variantes de genes

Nombre Estudiante
Mauro Javier Oruezábal Moreno

Nombre Consultor/a
Romina Astrid Rebrij

Nombre Profesor/a responsable de la asignatura
David Merino Arranz

Fecha Entrega: 16 de junio de 2018

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Enrich_Gen:* *Plataforma web para el enriquecimiento clínico y farmacológico de variantes de genes*
Nombre del autor:	*Mauro Javier Oruezábal Moreno*
Nombre del consultor/a:	*Romina Astrid Rebrij*
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	16/2018
Titulación:	Master Universitario en Bioinformática y Bioestadística
Área del Trabajo Final:	*Computación e Inteligencia Artificial en problemas biológicos y clínicos*
Idioma del trabajo:	*español*
Palabras clave	*Data mining, Análisis semántico latente, Descomposición Valores Singulares*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Hoy día la aplicación clínica de un estudio genómico se limita al análisis de un número muy pequeño de genes y muy dirigidos hacía la patología que tiene el paciente, debido a que la interpretación de múltiples mutaciones en un mismo estudio es muy completa y a veces imposible. El gran desconocimiento sobre las relaciones entre genes y el fenotipo-genotipo, está provocando que a pesar de disponer de la tecnología para secuenciar múltiples genes e incluso el exoma completo, su utilidad clínica sea muy limitada.

La plataforma *Enrich Gen* permite, mediante la integración de la información procedente de diversas bases de datos y la visualización en tablas o gráficos, priorizar los genes más relevantes de un análisis; conocer la similitud o distancia entre los genes; disponer del listado de fármacos con actividad para cada gen; conocer los ensayos clínicos desarrollados por gen, y por último, evaluar la evidencia científica de cada gen.

Los tres procedimientos fundamentales de la plataforma son: los metadatos, el análisis semántico latente basado en la descomposición de una matriz en sus valores singulares; y el análisis semántico basado en la ontología. El primero, los metadados, se caracterizan por ser datos altamente estructurados debido a las reglas incluidas para su extracción, clasificación y adaptación a la estructura de visualización que permite ganar en eficiencia y/o mejorar la interpretación; el análisis semántico latente basado en la descomposición en valores singulares es usado para la construcción de un gráfico nodo-arco a partir de una busqueda en la plataforma PubMed; y el análisis semántico basado en la ontología, se utiliza para evaluar la similitud de genes basados en los términos GO.

Enrich Gen es una herramienta web que permite al usuario extraer información clínica,

farmacológica, biológica, y facilita la anotación de un conjunto de genes tras conocer su similitud basada en los términos GO, por lo tanto, es de gran utilidad en la práctica clínica.

Abstract (in English, 250 words or less):

Nowadays the clinical application of a genomic study is limited to analysis of a very small number of genes and very directed towards patient's disease, because the interpretation of multiple mutations in the same study is very complicated and sometimes impossible. The great ignorance about the relationships between genes and phenotype-genotype, are causing that in spite of having the technology to sequence multiple genes and even the whole exome, its clinical usefulness is very low.

The Enrich Gen platform allows, through the integration of information from various databases and visualization in tables or graphs, to prioritize the most relevant genes in an analysis; to know the similarity or distance between the genes; to dispose a list of drugs with activity for each gene; to know the clinical trials available by gene, and finally, evaluate the scientific evidence of each gene.

The three fundamental procedures of the platform are: the metadata, the latent semantic analysis based on the decomposition of a matrix in its singular values; and semantic analysis based on ontology. The first, metadata, is characterized by highly structured data due to the rules included for its extraction, classification and adaptation to the visualization structure that allows to gain in efficiency and / or improve interpretation; the latent semantic analysis based on the decomposition in singular values is used for the construction of a node-arc graph from a search in the PubMed platform; and semantic analysis based on ontology, is used to evaluate the similarity of genes based on the terms GO.

Enrich Gen is a web tool that allows user to extract clinical, pharmacological, biological information, and facilitates the annotation of a set of genes after knowing their similarity based on the GO terms, therefore, it is very useful in clinical practice.

Índice

1. Introducción

1.1 Contexto y justificación del Trabajo

La traducción de la secuencia del genoma en información médicamente procesable es un desafío clave, pero nos encontramos con distintos problemas que hasta ahora no se han resuelto y precisan de nuevos abordajes.

Por una parte, existe una percepción errónea sobre el valor diagnóstico de la secuenciación del genoma y que resulta de una simplificación excesiva en la que se supone que hay un "gen" para la enfermedad, cuando en realidad lo más probable es que intervengan múltiples genes que confieren un riesgo aumentado, que puede ser sutil individualmente, y que se manifiesta sólo en el contexto de antecedentes genéticos específicos o una determinada exposición ambiental.

Por otra parte, también existe el riesgo de una interpretación excesiva incluso para mutaciones con efectos aparentemente grandes. Para los pacientes afectados en los que existe una gran probabilidad previa de que la mutación genética sea causal debido a una historia familiar positiva y/o un fenotipo clínico específico, la interpretación puede ser directa. Sin embargo, si las mutaciones no son totalmente penetrantes, habrá portadores en la población que estén sanos. Gran parte de nuestro conocimiento sobre la penetración de mutaciones hasta la fecha se basa en datos familiares y, por lo tanto, adolece de un sesgo de determinación. Sin un conocimiento imparcial del efecto de las mutaciones, la interpretación a nivel de la población será intrínsecamente problemática. Si bien las políticas para restringir las pruebas genéticas a las poblaciones de alto riesgo fueron impulsadas inicialmente por restricciones presupuestarias, y la disponibilidad más generalizada de las pruebas se considera una ventaja de la disminución de los costos, otra consecuencia es que la interpretación de la importancia clínica de una mutación es mucho más difícil si se encuentra sin el sesgo de evaluación mencionado anteriormente. Es decir, predecir los efectos de una nueva mutación BRCA2 en el contexto de una fuerte historia familiar de la mutación que se segrega con la enfermedad en la familia, es mucho más fácil que cuando se descubre en la población general.
Sin el apoyo de la segregación en las familias, la asignación de la patogenicidad puede ser problemática; por ejemplo, las grandes duplicaciones, la mayoría de las mutaciones sinónimos y algunas de sentido erróneo, variantes intrónicas, y la mayoría de las variantes en el promotor son particularmente difíciles de interpretar. Predecir las consecuencias funcionales de las variantes que alteran la secuencia de codificación de proteínas también puede ser un desafío. Una variante puede afectar un sitio de unión del factor de transcripción, un sitio objetivo de microARN, afectar el empalme o estabilidad del ARN o truncar una proteína. Finalmente, la cuestión del desequilibrio de ligamiento (donde las variantes benignas se encuentran cerca de una variante que predispone a la enfermedad) puede complicar la interpretación de las variantes de riesgo recurrentes [1].

A todo esto hay que añadir que independientemente de si los modelos animales pueden imitar adecuadamente la enfermedad humana, tales sistemas modelo son inherentemente inadecuados para determinar las consecuencias de las mutaciones específicas como actividad rutinaria. Si bien los sistemas de levadura y línea celular

pueden usarse para evaluar la funcionalidad de las mutaciones del gen de reparación del ADN, la aplicabilidad general de tales sistemas modelo es limitada. En vista de estos factores, se confía cada vez más en la implementación de herramientas in silico para inferir las consecuencias funcionales de las mutaciones. Aunque tales algoritmos pueden ayudar a predecir la posible patogenicidad de las variantes, a menudo diferentes herramientas concluyen en direcciones opuestas y sin una relación establecida entre la disfunción génica y el fenotipo de la enfermedad, por lo que la predicción de riesgo es problemática [2].

Otra iniciativa ha sido catalogar y asignar la patogenicidad a variantes / mutaciones en varios genes específicos. Ejemplos de tales bases de datos incluyen InSiGHT (Sociedad Internacional de Tumores Hereditarios Gastrointestinales Incorporated), LoVd (base de datos de variantes abiertas de Leiden) Decipher y DMuDB (la base de datos de mutaciones de diagnóstico) y el locus Reference Genomic Colaboración. Estos recursos brindan a los profesionales de la salud información valiosa para los procesos de toma de decisiones. Si bien los informes publicados son fuentes valiosas para tales bases de datos, su administración depende en gran medida de la presentación de variantes individuales y los datos clínicos patológicos asociados mediante la secuenciación de los laboratorios. De este modo, actualmente estas bases de datos están limitadas a la conservación de un número restringido de genes. Incluso aquí traducir la secuencia genómica en información médicamente procesable puede consumir mucho tiempo.

En conclusión, para satisfacer las necesidades futuras, se deben desarrollar y mantener recursos integrales con un alcance mucho más amplio. Nuestra hipótesis se basa en que para investigar los genes relacionados con el cáncer, se puede investigar la literatura que contiene una gran cantidad de información, en forma de artículos científicos, no facilmente relacionable. Sin embargo, una consulta bruta de un término o varios términos indexados a PubMed puede recuperar más de 250,000 artículos, lo que hace imposible obtener una visión completa, real e interpretable al leerlos. La tendencia es que el número de artículos de PubMed está aumentando constantemente, y también lo están los artículos sobre el cáncer que mencionan nombres de genes argumentados como posibles biomarcadores. Por lo tanto, el uso de técnicas de minería de textos para recopilar nuevos conocimientos de muchas fuentes científicas existentes y la utilización del análisis de la Semántica Latente o Latent Semantic Analysis (LSA) que aprovecha un fenómeno que se suele cumplir en el lenguaje natural: las palabras del mismo campo semántico suelen aparecer juntas o en similares contextos, puede ser una forma efectiva de investigar la literatura sobre la utilidad clínica de las variantes encontradas por secuenciación genómica [3,4].

1.2 Objetivos del Trabajo

El objetivo fundamental es desarrollar una aplicación web que permita conocer las relaciones genes-procesos biológicos, genes-fármacos, genes-fenotipos y genes-genes, con el fin último de determinar la utilidad clínica de las variantes encontradas por secuenciación genómica.

Una ventaja de esta herramienta es que puede usarse para la interpretación de paneles genéticos realizados a pacientes con cáncer y poder interpretar mutaciones en genes cuya evidencia cientifica basadas en las recomendaciones internacionales es

pequeña, pero que son muy frecuentes y determina el alto porcentaje de estudios genéticos sin interpretar y la dificultad para implantar la medicina de precisión en la práctica clínica [10].

En este sentido, se construirán diversas redes a partir de las relaciones establecidas (genes-procesos biológicos, genes-fármacos, genes-fenotipos y genes-genes), en base a los artículos publicados y la extracción de términos de cada artículo y distribuidos en una matriz de ocurrencias.

1.2.1 OBJETIVOS GENERALES

1. Realizar una herramienta web que permita generar mapas de conocimiento de los genes implicados en cada tipo de cáncer basado en la integración de datos biológicos y clínicos.

2. Visualización gráfica de la relación de similitud de los genes implicados en cada tipo de cáncer, basado en los términos ontológicos de cada gen (GO) que aportan información sobre su componente celular, función molecular y procesos biológicos implicados.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Obtener una descripción general de los genes y relacionarlos con un determinado tipo de tumor, los fármacos aprobados y/o en desarrollo clínico y la relación de similitud de los distintos genes descritos.

2. Disponer de un listado de los procesos biológicos en los que están implicados cada gen.

3. Priorizar aquellos genes mutados con una mayor implicación en el desarrollo de la enfermedad.

4. Ayudar a seleccionar uno o varios fármacos basados en los genes implicados en cada tumor.

5. Identificar fácilmente información útil y pertinente.

6. Identificar conceptos relevantes.

1.3 Enfoque y método seguido

Las pruebas genéticas se han restringido tradicionalmente al análisis de un pequeño número de genes generalmente seleccionados sobre la base de alta probabilidad de ser mutado. Sin embargo, este enfoque como hemos visto tiene varias limitaciones. En primer lugar, muchas enfermedades hereditarias son genéticamente heterogéneas y el análisis mutacional secuencial de genes individuales es lento y costoso. En segundo lugar, aunque los subconjuntos de algunas enfermedades comunes pueden ser causados por mutaciones en un único gen, los métodos tradicionales para seleccionar a quién someter a prueba en función de las características de la

enfermedad o antecedentes familiares son sesgados y tienen una alta tasa de falsos negativos.
Por otra parte, los recursos tradicionales que están disponibles para recuperar información funcional del NCBI (Centro Nacional de Información Biotecnológica, https://www.ncbi.nlm.nih.gov/gene/), están diseñados típicamente para un solo gen. Se ha creado una nueva generación de recursos para facilitar la recuperación de información por lotes para conjuntos de genes. Un ejemplo de esto es BIOMART (https://www.ensembl.org/biomart/martview/), en el cual los usuarios pueden realizar una búsqueda y recuperación de información del genoma para conjuntos de genes. BIOMART cubre un amplio espectro de información funcional perteneciente a atributos específicos de genes y proteínas, así como a enfermedades, expresiones, variaciones de secuencia y atributos de especies cruzadas. A pesar de ser una excelente herramienta de recuperación de información por lotes, BIOMART no ayuda a los profesionales sanitarios a explorar de manera eficiente la abundante información asociada con un conjunto de genes y poder tomar decisiones relevantes para los pacientes [5].

Una forma que puede ayudar a explorar conjuntos genéticos grandes es organizar los genes basándose en características funcionales comunes, tales como categorías de Ontología Genética (GO) o vías bioquímicas. Se han desarrollado varias herramientas bioinformáticas para organizar conjuntos de genes basados en GO. La mayoría de estas herramientas también han implementado pruebas estadísticas para identificar categorías enriquecidas de GO y para sugerir las áreas biológicas más importantes asociadas con un determinado conjunto de genes. Aunque el uso de métodos ontológicos para estructurar el conocimiento biológico es un área activa de investigación y desarrollo, el cuerpo del conocimiento biológico asociado con cualquier conjunto de genes se extiende mucho más allá de GO. Además de organizar conjuntos de genes dentro del contexto de GO, MAPPFinder, DAVID y GFINDer ofrecen la opción de organizar y visualizar conjuntos de genes en el contexto de las vías bioquímicas KEGG (Enciclopedia de Kioto de Genomas y Genomas, http: //www.genome.ad.jp/kegg). DAVID y GFINDer también pueden organizar conjuntos de genes basados en información de dominio de proteínas. Otras características, como la ubicación cromosómica, el patrón de expresión tisular y la asociación en la publicación, también podrían usarse para organizar un conjunto de genes. Sin embargo, estas características no se implementan en las herramientas actuales de análisis de conjuntos de genes [6].

En conclusión, todos estos problemas hacen que los enfoques del análisis de la relación existente entre multiples genes analizados a partir de paneles genéticos o del análisis del genoma completo, y la integración de la relación gen-gen con gen-fenotipo, gen-proceso biológico o gen-fármacos, sean propuestas atractivas que permitirán conocer la utilidad clínica. Esta mejor comprensión de los procesos biológicos subyacentes y la similitud de un conjunto de genes constituye un gran desafío en el momento actual, pero esencial para poder implantar la medicina de precisión en la práctica clínica.

El presente trabajo propone un enfoque de integración de la información disponible en múltiples bases de datos, y presentación en tablas y gráficos que combinen la evidencia disponible en estudios científicos, ensayos clínicos, términos geontológicos

relacionados con procesos biológicos, función molecular, localización celular o diversas ómicas [7].

La idea subyacente del proyecto es que los genes no actúan de forma aislada sino que interactúan en redes génicas de señalización o regulación. Se propone, por todo ello, un procedimiento de extracción y procesamiento de la información de la literatura para identificar posibles relaciones entre genes, así como su asociación con el fenotipo, procesos biológicos y sensibilidad a fármacos, con el fin último de establecer la relevancia clínica [8,9].

Se utilizarán métodos visuales para mostrar la información que ayudará a la visualización de los datos y permitirá reconocer patrones y tendencias.

1.4 Planificación del Trabajo

El trabajo se estructuró siguiendo el planteamiento propuesto, de este modo se dividió en los siguientes apartados:

- Definición de contenidos: en esta etapa se definió la temática del trabajo, el título del proyecto, la problemática a resolver y los objetivos, después de una amplia revisión bibliográfica.

- Plan de trabajo: en el periodo que comprendió de marzo a mediados de mayo de 2018, se diseño lo que es la plataforma; se eligió el lenguaje de programación; se seleccionaron las librerías que contenían las funciones necesarias para el análisis; se dividió el proyecto en módulos en relación a las funcionalidades que la plataforma debía incorporar, y finalmente se depuraron los errores detectados durante la fase de prueba. En este sentido, estructuramos este trabajo en las siguientes etapas e identificamos los siguientes hitos:

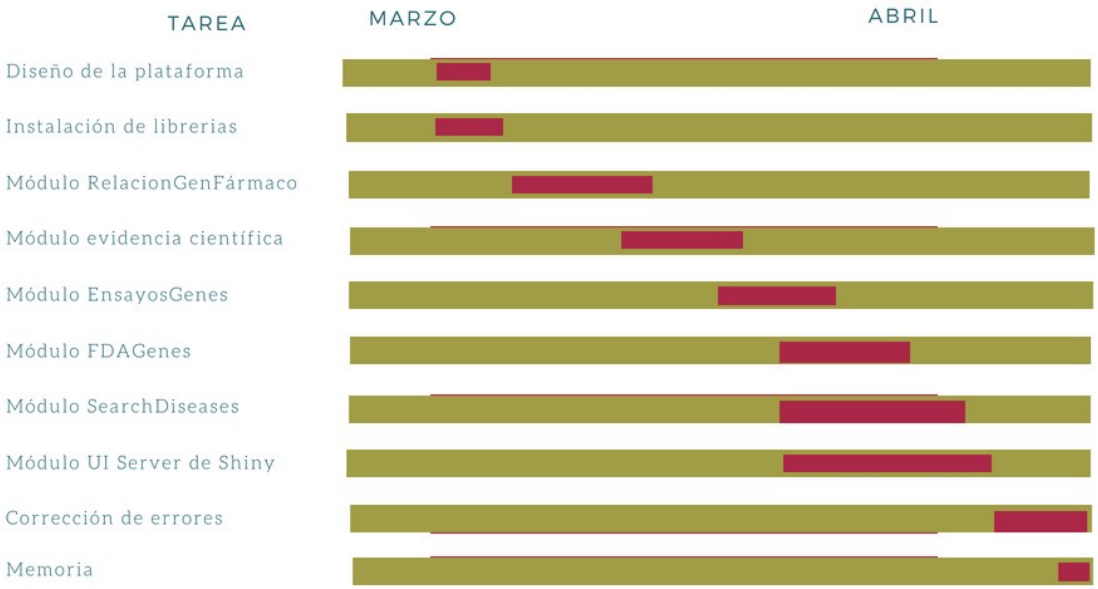
- Diseño de la plataforma: se proyectó un diseño sencillo y amigable que permitíera un rápido aprendizaje.
- Instalación de librerías: la busqueda de las librerias se basó en la necesidad de disponer de funciones y procedimientos matemáticas para la descomposición de una matriz en valores singulares, el análisis semántico latente, la busqueda de similitud de genes basado en su ontología, el diseño de gráficos de nodos y enlaces y la estructuración de la información en tablas con un diseño sencillo però atractivo que facilitara la visualización de los datos. Se utilizó también la librería shiny con el fin de crear una aplicación web interactiva desde R.
- Estructuración modular de la plataforma con los siguientes apartados:
    - Módulo de la Relación de Genes – Fármaco: el hito de este módulo fue disponer de una tabla resumen que mostrara el global de fármacos experimentados para cada uno de los genes. Se optó por incorporar sólo el código PMID dado que era una forma rápida de conocer la evidencia de cada fármaco.
    - Módulo evidencia científica: el objetivo fue disponer de una tabla resumen de las publicaciones de PubMed para cada gen en un periodo de 10 años y hasta el momento actual. Un hito fue disponer en esta tabla de un resumen del artículo que agilizará la revisión de la evidencia científica, lo cual se

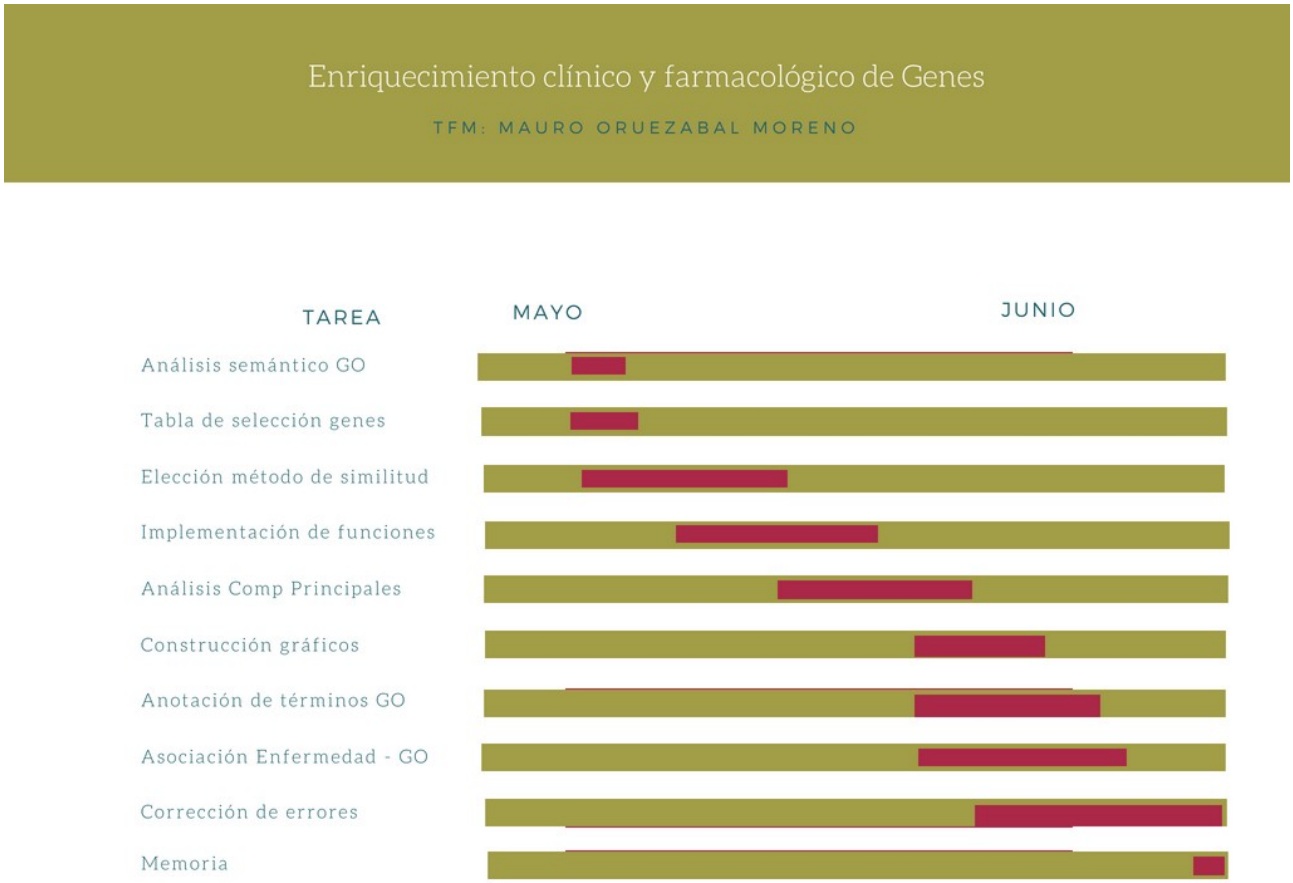
consiguió mediante la busqueda de palabras claves en el abstract de cada publicación.

- Módulo Ensayos-Genes: el objetivo fue estructurar la información disponible en la base de datos ClinicalTrials.org para cada gen y mostrar el título del artículo, la situación actual del ensayo clínico y el momento de apertura y cierre, dado que es una información que puede ayudar en el desarrollo posterior de un informe clinico.
- Módulo FDA-Genes: este módulo es la fase final de los anteriores dado que muestra sólo los fármacos indicados para cada gen, y es el que muestra la información con mayor evidencia científica dado que los fármacos que aparecen en esta tabla han sido aprobados por la agencia americana del medicamento (FDA).
- Módulo SearchDiseases: este módulo al final se estructuró en dos submódulos y fueron desarrollados en Mayo de 2018. Uno primero que fue la construcción de una red de genes por enfermedad; y otro que fue la representación gráfica de la similitud de genes basado en su ontología. La complejidad de este módulo fue la compresión del procedimiento matemático que permitiera el análisis, y la reflexión de cómo se debía estructurar los datos para que se pudiese realizar una representación mediante gráficos de nodos y enlaces.
- Módulo UI Server de Shiny: el objetivo final del proyecto fue la construcción de una plataforma web, al considerarlo como el mejor medio de presentación de la información y permitir la interacción con los usuarios, por lo que se utilizó la libreria Shiny de R, tanto para el diseño del interface del usuario como el servidor. Por otra parte, al estar estructurado el proyecto en módulos, se planificó el código en las diferentes partes que componen la plataforma. Se tuvo en cuenta los nombres de variables entre las entradas y salidas de unos u otros módulos, dado que había módulos embebidos unos en otros, y con el fin de evitar la colisión de variables.
- Corrección de errores: Una vez desarrollada la plataforma fueron descubiertos algunos errores en relación al número de valores que podía asignarse a cada variable. De este modo, hubo variables a las que inicialmente se les dió una relación uno a uno, pero al extraer la información de las bases de datos, la relación era distinta, por lo que estos errores fueron depurados. La identificación de todos los errores es compleja dado que estos se van detectando a medida que se va utilizando la plataforma, y puesto que la información extraida es enorme, y no siempre todas las variables tienen el mismo comportamiento en todas las ocaciones por el mismo tipo de datos extraido. Por ejemplo, en la bases de datos FDA, se comprobó que un medicamento podía tener distintos nombres comerciales por lo que se modificó el algoritmo de busqueda de información y almacenamiento.

- Redacción de la memoria y presentación: se siguió el esquema propuesto por la Universidad y se realizó un video que incorporó una imagen del alumno durante la lectura del proyecto.


TAREA
FEBRERO
MARZO
ABRIL
MAYO
JUNIO
Definición de contenidos
Plan de trabajo
Desarrollo del trabajo
Desarrollo del trabajo
Redacción de la memoria
Elaboración de la presentación
Defensa publica


TAREA
MARZO
ABRIL
Diseño de la plataforma
Instalación de librerias
Módulo RelacionGenFármaco
Módulo evidencia científica
Módulo EnsayosGenes
Módulo FDAGenes
Módulo SearchDiseases
Módulo UI Server de Shiny
Corrección de errores
Memoria



1.5 Breve sumario de productos obtenidos

La plataforma Enrich_Gen es el producto obtenido del trabajo de fin de Master cuya mayor utilidad es servir de herramienta para la interpretación de un panel genético o un análisis del exoma en pacientes diagnósticos de cáncer, dado que permite un enriquecimiento clínico y farmacológico de las variantes identificadas como patológicas.

Por otra parte, también posibilita la priorización de las variantes identificadas mediante el análisis de la similitud de genes basados en su ontología , lo cual nos dará una visión de que procesos biológicos y rutas metabólicas están alteradas. Así mismo, esta priorización también puede realizarse a nivel clínico mediante el conocimiento de los fármacos disponibles para cada gen y la evidencia científica que lo sustenta, lo cual es vital para poder aplicar los estudios genéticos en la práctica médica habitual y, por ende, el desarrollo de la medicina personalizada.

La plataforma Enrich_Gen es, por todo ello, muy útil para poder realizar informes de estudios genéticos que incorporen múltiples variantes patológicas, puesto que la interpretación es más objetiva en comparación a cómo se realiza actualmente, debido a que no sólo incorpora e integra una información extensisima y de distintas bases de datos, sino que además, permite el análisis semántico de esta información. A su vez, ayuda a la toma de decisiones en los pacientes puesto que clasifica y prioriza la información con el fin de mostrarla de una manera gráfica, visible y de fácil comprensión e interpretación.

1.6 Breve descripción de los otros capítulos de la memoria

En el siguiente capítulo se tratarán los siguientes puntos:

2. Metodología:

- 2.1. Fuente de información
- 2.2. Diseño de la plataforma
- 2.3. Instalación de las librerías utilizadas
- 2.4. Integración de bases de datos
- 2.5. Minería de datos mediante análisis de semántica latente
- 2.6. Similitud de genes basado en su ontología
    - 2.6.1 Términos de la Ontología de Genes
    - 2.6.2 Divisiones de la Ontología de Genes
    - 2.6.3 Estructura de las ontologías
    - 2.6.4 Anotaciones de la Ontología de Genes
    - 2.6.5 Análisis semántico basado en la Ontología de Genes
    - 2.6.6 Similitudes de Lin, Jiang y Conrath, y Resnik
    - 2.6.7 Basadas en la representación del DAG como GO–árbol
    - 2.6.8 Similitud entre Gen–Productos o proteínas
    - 2.6.9 Busqueda de enfermedades para un conjunto de genes
- 2.7. Servidor Shiny

3. Resultados

1. Visión general de la plataforma de enriquecimiento clínico y farmacológico de genes
2. Evidencia científica por genes
3. Relación genes y fármacos
4. Ensayos clínicos por genes
5. Indicación FDA por fármacos
6. Red de genes por enfermedad
7. Similitud de genes
    - 7.1 Panel de Genes
    - 7.2 Similitud de genes
    - 7.3 Ontología Génica
    - 7.4 Enfermedades asociadas
8. Ejemplo: secuenciación del exoma completo (WES) de un cáncer de mama metastásico. Informe realizado utilizando Enrich_Gen.

2. Metodología

2.1 Fuente de información

Para construir el mapa de conocimiento, se utilizará información de las siguientes bases de datos:

- PubMed (https://www.ncbi.nlm.nih.gov/pubmed/)
- Clinical Trials (https://clinicaltrials.gov/)
- FDA.gov (https://www.fda.gov/)
- Clearity Foundation (http://www.clearityfoundation.org/)
- CenterWatch's Drugs in Clinical Trials Database (https://drugs.centerwatch.com/).
- Guide To Pharmacology (IUPHAR/BPS) (http://www.guidetopharmacology.org/)
- My Cancer Genome (https://www.mycancergenome.org/)
- Therapeutic Target Database (http://bidd.nus.edu.sg/group/cjttd/)
- Memorial Sloan Kettering Cancer Center (https://www.mskcc.org/)
- GO Ontology (http://www.geneontology.org/)

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 ofrecida por la Biblioteca Nacional de Medicina de los Estados Unidos como parte de Entrez. MEDLINE tiene alrededor de 4800 revistas publicadas en Estados Unidos y en más de 70 países de todo el mundo desde 1966 hasta la actualidad. PMID, es el acrónimo de «PubMed Identifier» o «PubMed Unique Identifier», el cual es un número único asignado a cada cita de un artículo de revistas biomédicas y de ciencias de la vida que recoge PubMed.

Clinical Trials: CinicalTrials.gov (CT) es un sitio web desarrollado por U.S. National Institutes of Health (NIH) que ofrece regularmente información actualizada sobre ensayos clínicos y estudios observacionales. Actualmente su base de datos la forman más de 96 700 estudios de 174 países.

FDA.gov: La FDA (Food and Drug Administration: Administración de Medicamentos y Alimentos o Administración de Alimentos y Medicamentos) es la agencia del gobierno de los Estados Unidos responsable de la regulación de alimentos (tanto para personas como para animales), medicamentos (humanos y veterinarios), cosméticos, aparatos médicos (humanos y animales), productos biológicos y derivados sanguíneos. La base de datos de la FDA dispone de toda la documentación de medicamentos o productos biológicos aprobados en seres humanos y animales.

Clearity Foundation: The Clearity Foundation es una organización sin fines de lucro que ofrece servicios de asistencia al paciente, incluida la coordinación de pruebas de laboratorio, la interpretación de estudios genéticos y la identificación de ensayos

clínicos. Dispone de una amplia base de datos con información de fármacos en ensayo clínico o fase preclínica.

CenterWatch's Drugs in Clinical Trials Database: La base de datos de Drugs in Clinical Trials contiene más de 5.000 nuevos tratamientos en investigación, en ensayos de Fase I a Fase IV en todo el mundo. Se actualiza semanalmente y los perfiles de los medicamentos incluyen indicaciones de uso, iniciaciones de prueba actuales y resultados, estado de la fase de estudio e información de contacto con el fabricante.

Guide To Pharmacology (IUPHAR/BPS): La Guía IUPHAR / BPS de farmacología es un sitio web de acceso abierto que actúa como un portal de información sobre los objetivos biológicos de los medicamentos con licencia y otras moléculas pequeñas. La Guía de farmacología (con GtoPdb como abreviatura estándar) se desarrolla como una empresa conjunta entre la Unión Internacional de Farmacología Básica y Clínica (IUPHAR) y la Sociedad Farmacológica Británica (BPS).

My Cancer Genome: My Cancer Genome es un recurso de precisión sobre el conocimiento de la genética del cáncer orientado a médicos, pacientes, cuidadores e investigadores. My Cancer Genome brinda información actualizada sobre las mutaciones relacionadas con cáncer y las implicaciones terapéuticas, incluidos los ensayos clínicos disponibles.

Therapeutic Target Database: Therapeutic Target Database (TTD) es una base de datos proporcionada por Bioinformatics and Drug Design Group en la Universidad Nacional de Singapur e Innovative Drug Research and Bioinformatics Group (IDRB) en la Universidad de Zhejiang, que proporciona información sobre proteínas y nucleótidos con implicaciones terapéuticas. También se incluyen en esta base de datos enlaces a bases de datos relevantes que contienen información sobre su función, secuencia, estructura tridimensional, propiedades de unión a ligando, nomenclatura enzimática y estructura de fármacos, clase terapéutica y estado de desarrollo clínico.

Memorial Sloan Kettering Cancer Center: MSKCC es uno de los mayores centros oncológicos del mundo y disponen de una base de datos pública con información procedente tanto de práctica clínica habitual como de ensayos clínicos.

GO Ontology: El proyecto Ontología Génica (en inglés Gene Ontology cuya sigla es GO) provee un vocabulario controlado que describe el gen y los atributos del producto génico en cualquier organismo. En realidad son tres ontologías, cada cual representando un concepto clave en biología molecular: la función molecular de los productos génicos; su rol en los procesos biológicos de múltiples direcciones; y su localización en componentes celulares. Las ontologías son continuamente actualizadas, y se dispone de nuevas versiones mensualmente.

2.2 Diseño de la plataforma

El diseño de la página es muy simple con el fin de que pueda ser muy intuitivo y de fácil aprendizaje.
Los campos se organizan en tres partes diferenciadas: la cabecera donde se sitúa el título de la plataforma (“Enriquecimiento clínico y farmacológico de genes”) y el logo de la plataforma (“EnrichGen”); la columna lateral izquierda que contiene el menú de navegación; y la parte central donde se muestran los resultados.
El menú de navegación lateral presenta las siguientes posibilidades de búsqueda:

1. Un selector sobre el tipo de información que se quiere cargar en la página central y que es:
    1. Evidencia científica de genes
    2. Relación Genes-Fármacos
    3. Ensayos clínicos por genes
    4. Autorizaciones de la FDA por genes
    5. Red de genes basado en el cálculo sobre una matriz de coincidencia de genes.
    6. Similitud de genes basado en su geontología.
    7. Filtrado de Genes desde panel
2. Las cuatro primeras opciones permiten hacer la búsqueda por los genes que son mostrados en un desplegable. Por defecto se cargará un panel de 460 genes pero el número es ilimitado dado que la búsqueda en esta opción es por cada gen individualmente.
3. La última opción, “Filtrado de Genes desde panel”, muestra un selector de un archivo de texto cuya utilidad es poder subir un panel de genes y hacer una búsqueda masiva para todo el panel. Hay que tener en cuenta que el tiempo de ejecución aumenta de manera logarítmica al incrementar el tamaño del panel, por lo que se recomienda que el archivo cargado no sea excesivamente grande. El archivo de texto debe contener el nombre de los genes como símbolo o abreviatura según el Comité de Nomenclatura de Genes de HUGO y que se puede encontrar en el campo “Gene” de la base de datos NCBI, y separado por comas.

Lo fundamental de esta plataforma creada con el paquete Shiny es que es reactiva lo cual indica que todo el proceso se renueva por cada cambio en un objeto, y generalmente tienen que ver con variaciones en un Input desde el interfaz del usuario y que será enviado al servidor para sufrir una transformación mediante llamadas a los distintos módulos, y de nuevo volverán al cliente mediante distintos objetos Outputs.
La visualización de los distintos objetos Outputs que se van a generar se manejan mediante las opciones disponibles para un objeto de salida: si la variable suspendWhenHidden toma el valor TRUE (valor predeterminado), el objeto de salida se suspenderá (no se ejecutará) cuando esté oculto en la página web. Cuando es FALSO, el objeto de salida no se suspenderá cuando esté oculto, y si ya estaba oculto y suspendido, se reanudará inmediatamente.
Otro aspecto importante es que se trabaja con módulos por lo que es necesario inicialmente usar la función NS que permite crear un espacio de nombres únicos donde se almacenarán cada identificador de entradas y salidas (ID) y evitarán, de este modo, la colisión. Cada módulo consiste en una función para crear la interfaz de usuario y una

función para llamar dentro de la función del servidor usando callModule. El Cliente debe incluir todos los ID de entrada y salida usando NS (). La función callModule maneja el prefijo para el componente del servidor.

A continuación explicaremos las características técnicas de cada tipo de información que se muestra en la página central:

Tabla Evidencia científica por genes

El módulo *tables_evidencia_cientifica* es llamado desde el objeto *gen_search_evidencia_cientifica* que es reactivo cuando se selecciona un nuevo gen y es utilizado como argumento mediante la función *get("input")[["gen_id"]]*. La información es organizada de modo tabular en cuatro columnas que incluyen *"Gen"*, *"Artículo","PMID" y "Resumen"*, y se utiliza el paquete *kable y kableExtra* que permite construir tablas complejas comunes y manipular estilos de tabla. Este paquete importa el símbolo de %>% de *magrittr* y verbaliza todas las funciones, por lo que básicamente se puede agregar "layers" a una salida *kable.*
Una de las peculiaridades de esta tabla es que el resumen es generado de manera automática desde el Abstract del artículo, buscando como expresión regular la palabra "*conclusion*", de este modo, uno de los inconvenientes es que cuando no aparece en el abstract esta palabra, el resultado será un vector de caracteres vacio.

Tabla Relación Genes Farmacos

Para construir esta tabla se hace una llamada al módulo *tables_generation* que contiene el script para hacer la busqueda en PubMed, considerando como argumento el nombre de un gen que se obtiene mediante la función *get("input")[["gen_id"]] o input$gen_id*.
El interés técnico este módulo es cómo está organizada la información procedente de la función *queryDGIdb(data)* del paquete *pubmed.mineR*. De este modo, se han considerado cuatro variables *'Gen','Droga','Tipo de interacción', 'PMID'* y ordenando la tabla por la cuarto columna según el número de identificadores *PMID* que tiene cada fármaco, lo cual nos da una idea de la evidencia científica disponible. De este modo, el fármaco que se sitúa en la primera fila es el que más citas bibliográficas tiene en PubMed.
Se utiliza el paquete *formattable* para formatear la tabla en HTML y de este modo mejorar la lectura de los datos presentados en forma tabular.

Tabla Ensayos clínicos por genes

El objeto reactivo *gen_search_clinical_trial* permite la llamada al *módulo tables_clinical_trial* cuando es seleccionado un nuevo gen, y es utilizado nuevamente como argumento mediante la función *get("input")[["gen_id"]]*. La información se obtiene gracias al paquete *rclinicaltrials*, y mediante la función *clinicaltrials_download* y es organizada en una tabla que incluye los siguientes campos: *"titulo", "indicación", "estado","fecha de inicio","tipo de estudio"*. La complejidad de esta tabla es debida a tener que filtrar la información para seleccionar la que consideramos necesaria, y mostrarla en una tabla con formato HTML para lo que se ha utilizado el paquete k*able y kableExtra.*

Tabla Indicación FDA por fármacos

De nuevo es llamado el módulo *tables_generation* cuando hay algún cambio en el campo *gene_id* y es usado como argumento mediante la función *get("input") [["gen_id"]]*. Lo complejo de esta tabla es extraer la información y organizarla en una tabla de cuatro columnas que incluye los siguientes campos: *"Nombre comercial","Fármaco","Indicación por la FDA"*. El objetivo es extraer toda la información disponible de la base de datos FDA para el gen seleccionado. Con el fin de agilizar la busqueda lo más recomendable es descargar en local las bases de datos de la FDA y consultarlas posteriormente. Otro de los problemas que se soslayó fue que cada fármaco puede tener varias denominaciones, por lo que se utiliza una función iterativa para mostrar la información de cada uno de las denominaciones del fármaco y que inicialmente son almacenadas en una lista. Por otra parte, se utiliza una expresión regular para encontrar en el texto la indicación autorizada para cada fármaco. El resultado es una tabla de tres columnas y formateadas mediante funciones del paquete k*able y kableExtra.*

Red de genes por enfermedad

El módulo *Search_diseases* se inicia cuando es completado el campo *disease_id* y es llamado al cliclear el botón *Buscar*, usando como argumento *get("input") [["disease_id"]]*. Se utiliza la función *observeEvent* que reacciona cuando detecta un cambio en el objeto *input$Seeking_Button*.
Con el fin de homogeneizar los términos utilizados, se hace primeramente una busqueda del término más proximo en el tesauro *UMLS*. *UMLS* es un repositorio de vocabularios biomédicos desarrollado por la Biblioteca Nacional de Medicina de EE. UU. El *UMLS* integra más de 2 millones de nombres para unos 900,000 conceptos de más de 60 familias de vocabularios biomédicos, así como 12 millones de relaciones entre estos conceptos. Los vocabularios integrados en el metatesauro UMLS incluyen la taxonomía *NCBI, Gene Ontology, Medical Subject Headings (MeSH), OMIM* y la base de conocimiento simbólico digital anatomista.
Se usa un algoritmo que permite obtener el término UMLS más próximo a los vocablos introducidos, de tal modo que cuando no encuentra ninguno, el sistema lanza el aviso de que el “término no ha sido encontrado y que introduzcas otro término”.
Se utiliza la función *updateTextInput* para actualizar el campo *disease_id*, una vez se ha introducido un término y es mostrado el resultado.
Se emplea la función *EutilsSummary* del paquete *pubmed.mineR* con el fin de realizar una busqueda en *PubMed*, incluyendo los *Abstract* que contengan los términos buscados y el resultado es almacenado en la clase *Abstracts_pubmed*, utilizando los *slots* especificados en la representación. A continuación se utilizan dos funciones de atomización, *word_atom* y *gene_atom*. La primera permite separar en un texto cada una de sus palabras y elimina aquellos vocablos considerados comunes por encontrarlos en un diccionario de términos de uso muy habitual, el cual contiene 106 palabras. El resultado es una tabla de frecuencias que muestra cada palabra y el número de veces que ha sido observada en el objeto *volume@Abstract.*
La función *gene_atom* permite buscar dentro de las palabras que componen el *corpus* de la busqueda, aquellas que corresponden a genes al comparar cada término con los nombres de los genes presentes en el objeto *HGNCdata$Approved.Symbol*, el cual contiene 19064 términos. El resultado es una tabla que contiene cada gen y su frecuencia de aparición en el *corpus* considerado.

Uno de los puntos claves de esta tabla es la utilización de la función *lsa_propia* que permite factorizar la matriz *tdm*, que contiene como filas, el nombre de los genes y como columnas, el nombre del objeto *volume@Abstract* (de tal modo, que podemos ver en que documento es encontrado cada gen), en dos matrices ortonormales cuyos vectores columnas son linealmente independientes. De este modo, la descomposición de valores singulares (DVS) permite a partir de una matriz A de $m \times n$ ($m \geq n$) factorizarla como $A = U\Sigma V^T$,donde U es una matriz con columnas ortogonales de $m \times n$, V es una matriz ortogonal de $n \times n$, y Σ una matriz "diagonal" de $n \times n$. Los vectores de las columnas de U se denominan vectores singulares por la izquierda de A, mientras los vectores de las columnas de V se llaman vectores singulares por la derecha de A. Las matrices U y V no están determinadas en forma única por A, en cambio Σ si, porque contiene los valores singulares de A. La DVS de una matriz A da mucha información acerca de A, como que:

(i) el rango de A es r.
(ii)$\{u_1, u_2, \ldots, u_r\}$ es una base ortonormal de R(A)
(iii)$\{u_{r+1}, u_{r+2}, \ldots, u_n, \ldots u_m\}$ es una base ortonormal de $N(A^T)$
(iv)$\{v_1, v_2, \ldots, v_r\}$ es una base ortonormal de $R(A^T)$
(v) Si $r < n$, $\{v_{r+1}, v_{r+2}, \ldots, v_n\}$ es una base ortonormal de N(A)

Posteriormente se obtiene la matrix A que corresponde al producto interno de las matrices $U\Sigma V^T$ y se calcula el coseno entre dos vectores o entre todos los vectores de columna de una matriz. Esta medida representa la similitud de dos vectores, lo que en términos estadísticos es una medida de correlación entre dos valores. Los valores cercanos a 1 representan una alta similitud o correlación, mientras que los valores cercanos a 0 indican que son vectores perpendiculares y no correlacionados, o lo que es lo mismo, son dos vectores linealmente independientes. Se considera, de modo arbitrario, que dos vectores están asociados si presentan una correlación de al menos 0.7. De este modo, la descomposición del valor singular (SVD) permite reducir el número de filas y preservar la estructura de similitud entre las columnas (figura 1 y 2).
El siguiente paso es representar esta matriz en un gráfico de nodos y enlaces. Para ello se utiliza el paquete *igraph*, de tal modo que, los nodos están representados por los nombres de los genes que corresponden a los nombres de las filas de la matriz de similitud (A) y los enlaces corresponden a los genes que presentan una correlación con cada gen de al menos 0.7. El gráfico es construido a partir de una data.frame que tiene en las dos primeras columnas el nombre de los genes, la primera representa el nodo de partida, y la segunda columna, el nodo de enlace. La tercera columna tiene el valor de correlación entre los dos vectores, y es utilizada para situar los distintos nodos en el gráfico de acuerdo a su similitud. El tamaño de cada nodo vendrá determinado por el número de enlaces que presente.
En conclusión, mediante esta representación nodo-enlace se puede conocer para un término médico el conjunto de genes que están descritos en la literatura y las relaciones entre los mismos de acuerdo a la ocurrencia de estos genes entre todos los documentos analizados. La interpretación técnica es que genes que están descritos conjuntamente en un documento es porque probablemente están relacionados por intervenir en un mismo proceso biológico, una misma localización celular o proceso bioquímico.

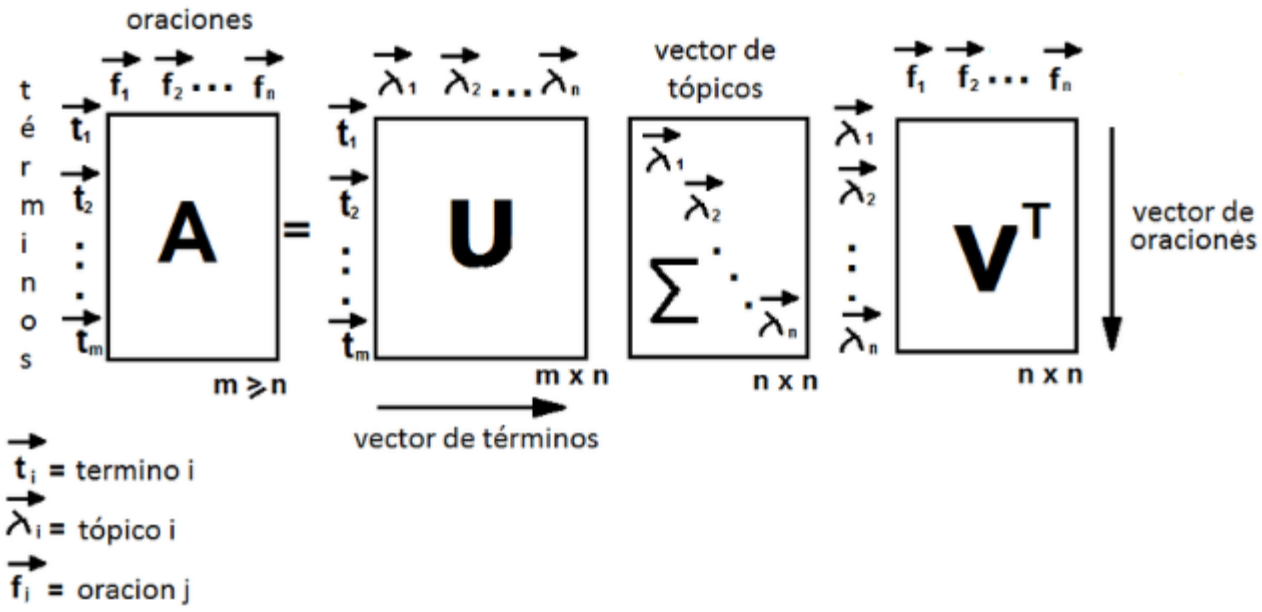


Figura 1. Representación geométrica de las matrices obtenidas de la factorización de la matriz A

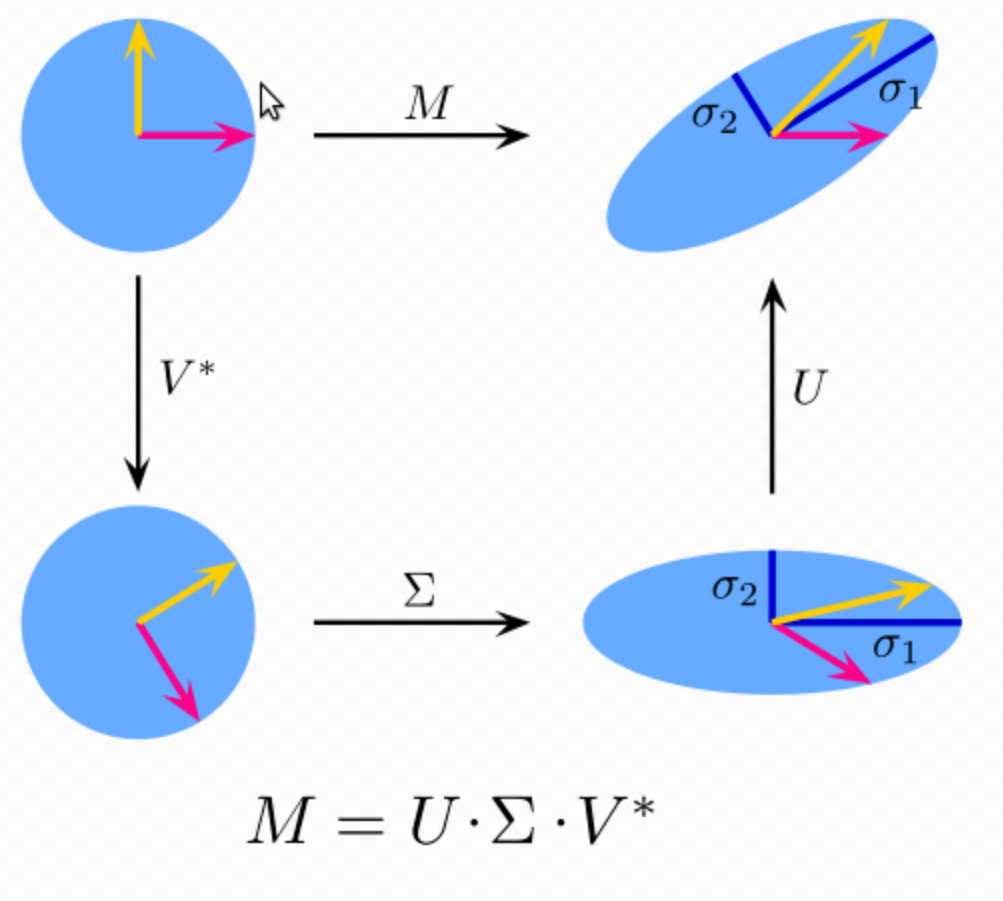


Figura 2. Interpretación geométrica de la descomposición del valor singular

Similitud de genes

El módulo *datafile_Similarity_genes_GO* se inicia cuando se llama a la función *tables_Similarity_genes_GO*. En este caso los genes seleccionados en la tabla *genes_table* son almacenados como una variable de entorno llamada *'gen_tb_choice'*, y de este modo, será utilizada por la función *tables_Similarity_genes_GO* como argumento. Se emplea el paquete *org.Hs.egALIAS2EG* para obtener el número de identificación *GenBank* a partir del nombre del gen.
Los términos GO asociados a cada gen se obtienen a partir de la función *getGOInfo* del paquete *GOSim*. El objeto que se obtiene y sobre el que se trabaja es una *data.frame* que tiene 4 filas etiquetadas como *go_id, Term, Definition, IC* ,y un número de columnas igual al número de genes considerados. Los elementos de esta data.frame son listas.
La similitud entre genes basados en términos GO, o bien entre los términos GO, se realiza mediante el cálculo de una matrix de correlación a partir de la matrix que tiene por valores, la probabilidad de similitud de ambos términos basada en la identificación del ancestro común más proximo (figura 3). Este valor puede variar según el modelo de similitud considerado, por ejemplo, en el modelo propuesto por *Resnik*, este valor es el logaritmo negativo de la probabilidad de ocurrencia de dos términos. La probabilidad de que dos términos estén asociados se calcula mediante una distribución de probabilidad hipergeométrica, puesto que la probabilidad de que un término este relacionado entre ambos genes depende del resultado de la relación de los términos anteriores, y esta probabilidad se va incrementando para cada término (figura 4 y 5).
A continuación se simplifica la matrix de correlación mediante el análisis de los componentes principales que permite transformar el conjunto original de variables, en otro conjunto de nuevas variables incorreladas entre sí. El primer componente se calcula eligiendo a_1 de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $aa_1 = 1$.
El gráfico de dos dimensiones, que muestra las coordenadas de cada gen o cada término GO, corresponden al valor que toma el primer componente principal y el segundo componente. De este modo, se puede visualizar la proximidad de cada gen o término GO.
Una peculiaridad del script es que las funciones renderPlot sólo pueden enviar un gráfico desde el servidor al interfaz del usuario. Para poder enviar tantos gráficos como número de genes seleccionados, el servidor devuelve una lista de objetos gráficos y son colocados en una división en linea que es devuelta al cliente como un objeto reactivo HTML.

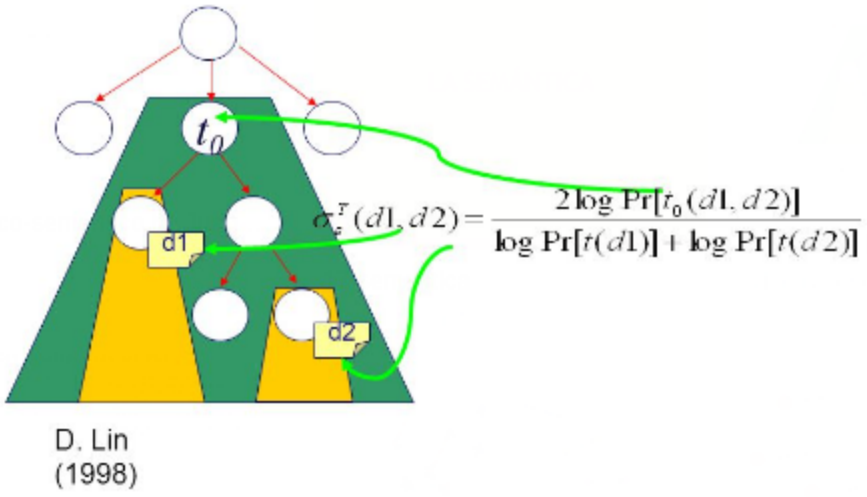


Figura 3. Representación de una ontología y busqueda del ancestro común más próximo

Definición: Distribución Hipergeométrica

Sean **N:** Cantidad de elementos del conjunto del que se toma la muestra
K: Cantidad de elementos existentes que se consideran "éxitos"
n: Tamaño de la muestra
X: Variable aleatoria discreta (es la cantidad de resultados considerados "éxitos" que se obtienen en la muestra)
x = 0, 1, 2, ..., n (son los valores que puede tomar **X**)

Entonces, la distribución de probabilidad de **X** es

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad x = 0,1,2,\ldots,n$$

Demostración

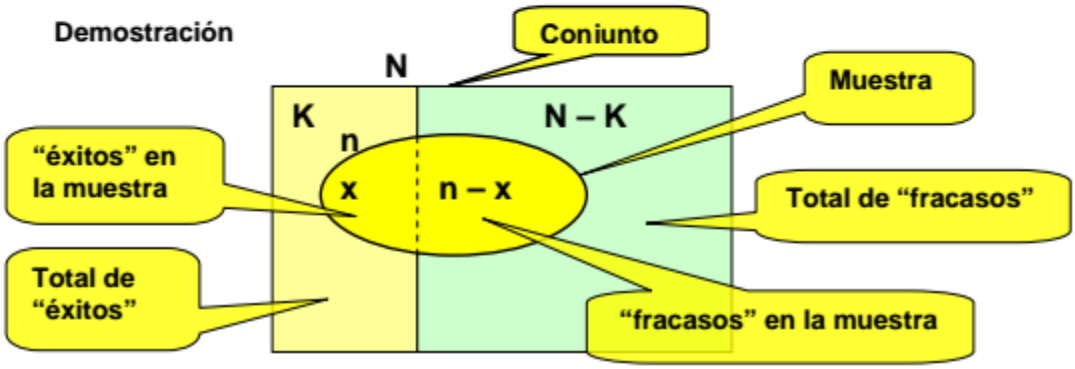


Figura 4. Cálculo de la probabilidad de ocurrencia de dos términos basado en una distribución hipergeométrica

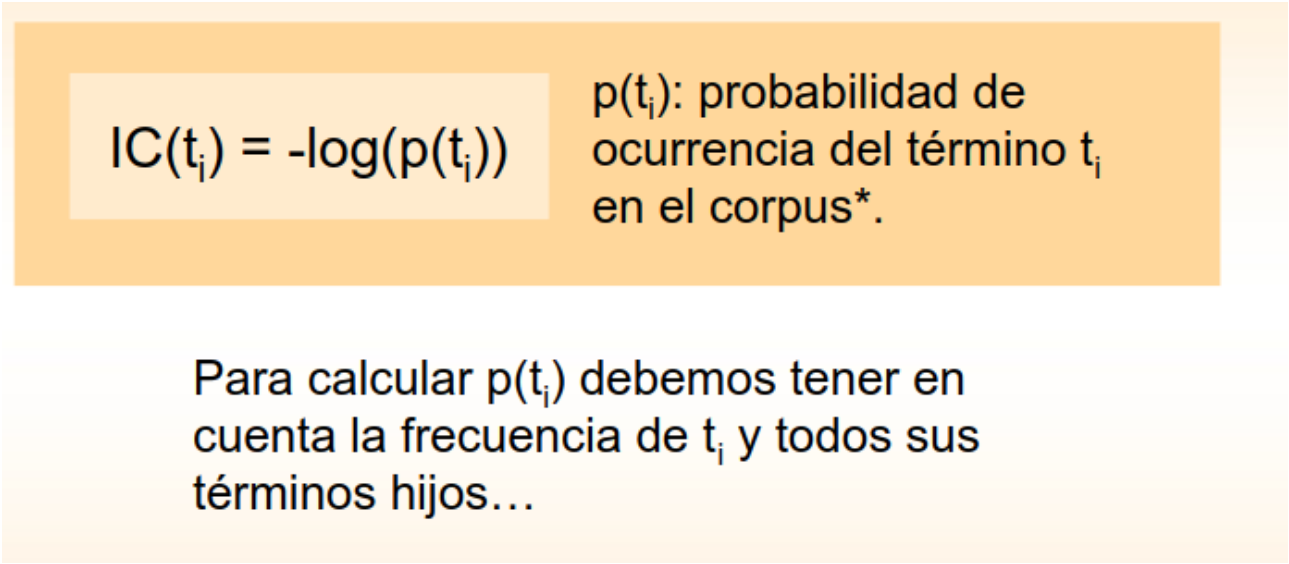


Figura 5. Cálculo de la similitud mediante el coseno de dos vectores que están representados por el logaritmo negativo de la probabilidad de ocurrencia de cada término basado en una distribución hipergeométrica (según el modelo Resnik)

Filtrado de genes desde panel

Este módulo muestra la misma información ordenada, según el número de identificadores de *PubMed* (*PMID*), que la tabla "*Relacion Genes Farmacos*" pero se diferencia en que el módulo es iniciado cuando se carga un archivo de texto que contiene el nombre de los genes separados por comas.
La utilidad de este módulo es permitir la priorización de genes dentro de un panel basado en la evidencia científica disponible. Se debe tener en cuenta que el tiempo de ejecución aumenta según el número de genes consultados por lo que puede ser un problema que limite la efectividad de la plataforma.

2.3 Instalación de las librerías utilizadas

2.3.1 OpenFDA

Este paquete proporciona funciones para acceder a la API de OpenFDA desde R. Utiliza los paquetes *jsonlite* y *magrittr* para convertir las consultas de OpenFDA a una data.frame que será utilizada para el análisis.
OpenFDA está diseñado principalmente para consultas en tiempo real. Sin embargo, como nuestra aplicación excedía los límites de consulta o de resultados (25000 registros por consulta) he creado una instancia propia de openFDA y lo ejecuto desde el propio servidor u ordenador local. En este sentido he descargado los archivos de openFDA, en formato JSON.
Una desventaja de esto es que las descargas están divididas en partes. Algunos puntos finales tienen millones de registros. Para esos puntos finales, los datos se dividen en muchas partes pequeñas. Entonces, aunque algunos puntos finales tienen todos sus datos disponibles en un solo archivo JSON, otros tienen docenas de archivos.
Por otra parte, para mantener los datos descargados actualizados, se necesita volver a descargar los datos cada vez que se actualicen (lo cual sucede de manera regular), y es posible que cada registro cambie debido a correcciones o mejoras.

2.3.2 pubmed.mineR

Pubmed es una base de datos de consulta pública de artículos de revistas publicadas puesta a disposición por National Institutes of Health.
El paquete R, pubmed.mineR, permite extraer datos de resúmenes de PubMed utilizando algoritmos de minería de texto, y varias funciones existentes de otros paquetes R como RISmed, lo que permite extraer el contenido de la interfaz Entrez Programming Utilities (E-Utilities) a la consulta de PubMed y la base de datos de NCBI.
Para iniciar pubmed.mineR debemos guardar localmente los resúmenes de los artículos previamente descargados y utilizando el motor de búsqueda PubMed, en formato texto o XML.
Las dos funciones clave son - readabs () (para texto formato) y xmlreadabs () (para formato XML) - que permiten importar ya sea el archivo de texto o el archivo XML en el objeto S4 de Clase 'Abstracts'. Este objeto de datos S4 es el punto de partida para todo el procesamiento posterior.
En nuestra caso como queremos hacer una aplicación web, hemos construido un script que permite llamar a la función EUtilsSummary () del paquete RISmed para producir un objeto que almacenaremos en un objeto de clase mediante el comando setClass. El comando toma varios argumentos como “Title, Abstract y PMDI”.
Se crearon dos funciones “word_atom” y “gene_atom” que siguen el mismo algoritmo que las funciones “word_atomization” y “gene_atomization” del paquete pubmed.mineR pero adaptada a la estructura de nuestro objeto de clase.
La función “gene_atom” permite la búsqueda automatizada de genes en formato HGNC desde el objeto de clase y fragmentado en cadenas de caracteres. El resultado final es una data.frame con tres columnas: la primera con el nombre del gen en formato HGNC, la segunda el nombre completo del gen, y la tercera el número de veces que se ha encontrado este gen en el objeto de clase analizado.
Otra función que hemos creado es “conclusión” que permite extraer un resumen del abstract y que se mostrará en la tabla web de visualización de los resultados.

2.3.3 Rclinicaltrials

Clinicaltrials.gov es una base de datos de registro y resultados de estudios clínicos públicos y privados realizados en pacientes de todo el mundo. Los usuarios pueden buscar información y resultados de esos ensayos.
Este paquete *Rclinicaltrials* proporciona un conjunto de funciones para interactuar con las funciones de búsqueda y descarga. Los resultados se descargan en directorios temporales y se devuelven como objetos R.
La función principal es clinicaltrials_search (). Esto devuelve una información básica sobre los ensayos. La consulta puede ser de una sola cadena que se pasará al campo "términos de búsqueda" en clinicaltrials.gov, o se pueden combinar los términos utilizando los operadores lógicos AND, OR y NOT.
Los datos provienen de una base de datos relacional con muchos campos de texto, por lo que puede tardar un largo tiempo en devolver los resultados de la búsqueda, que se estructurarán como una lista de data.frames. La información requerida debe ser extraída a continuación según el criterio del usuario.

2.3.4 DGIdb

La base de datos Drug Gene Interaction (*DGIdb*) integra datos de 15 fuentes primarias que cubren genes humanos, medicamentos, interacciones fármaco-gen y fármacos potencialmente relevantes para la enfermedad. Actualmente, DGIdb contiene 2.611 genes y 6.307 fármacos implicados en más de 14.144 interacciones fármaco-gen y 6.761 genes que pertenecen a una o más de las 39 categorías de genes potencialmente farmacorresistentes. Cada asociación de gen de fármaco o categoría de gen está vinculada a su base de datos primaria o fuente de literatura. DGIdb organiza los genes del genoma farmacorresistente en dos clases principales. La primera clase incluye genes con interacciones medicamentosas conocidas obtenidas mediante la extracción de la literatura o mediante el análisis de revisiones y bases de datos disponibles públicamente. La segunda clase incluye genes que pueden no estar actualmente dirigidos terapéuticamente pero que son "potencialmente" dirigibles de acuerdo con su pertenencia a una determinada categoría de genes asociados con la farmacología (por ejemplo, quinasas).
En la plataforma creada el usuario puede introducir un solo gen y explorar el estado actual de conocimiento con respecto a la farmacorresistencia de ese gen. Alternativamente, puede incorporar una gran lista de genes (panel de genes) para identificar el subconjunto con potencial terapéutico.
La función *queryDGIdb()* proporciona una interfaz para consultar DGIdb desde R utilizando la API DGIdb.

2.3.5 Shiny

Shiny es una herramienta para crear fácilmente aplicaciones web interactivas que permiten a los usuarios interactuar con sus datos sin tener que manipular el código.
Shiny se basa la programación Reactiva que vincula los valores de entrada con los de salida.
Cuando una entrada (input) cambia, el servidor reconstruye cada salida (output) que depende de ella (también si la dependencia es indirecta). Se puede controlar este comportamiento a través de la cadena de dependencias.
La programación Reactiva enfatiza el uso de:

- Valores que cambian en el tiempo
- Expresiones que registran esos cambios

Cada aplicación es una carpeta que contiene los siguientes 2 archivos:

server.R: Instrucciones que constituyen los componentes de R de la aplicación.

ui.R: Una descripción de la interfaz (UI) de la aplicación y la secuencia de comandos de la interfaz de usuario.

2.3.6 KableExtra

El objetivo de *kableExtra* es ayudar a crear tablas complejas y manipular estilos de tabla. Importa el símbolo de %>% de *magrittr* y verbaliza todas las funciones, por lo

que básicamente puedes agregar "layers" a una salida kable de una manera similar a ggplot2 y plotly.

2.3.7 Igraph

Los dos aspectos principales de las redes son una multitud de entidades separadas y las conexiones entre ellas. Las entidades se conocen como nodos o vértices de un gráfico, mientras que las conexiones son bordes o enlaces.
Los paquetes de análisis de red necesitan que los datos estén en una forma particular para crear el tipo especial de objeto utilizado por cada paquete. Las clases de objetos para la red en igraph están todas basadas en matrices de adyacencia, también conocidas como sociomatrices. Una matriz de adyacencia es una matriz cuadrada en la que los nombres de columna y fila son los nodos de la red. Dentro de la matriz, un 1 indica que hay una conexión entre los nodos, y un 0 indica que no hay conexión. Las matrices de adyacencia implementan una estructura de datos muy diferente a una data.frame.
Una lista de bordes es un marco de datos que contiene un mínimo de dos columnas, una columna de nodos que son la fuente de una conexión y otra columna de nodos que son el objetivo de la conexión. Los nodos en los datos se identifican mediante ID únicos. Si la distinción entre fuente y objetivo es significativa, la red se dirige. Si la distinción no es significativa, la red no está dirigida.
Para generar las relaciones entre los genes cree un script R que a partir de la matriz de palabras asociadas generadas en el análisis semántico latente, se distribuían en una data.frame con tres columnas: la primera con los nodos origen, la segunda con los nodos que reciben y la tercera con el coseno de los vectores de genes y que establece la correlación entre ambos genes.

2.3.8 org.Hs.eg.db

El paquete org.Hs.eg.db es un objeto R que contiene las asignaciones entre los identificadores de Entrez Gene y los números de acceso de GenBank. Utilizamos determinadas funciones de org.Hs.eg.db para convertir fácilmente los identificadores Entrez Gene.
Es necesario realizar este proceso de conversión dado que por razones históricas, el mismo gen puede tener más de un símbolo, y esto generalmente complica mucho cualquier análisis. El bimap ALIAS2EG permite este tipo de conversión. Por otra parte, también es importante recordar que no solo el mismo gen puede tener más de un símbolo, sino que también el mismo símbolo puede coincidir con múltiples identidades entre. Por ejemplo, con el código siguiente podemos obtener qué símbolos coinciden con más de una identificación en la especie humana:

```
as.data.frame(org.Hs.egALIAS2EG) %>%
count(alias_symbol) %>%
arrange(-n)
```

```
alias_symbol     n
  <chr>       <int>
 1 VH            36
 2 TARNN-GUU     20
 3 MT1           12
 4 GPCR          11
```

```
 5 PR           11
 6 HOX1         10
 7 HOX2          9
 8 NAP1          9
 9 PPIase        9
10 Protease      9
```

Lo único seguro que se puede hacer en estos casos es identificar los símbolos duplicados y descartarlos o curarlos manualmente.

```
mygenes <- c("ALK","ROS1","EGFR","MGAT3")
as.data.frame(org.Hs.egALIAS2EG) %>%
count(alias_symbol) %>%
filter(alias_symbol %in% mygenes)
```

```
 alias_symbol     n
  <chr>       <int>
1 ALK             1
2 EGFR            1
3 MGAT3           2
4 ROS1            1
```

```
AnnotationDbi::select( org.Hs.eg.db, keys=mygenes, keytype='ALIAS',
columns=c('ENTREZID', 'ENSEMBL')) 'select()' returned 1:many mapping between
keys and columns
```

```
 ALIAS ENTREZID        ENSEMBL
1   ALK      238 ENSG00000171094
2  ROS1     6098 ENSG00000047936
3  EGFR     1956 ENSG00000146648
4 MGAT3     4248 ENSG00000128268
5 MGAT3   346606 ENSG00000106384
```

Este paquete se actualiza semestralmente.

2.3.9 GOSim

El paquete GOSim implementa diferentes métodos para calcular las similitudes funcionales entre los productos de los genes en función de las similitudes entre los términos GO asociados. Esto puede utilizarse, por ejemplo, para agrupar genes de acuerdo con su función biológica y, por lo tanto, puede ayudar a comprender mejor los aspectos biológicos cubiertos por un conjunto de genes.
GOSim se concentra en los conceptos de similitud para los términos GO derivados de la teoría de la información. Una de las medidas de similitud de la teoría de la información más conocida fue la introducida por Resnik. Se basa en la noción del llamado mínimo subsumer de dos términos GO t y t ', que es el ancestro común más bajo en la jerarquía GO. Su contenido de información IC_{ms}, que se puede entender como una medida de similitud entre t y t ', viene dado por:

$sim(t,t') = ICms(t,t') := \max t\hat{} \in Pa(t,t') IC(t\hat{})$

Aquí Pa (t, t ') denota el conjunto de todos los ancestros comunes (también indirectos) de los términos GO t y t', mientras que IC (t) denota el contenido de información del término t. Se define como:

IC(t) = -log P(t)

es decir, como el logaritmo negativo de la probabilidad de observar t y cuya función sigue una distribución hipergeométrica. El contenido de información de cada término GO se puede calcular con GOSim para cada una de las taxonomías función molecular, proceso biológico y componente celular. El cálculo se basa en el recuento en la que un término GO específico o cualquiera de sus descendientes directos o indirectos aparecen en un producto genético anotado. La asociación entre los productos genéticos y los identificadores GO se informa regularmente por el Consorcio GO. El Consorcio GO además proporciona códigos de evidencia en las anotaciones, que se pueden usar para calcular los contenidos de información de todos los términos de GO sobre una base diferente. GOSim almacena los contenidos de información de todos los términos GO en los archivos de datos para acelerar todos los cálculos siguientes. Por defecto, para algunas combinaciones de códigos de evidencia, los contenidos de la información ya están precalculados.
De este modo, dado dos genes g y g' anotados con los términos GO t_1, ..., t_n y t'_1, ..., t'_m definimos la similitud funcional entre g y g' como:

$simgene(g,g') = \max i=1,,...,nj=1,...,msim(ti,t'j)$

donde *sim* es una medida de similitud para comparar términos GO t i y t'j. En GOSim, el valor resultante se puede normalizar aún más para tener en cuenta una cantidad desigual de términos GO para ambos genes.

Debido a que los vectores de características son muy dimensionales, generalmente realizamos un análisis de componentes principales (PCA) para proyectar los datos en un subespacio dimensional inferior. El número de componentes principales se elige por defecto de manera que se pueda explicar al menos el 95% de la varianza total en el espacio de características, y los vectores de características se normalizan a la norma 1.
GOSim tiene la posibilidad de evaluar una agrupación dada de genes o términos mediante sus similitudes GO. GOSim usa la similitud funcional entre los genes para calcular para cada grupo la mediana dentro de la similitud del clúster y la desviación absoluta media. Además, GOSim también proporciona una visualización a través de siluetas de clúster.
Además de la medida de similitud del término GO de Resnik, existen extensiones de Lin, y Jiang y Conrath, que se incluyen también en GOSim.

2.3.10 DOSE

DOSE proporciona la función doSim para calcular la similitud semántica entre los términos DO. DOSE implementa cuatro algoritmos basado en el contenido de información propuestos por Resnik (Resnik, 1999), Lin (Lin, 1998), Jiang y Conrath

(Jiang y Conrath, 1997) y Schlicker (Schlicker et al., 2006), respectivamente, y uno basado en algoritmos gráficos propuesto por Wang (Wang et al., 2007) para medir la similitud semántica entre los términos DO.

La función geneSim mide las similitudes semánticas entre los genes en función de sus términos anotados DO. Se implementan cuatro estrategias combinadas para agregar puntuaciones semánticas de similitud de múltiples DO relacionados con genes, incluido max que calcula el puntaje máximo de similitud sobre todos los pares de términos DO, promedio que usa el promedio de puntuaciones de similitud sobre todos los pares de términos DO. rcmax que mide el máximo de RowScore y ColumnScore, donde RowScore (ColumnScore) es el promedio de similitud máxima en cada fila (columna) y el mejor promedio que mide el promedio de puntuaciones de similitud máxima en cada fila y columna. Los resultados de similitud semántica obtenidos de doSim y geneSim se pueden visualizar mediante la función simplot.

DOSE proporciona un modelo hipergeométrico para evaluar asociaciones de enfermedades de genes expresados diferenciales. La función enrichDO permite seleccionar un fondo apropiado de genes como referencia. La función gseAnalyzer es compatible con GSEA para evaluar la relevancia de la enfermedad de los datos de alto rendimiento. Estos enfoques pueden usarse para verificar si los genes implicados en el experimento biológico están asociados a la enfermedad y para identificar asociaciones de enfermedad inesperadas. También se incorporan correcciones de comparación múltiple que incluyen Bonferroni, Benjamini, False Discovery Rate y q-values. Las asociaciones de enfermedades entre diferentes grupos de genes o listas de genes de diferentes condiciones se pueden comparar utilizando el paquete R clusterProfiler. Se implementan varias funciones de visualización que incluyen barplot y cnetplot para visualizar asociaciones de enfermedades significativas y una red de asociación de genes y enfermedades, respectivamente. La suma de las puntuaciones de enriquecimiento y su asociación con el fenotipo se puede visualizar utilizando la función gseaplot.

2.3.11 disgenet2r

Este paquete es utilizado para conocer el término asociado a una determinada enfermedad en el Sistema de lenguaje médico unificado (UMLS) mediante la función *getUMLs*. Esto permite emplear identificadores únicos para anotar homogéneamente enfermedades obtenidas de diferentes fuentes.
El Sistema de lenguaje médico unificado (UMLS) es un compendio de muchos vocabularios controlados en las ciencias biomédicas creado en 1986. Proporciona una estructura de mapeo entre estos vocabularios y, por lo tanto, permite traducir entre los diversos sistemas de terminología; también se puede ver como un tesauro completo y una ontología de conceptos biomédicos. UMLS además proporciona instalaciones para el procesamiento del lenguaje natural.
El UMLS fue diseñado y mantenido por la Biblioteca Nacional de Medicina de EE. UU. Se actualiza trimestralmente y se puede usar de forma gratuita.

2.4 Integración de bases de datos

Las bases de datos que se van a utizar son *PubMed, Drug Gene Interaction (DGIdb), Clinical Trials y FDA.* La integración de las mismas mejora la fiabilidad y la interpretación de los datos, por medio de los metadatos, dado que proveen un gran

potencial al permitir cerciorar si un dato o conjunto de datos son apropiados para una necesidad.

La información de estas bases de datos cambia con frecuencia, por lo tanto, un enfoque útil del que se parte es diseñar un sistema de extracción de los datos directamente de las fuentes de datos individuales. El siguiente paso será combinar datos de diferentes fuentes, posiblemente heterogéneas entre sí, autónomas y distribuidas, con el objetivo de proporcionar a los usuarios una visión unificada de esos datos. Este proceso de transformación se logra mediante:

- Aplicando filtros, con el fin de recoger únicamente los registros con determinados datos.
- Cambiando el formato y adaptarlo a la nueva estructura de datos.
- Llevando a cabo agregaciones que permitan adaptar los datos según la estructura que interese.
- Estableciendo uniones de los datos.

Posteriormente, se vuelcan los datos procesados a las estructuras finales de las tablas que presentarán la información.

2.5 Minería de datos mediante análisis de semántica latente

Para poder construir el módulo que hemos llamado "*Red de genes*" se ha utilizado el análisis de semántica latente (LSA) que emplea una técnica para disminuir las dimensiones de sus matrices llamada descomposición de valores singulares (Singular Value Descomposition; SVD), la cual pretende encontrar un espacio semántico latente mediante la factorización de matrices. Esta técnica descompone una matriz de término-documento en un conjunto de factores ortogonales, desde los cuales la matriz original puede aproximarse por una combinación lineal.

El LSA comienza procesando un texto de grandes dimensiones que llamaremos corpus lingüístico. Dicho corpus contiene miles de palabras, párrafos y frases. Además, se representa como una matriz de frecuencias cuyas filas son las distintas palabras del corpus y cuyas columnas aparecen los distintos párrafos o frases. De esta forma la matriz contiene el número de veces que cada palabra aparece en el texto. A continuación, se realiza una ponderación con el fin de restar importancia a los términos muy frecuentes ya que en cualquier texto aparecen reiteradas veces artículos y determinantes que no aportan información relevante; y aumentarla a los menos frecuentes debido a que las palabras excesivamente frecuentes no nos sirven para seleccionar bien la información relevante del párrafo y las que aparecen de forma moderada sí.
El siguiente paso es aplicar un algoritmo denominado Descomposición en Valores Singulares (Singular Value Descomposition; SVD) con el fin de disminuir la dimensión de la matriz a una cifra más accesible sin perder información importante de la original.

Otro propósito interesante de este algoritmo es el de obtener una matriz que contenga únicamente los vectores con información relevante.
La ventaja de representar el lenguaje vectorialmente es que los vectores son aptos a comparaciones por medio de distancias euclídeas, cosenos y otras medidas.

Además, a partir de las coordenadas de la matriz que tenemos se puede introducir en el espacio nuevos vectores que representen textos introducidos a posteriori llamados pseudodocumentos. Los pseudodocumentos son textos que añadimos al espacio semántico reducido definido y que no forman parte del corpus inicial.
El LSA permite realizar el proceso a este último y añadirlo a nuestra matriz reducida sin necesidad de realizar el proceso de nuevo con todos los documentos.

Pasos seguidos

1. En primer lugar, se parte de la matriz de términos-documentos de dimensión m x n, *A*, donde cada columna corresponde a un documento. Si el término *i* aparece *a* veces en el documento *j*, entonces *A[i,j] = a*.
2. A través de esta matriz A se obtiene la matriz términos – términos $B = A \cdot t(A)$ de dimensión m x m y la matriz documentos – documentos $C = t(A) \cdot A$ de dimensión n x n . Si los términos i y j aparecen juntos en el documento b, entonces B[i,j] = b. Por otro lado, si el documento i y j tienen c palabras en común, entonces C[i,j] = c.
3. Se define ahora las matrices:
    1. S : matriz de autovalores de B
    2. U : matriz de autovalores de C
    3. P : matriz diagonal cuyos elementos son las raices cuadradas de los autovalores de la matriz B

Donde: $A = S \cdot P \cdot t(U)$.

4. Por medio de la SVD se reducen las dimensiones de la matriz inicial eliminando los valores mas pequeños de la matriz P, que como están ordenados por relevancia de mayor a menor resulta muy fácil. Quedarán los k valores más grandes creando la matriz Pk. A consecuencia, se reduce también las matrices S y t(U). La matriz A se aproxima, de este modo, por: $Ak = Sk \cdot Pk \cdot t(U)$. La dimensión de Sk será m x k, de Pk será k x k y de t(U) será k x n. Luego, la matriz Ak tiene de nuevo dimensión m x n. Las palabras vendrán representadas por las filas de la matriz m x k $Sk \cdot Pk$, mientras que los documentos lo estarán por las columnas de la matriz k x n $Pk \cdot t(Uk)$.
5. La similitud de los términos vendrá dada por el centroide de los vectores.

2.6 Similitud de genes basado en su ontología

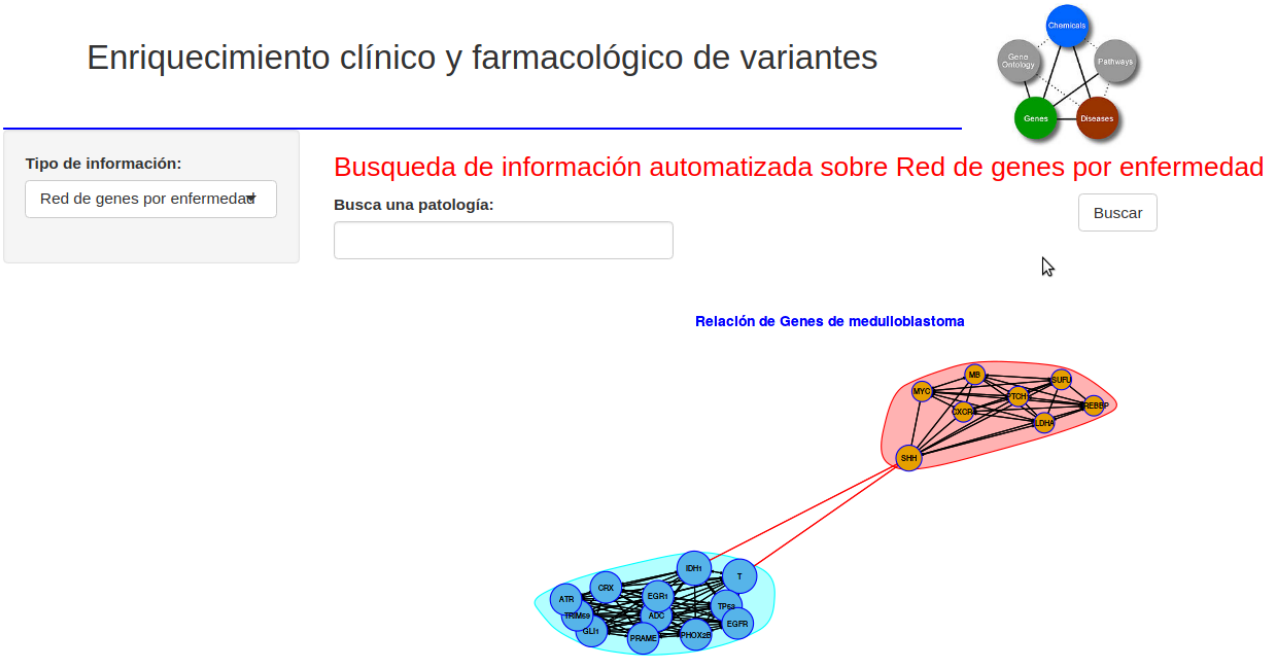


Figura 6. Red de genes creada basada en análisis de semántica latente

Para el desarrollo del módulo llamado “Similitud de genes basado en su geontología” hemos utilizado distintas medidas de similitud semántica que permiten obtener valores numéricos en función de la cercanía del significado entre términos de una ontología y que se explican a continuación (figura 6).

El proyecto Ontología de genes (GO, del inglés Gene Ontology) mantiene un vocabulario dinámico, estructurado, definido con precisión y controlado, de términos para representar las funciones y las localizaciones celulares de las proteínas de una manera independiente de la especie. Comprende tres ontologías ortogonales: componente celular (CC), función molecular (MF) y proceso biológico (BP). Estas ontologías están estructuradas como tres gráficos acíclicos dirigidos (DAG) en los cuales, los nodos corresponden a los términos que describen una cierta categoría semántica biológica y los bordes representan los enlaces entre los términos que describen relaciones definidas [11]. Las relaciones más comunes son 'is-a' y 'part-of'.

Los términos GO se han utilizado ampliamente para anotar proteínas, como las del proyecto de anotación genética (GO Ontology Annotation, GOA) [12]. Este procedimiento permite usar la similitud semántica para comparar proteínas basadas en la función en lugar de su similitud de secuencia [13,14].

2.6.1 Términos de la Ontología de Genes

Todos los términos de la GO tienen un nombre y un identificador único de la forma GO:nnnnnnn, la mayoría con una definición textual, con referencia a la fuente donde fue descrito. Cualquier observación necesaria se incluye en un campo de comentarios.
GO utiliza sinónimos en un sentido amplio, pues no es necesario que los nombres dentro del campo 'sinónimo' signifiquen exactamente lo mismo que el término al que están vinculados; esta flexibilidad resulta muy útil en búsquedas y varias aplicaciones como la minería de texto.
Muchas funciones, procesos y componentes no son comunes a todas las especies; sin embargo, el objetivo de GO es desarrollar un vocabulario capaz de describir cualquier organismo. Con este propósito, GO acordó incluir cualquier término aplicable a más de una clase taxonómica. Para especificar la clase de organismo en cuestión, GO utiliza el conector sensu, ("en el sentido de").
Los términos que se consideran obsoletos se marcan como 'obsolete', pero tanto el término como su identificador se mantienen en la base de datos de GO, por lo general, se añade un comentario que explica su caducidad y se sugiere un término actual para reemplazar el término obsoleto.

2.6.2 Divisiones de la Ontología de Genes

La Ontología de Genes agrupa realmente tres ontologías que se corresponden con tres aspectos diferentes de la biología celular: función molecular, proceso biológico y componente o localización sub-celular. Aunque la GO incluye fundamentalmente conceptos que se refieren al nivel sub-celular y celular, abarca también niveles superiores, como los correspondientes a sistemas órganos y a organismo.

Proceso biológico (PB): Los PBs implican generalmente transformaciones químicas o físicas que ocurren por la acción de un conjunto de funciones moleculares organizadas; es decir, el objeto que va a un PB sufre transformaciones que lo convierten en algo diferente. Los PBs pueden ser de un nivel más elevado o abstracto, como son el "crecimiento celular" o la "transducción de señales", o de un nivel menor o más específico como son el "metabolismo de pirimidinas" o la "biosíntesis de AMPc".

Función molecular (FM): Describe actividades que ocurren a nivel molecular; sus términos representan a las actividades y no a las entidades (moléculas o complejos moleculares) que llevan a cabo las acciones, sin especificar cuándo, dónde, o en qué contexto ocurren. Para evitar confusiones entre los nombres de los productos génicos y las FMs, muchos términos incorporan la palabra actividad ("*activity*").

Componente celular (CC): Se refiere al espacio celular donde se encuentra el producto génico. Un componente celular puede ser una estructura anatómica, como el retículo endoplasmático, el núcleo celular, o una estructura molecular más simple formada por productos génicos, como un ribosoma o un dímero proteico.

2.6.3 Estructura de las ontologías

Los términos en GO se organizan en un grafo acíclico dirigido (GAD) en el cual los términos son vértices o nodos y las relaciones entre ellos son los arcos. En un GAD, los arcos son unidireccionales, no existen ciclos y un nodo "hijo" puede relacionarse con diferentes nodos "padres". Los términos heredan las relaciones y propiedades de sus nodos padres. En un inicio, las relaciones entre los términos de GO eran de dos tipos fundamentales: *"is_a" y "part_of"*. En el 2008, el Consorcio agregó tres relaciones: "*regulates", "positively_regulates" y "negatively_regulates"* para representar la relación entre procesos que afectan otros procesos sin ser parte de ellos. Un año después, se incorporó la relación "*has_part"* que representa una relación parte-todo, pero desde la perspectiva de un nodo padre y es por tanto, un complemento lógico de la relación "*part_of*". GO no relacionaba las tres ontologías entre sí, recientemente se han establecido relaciones entre PB y FM: existen relaciones *"part_of"* entre FM y PB y regulates entre FM y PB.

2.6.4 Anotaciones de la Ontología de Genes

Toda anotación basada en la GO requiere de un código de evidencia que registra las condiciones en que se registra y se realizó la anotación. Los códigos de evidencia caen en cuatro categorías fundamentales: experimental, computacional, derivado indirectamente de cualquiera de las categorías anteriores, o desconocido. En estos momentos, se utilizan 17 códigos diferentes. Si se desconoce el proceso, función o localización de un gen, se anota en el nodo raíz con el código de evidencia ND ("*no biological data available"*). Las anotaciones ND permiten diferenciar genes no anotados de genes no caracterizados. Otra característica de GO es su actualización constante, incluyendo los mapeos entre términos GO y otros descriptores.

2.6.5 Análisis semántico basado en la Ontología de Genes

Las medidas de similitud semántica permiten obtener valores numéricos en función de la cercanía del significado entre términos de una ontología, o entre los conjuntos de anotados a determinadas entidades [15,16]. La aplicación de las medidas de similitud semántica entre las anotaciones de GO proporciona una medida de su similitud funcional. En la actualidad, están disponibles diversas propuestas para cuantificar la similitud semántica.

Los términos en una ontología con estructura de grafo como GO pueden compararse mediante dos vías fundamentales, dependiendo de que se utilicen los nodos o los arcos como fuente de datos.
Los métodos que emplean los arcos se basan fundamentalmente en contar el número de arcos entre dos nodos del grafo. Estas metodologías, aunque intuitivas; se basan en dos supuestos que son ciertos en muy raras ocasiones en la biología: (1) los nodos y los arcos están distribuidos uniformemente y (2) los arcos en el mismo nivel de la ontología se corresponden con igual distancia semántica entre los términos.

Los métodos basados en nodos utilizan las propiedades de los términos implicados, que pueden relacionarse con los propios términos, sus ancestros o descendientes. Uno de los conceptos que se emplea con más frecuencia en estos casos es el Contenido de Información (CI), el cual puede calcularse en base a la ocurrencia de un término en una

base de datos, o a partir del número de hijos que tiene un término en GO, aunque esta variante es menos empleada. Los métodos basados en el CI son menos sensibles que los métodos basados en arcos a la variabilidad de distancia semántica y densidad de nodos, pues el CI es una medida de la especificidad de un término independiente de su profundidad en la ontología. No obstante, el CI está sesgado por las tendencias actuales de la investigación biomédica, pues aquellos términos de interés científico tienen más probabilidad de estar anotados. El uso del CI todavía tiene sentido desde el punto de vista probabilístico, pues es más probable (y menos significativo) que dos productos génicos compartan términos usados con mucha frecuencia, independientemente de que el termino sea común por ser genérico o por estar relacionado a una temática de investigación activa.

El análisis ontológico puede realizarse con diferentes modelos estadísticos incluyendo el hipergeométrico, binomial, χ^2 (chi-cuadrado) y el test de Fisher, sin embargo, una de las más utilizadas para modelar la función de probabilidad es la distribución hipergeométrica basado en los siguientes supuestos:

1. Al realizar un experimento con este tipo de distribución, se esperan dos tipos de resultados (ancestro común compartido o no compartido).
2. Las probabilidades asociadas a cada uno de los resultados no son constantes (la probabilidad de un ancestro común compartido depende de cada término GO).
3. Cada ensayo o repetición del experimento no es independiente de los demás (los términos GO están relacionados unos con otros por lo que la probabilidad de tener un ancestro común va a depender de las anotaciones que se hayan hecho para cada término GO).
4. El número de repeticiones del experimento (n) es constante (el cálculo de probabilidad de los ancentros comunes se basan en todas las combinaciones posibles para cada ancestro teniendo en cuenta el número de arcos o nodos para ese ancestro común, por lo que no se permiten repeticiones ni un orden específico).

2.6.6 Similitudes de Lin, Jiang y Conrath, y Resnik

Las medidas más relevantes y usadas como base en multitud de trabajos son las propuestas por Resnik, Lin, Jiang y Conrath.
Concretamente, estas medidas basan la comparación entre términos buscando el ancestro común más bajo (Lowest Common Ancestor, LCA) dentro de la jerarquía GO.
Supóngase que la información contenida por un GO–term A es:

$$IC(A) = -\log(p(A))$$

Donde p(A) es la probabilidad de que un término ocurra en el conjunto de anotaciones bajo consideración:

$$p(A) = freq(A)/\ freq(root)$$

“Root” representa al término raíz de una de las tres ontologías y freq(root) es el número de veces que un gen es anotado con algún término de la ontología. Mientras que freq(A) es dado por:

f req(A) = |annot(A)| + ∑ c∈children(A) |annot(c)|

Siendo *children(A)* el conjunto de todos los términos hijos del termino A. Es decir, el conjunto de todos los términos para los que A es un término padre, ya sea directa o indirectamente.

Resnik calcula la similitud entre dos términos usando sólo la información contenida (IC) del LCA compartido entre dos términos A y B:

simRes(A, B) = IC(LCA(A, B))

Por otro lado, la medida de similitud de Lin toma en cuenta los valores de IC para cada uno de los términos A y B, además del LCA compartido por los dos términos:

simLin(A, B) = 2 × IC(LCA(A, B))) IC(A) + IC(B)

Mientras que Jiang y Conrath propusieron un IC basado en distancia semántica, la cual puede ser transformada en la siguiente medida de similitud: simJiang(A, B) = 1 IC(A) + IC(B) − 2 × IC(LCA(A, B)) + 1. Para cada una de estas medidas, cuanto mayor sea el valor obtenido mayor similitud semántica presentan los dos términos. El menor valor posible es el 0, mientras que el mayor valor es 1.

Debido a la estructura de gráfo acíclico dirigido de GO, es posible que un término GO presente diferentes padres con lo que se podría dar el caso que dos términos puedan compartir ancestros en diferentes caminos hacia el nodo raíz de la ontología. Lord et al. , además de adaptar las medidas de Jiang, Lin y Resnik, fueron los pioneros en proponer una nueva medida de similitud semántica teniendo en cuenta esta posibilidad de GO. Esta medida, que denominamos simLord, se basa en encontrar el ancestro común más informativo *(probability of the minimum subsumer, pms)*:

pms(A, B) = minc∈S (A,B)p(c)

Donde S (A, B) es el conjunto de términos ancestros compartidos por A y B. Quedando la similitud entre dos términos según Lord et al. Como:

simLord(A, B) = −ln(pms(A, B))

2.6.7 Basadas en la representación del DAG como GO–árbol

En 2004, Lee et al., abordaron una medida de similitud semántica basada en una novedosa variación de la estructura de GO [17]. La variación, denominada GO–árbol, consiste en mapear la estructura de grafo de GO a un árbol. Para ello, un GO–término en la estructura GO–árbol sólo podrá tener un padre aunque éste puede estar representado en diferentes partes del árbol. De esta forma, un GO–término aparecerá

tantas veces en el GO–árbol como caminos ascendentes diferentes tenga éste en el DAG para acceder al nodo raíz de la ontología en cuestión.
A partir de esta estructura, la medida de similitud, que denominaremos simLee, se calcula como el peso del LCA de los términos que le ocupa.

2.6.8 Similitud entre Gen–Productos o proteínas

Las medidas expuestas en el apartado anterior tienen la intención de medir la similitud entre dos GO–terminos, y deben ser extendidas para comparar gen–productos o proteínas [18,19]. Un gen–producto posee uno o más términos GO asociados, con lo que una aplicación directa de las medidas expuestas anteriormente no sería posible. Al conjunto de GO–terminos anotados o asociados a un gen–producto o proteína "gp" lo denominaremos anotaciones (Anotgp).

Lord et al. fueron los pioneros en proponer una solución al problema anterior. Éstos presentaron la similitud entre dos gen–productos A, B (simP(A, B)) como la combinación de las similitudes de los distintos GO–terminos asociados a éstos. Cada término asociado a A es comparado con todos los términos asociados a B, obteniendo por cada pareja un valor de similitud entre GO–terminos [14]. Estos valores son usados para producir una medida final de similitud entre pares de gen–productos. Concretamente, Lord et al. propusieron la media aritmética como combinación de las similitudes de todos las posibles parejas de GO–terminos.

$$simPavg(A, B) = \sum tA \in AnotA \wedge tB \in AnotB \; sim(tA, tB) \; |AnotA| \times |AnotB|$$

Posteriormente, en la literatura existen diferentes aproximaciones basadas en el mismo concepto de combinar las diferentes similitudes entre parejas de GO–terminos [20,21,22]. Estas combinaciones son:

- Máximo(simPmax): Selecciona el máximo nivel de similitud encontrado como similitud de los gen–productos.

$$simPmax(A, B) = max\{sim(tA, tB) | tA \in AnotA \wedge tB \in AnotB\}$$

- Suma(simPsum): Es calculado como la suma de las similitudes de todas las parejas de GO–términos. $simPsum(A, B) = \sum tA \in AnotA \wedge tB \in AnotB \; sim(tA, tB)$

- Match: Engloba a las combinaciones que sólo tiene en cuenta aquellas parejas de GO–términos que son iguales. En esta categoría se podrían encontrar las combinaciones de media aritmética (simPavg_M), máximo (simPmax_M)

- Best Pairs: Engloba a aquellas combinaciones que sólo tienen en cuenta aquellas parejas de GO–términos que maximizan la similitud sin repetición de GO–términos. De esta forma, se selecciona todos los términos de aquel soporte que contenga menos GO–términos, y éstos son asociados con el GO–término del soporte del otro gen–producto que maximice la similitud entre ambos. Además, cada GO–término, independientemente del soporte al que pertenezca, sólo puede ser usado una vez en el cálculo de similitud de términos GO. En esta categoría se

encuentran tanto la media aritmética (simPavg_BP) como la suma (simPsum_BP).

- Best Match Average(BMA)(simPavg_BM): Se determina como la media aritmética de las mejores similitudes. Con tal fin, se escoge el soporte con menor número de GO–términos y por cada uno de los términos en éste es seleccionado un término del soporte del otro gen–producto que maximice la similitud entre ambos.

2.6.9 Busqueda de enfermedades para un conjunto de genes

El uso de representaciones GO puede ampliar nuestra comprensión de las enfermedades al ofrecer formas alternativas de estudiar la similitud entre ellas. Estudiar las similitudes de las enfermedades puede arrojar luz sobre la etiología, revelar una fisiopatología común y/o sugerir un tratamiento que pueda ser apropiado de una enfermedad a otra. Se ha descubierto que varias enfermedades que previamente se creía que eran distintas compartían procesos biológicos en su etiología o en la manifestación de los síntomas. La información genética, los síntomas y el fenotipo junto con los modelos de penetrancia se han utilizado para encontrar la comorbilidad entre las enfermedades.

Una tendencia concomitante en la investigación ha sido el uso creciente de datos anotados con ontologías biomédicas. En particular, estos conjuntos de datos pueden ser explotados para revelar las relaciones entre las entidades biológicas (Gene Ontología) y patología de la enfermedad (UMLS). Los datos pueden combinarse en ontologías y similitudes entre términos cuantificados mediante el uso de métodos computacionales. La similitud semántica entre enfermedades se puede evaluar calculando la similitud entre los conjuntos de términos ontológicos asociados.

2.7 Servidor Shiny

El paquete Shiny proporciona las funciones que se ocupan de la comunicación entre el Cliente y el Servidor utilizando la programación reactiva (Reactive), y además genera el HTML5/JavaScript/CSS necesario para construir las Aplicaciones Web.
La aplicación Shiny está conformada por un archivo app.R que contine tanto los elementos de la interfaz como del servidor.
Shiny se encarga de generar el HTML/JavaScript y además ofrece toda la lógica de manejo de los eventos que se producen en pantalla (clicks a botones, cambios de menú, etc), cuando cambia un parámetro en pantalla, Shiny también se encarga de comunicarse con el servidor enviando el nuevo valor del parámetro y el servidor procesa la llamada, genera el resultado y lo envía al cliente.

En el objeto ui está todo el código HTML necesario para mostrar la información en la pantalla y los componentes y parámetros que la hacen interactiva. En el objeto server se encuentra definida la lógica que se ejecuta cada vez que el usuario hace un cambio o una interacción en pantalla.

La función fluidPage() genera código HTML. La función titlePanel(), la creación de una pantalla con una barra lateral y un panel principal, mediante las funciones sidebarPanel() y mainPanel() respectivamente. Ambos componentes se colocan dentro

de la llamada a la función sidebarLayout() que genera una página dividida en una barra lateral y un panel principal.

La forma en la cual se relacionan las interacciones o cambios de los usuarios con los resultados, se realizan en el servidor mediante las variables input y output.
En la interfaz hemos creado distintas variables input en la llamada a selectInput() denominada *'gen_id' y 'information'*. En estas variables están almacenadas las acciones seleccionadas. La forma de acceder a este valor desde el servidor es mediante la expresión input$*gen_id o get("input")[["gen_id"]]*. Adicionalmente se crean variables de salidas en la llamada a Output() que se genera en el servidor mediante la expresión output$name. Es decir, que el servidor utiliza una familia de funciones que convierten los objetos input$ en resultados para su interfaz de usuario. Esto se realiza mediante la función renderTipo() y son las siguientes:

- renderDataTable DataTable
- renderImage images (saved as a link to a source file)
- renderPlot plots
- renderPrint any printed output
- renderTable data frame, matrix, other table like structures
- renderText character strings
- renderUI a Shiny tag object or HTML

Los resultados Output() pueden ser de los siguientes tipos:

- dataTableOutput DataTable
- htmlOutput raw HTML
- imageOutput image
- plotOutput plot
- tableOutput table
- textOutput text
- uiOutput raw HTML
- verbatimTextOutput text

La aplicación Shiny posibilita una programación reactiva o dinámica lo que permite que la aplicación se actualice instantáneamente cada vez que el usuario realice un cambio. Cuando un usuario cambia algo en el navegador, el navegador envía el nuevo valor de entrada al servidor, lo que desencadena un evento de descarga e invalida todos los dependientes de la entrada. Luego, todos los puntos finales reactivos (como los observadores y las salidas) se actualizan, lo que también puede desencadenar una actualización de los conductores reactivos. Sin embargo, las cosas se vuelven más complicadas cuando se tienen llamadas para actualizar las entradas (como updateSelectInput). En estos casos no se ejecutarán hasta que todos los reactivos actualmente invalidados hayan terminado de ejecutarse, luego el servidor envía mensajes de vuelta al navegador para todas las llamadas de entrada de actualización. Si esto ocasiona algún cambio en los valores de entrada, continúa otro ciclo de actualización. Entonces, si no se tiene cuidado, es bastante fácil terminar con un ciclo infinito si las llamadas de entrada de actualización realmente cambian algunos valores de entrada .

3. RESULTADOS

PLATAFORMA WEB ENRICH_GEN

SECCIONES DE LA PLATAFORMA

1. Visión general de la plataforma de enriquecimiento clínico y farmacológico de genes
2. Evidencia científica por genes
3. Relación genes y fármacos
4. Ensayos clínicos por genes
5. Indicación FDA por fármacos
6. Red de genes por enfermedad
7. Similitud de genes

    7.1 Panel de Genes
    7.2 Similitud de genes
    7.3 Ontología Génica
    7.4 Enfermedades asociadas

8. Ejemplo: secuenciación del exoma completo (WES) de un cáncer de mama metastásico. Informe realizado utilizando Enrich_Gen.

1. Visión general de la plataforma de enriquecimiento clínico y farmacológico de genes

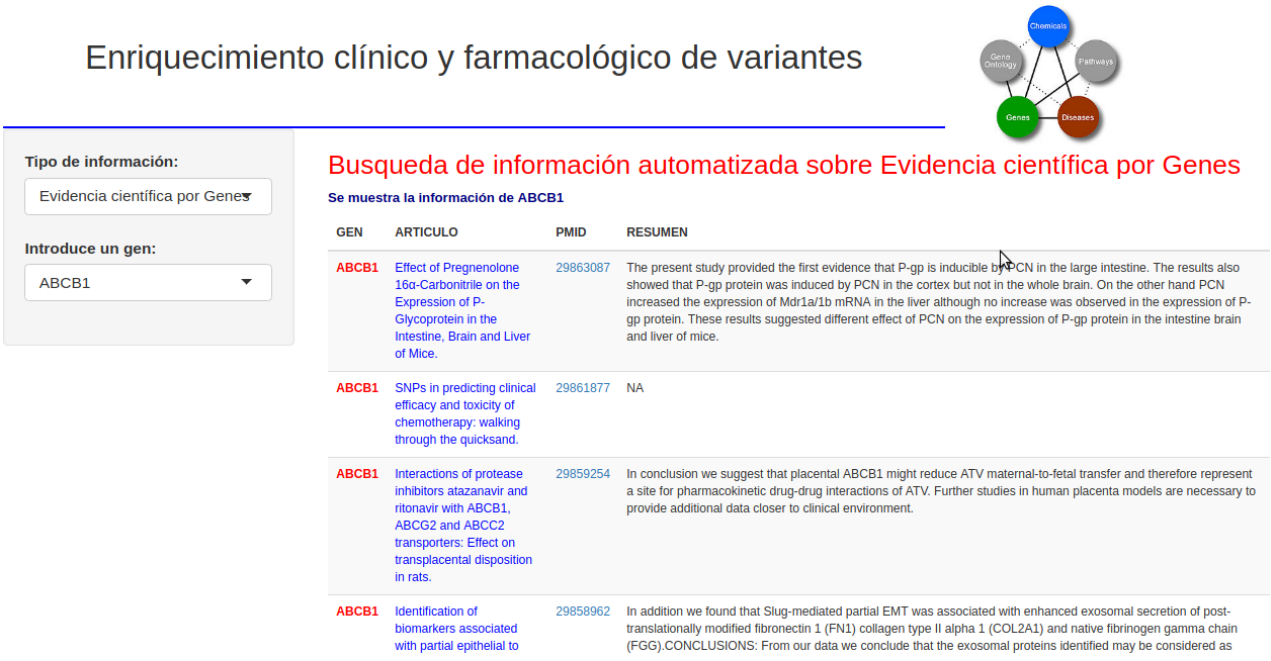
La plataforma permite una anotación clínica y farmacológica de un conjunto de genes uno a uno, o a partir de un panel de genes que puede ser cargado como un fichero en formato de arhivo de texto (.txt). Dicho fichero tiene el siguiente formato: el nombre de cada gen organizados en lineas y separados por comas.

EnrichGen es una herramienta web que permite al usuario extraer información clínica, farmacológica y biológica de un conjunto de genes por medio de dos tipos de consultas diferentes. El primer tipo de consulta se realiza a partir de la busqueda de información en Pubmed y proporcionando el listado de los artículos más relevantes asociados a dicho gen, junto con un resumen del mismo. En las bases Clinical Trials, la base de datos de interacción de genes y fármacos (DGIdb) y la FDA podemos encontrar la información necesaria para tener suficiente conocimiento sobre un conjunto de genes y poderlos anotar.

El segundo tipo de consulta se basa en la verificación de la similitud de un grupo de genes con respecto al conjunto de términos GO asociados a cada gen, y la busqueda de las enfermedades asociadas a cada gen en base a que dichas enfermedades comparten términos GO similares.

2. Evidencia científica por genes

La información es mostrada en una tabla con cuatro columnas: en la primera se muestra el nombre del gen; en la segunda, el título del artículo; en la tercera, el identificador de PubMed; y en la cuarta columna, la conclusión del artículo.



3. Relación genes y fármacos

En esta tabla se muestran los fármacos asociados a un gen en particular. La información procede de una consulta a la base de datos DGIdb que ayuda a anotar genes de interés con respecto a las interacciones conocidas entre el fármaco y el gen, y la potencial farmacobilidad.

Una interacción es una asociación entre un gen y un fármaco en particular. Por ejemplo, la interacción fármaco-gen, SUNITINIB-FLT3, se informa como de tipo 'inhibidor'. El tipo de interacción, que es usada en DGIdb, es muy amplio. Actualmente hay definidos decenas de tipos de interacción, y muchos tipos de interacción describen el mecanismo de acción entre una molécula pequeña y una proteína. Sin embargo, también podrían usarse otros tipos más amplios de "interacción". p.ej. Gene X como 'resistencia' o 'sensibilidad' a la droga Y.

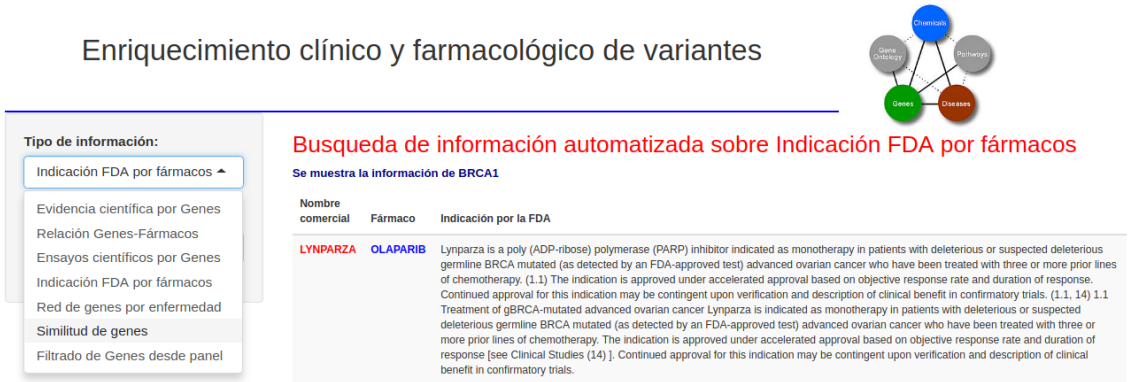
Esta base de datos se limita a genes humanos, y lo ideal es utilizar los símbolos según el Comité de Nomenclatura de Genes de HUGO y que se puede encontrar en el campo “Gene” de la base de datos NCBI, y separado por comas.

4. Ensayos clínicos por genes

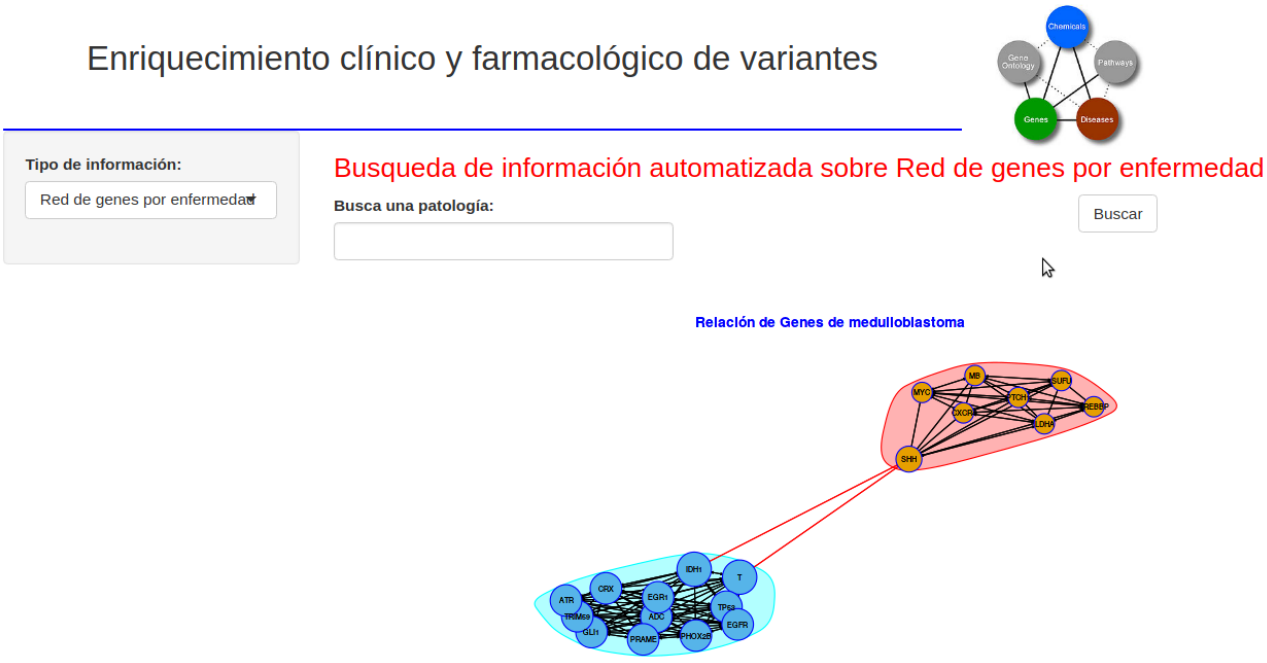
Esta tabla proporciona información de los ensayos clínicos disponibles para un gen particular, la cual se ha extraido de la base de datos *ClinicalTrials.gov*. *ClinicalTrials.gov* es una base de datos de estudios clínicos financiados de forma privada y pública realizada en todo el mundo. Actualmente dispone información de 273.805 estudios de investigación de 204 países.
La información está estructurada en cinco columnas: titulo del ensayo, indicación, estado, fecha de apertura o cierre del ensayo y tipo de estudio. Esta información está totalmente actualizada dado que EnrichGen realiza una consulta en tiempo real.

5. Indicación FDA por fármacos

Esta tabla muestra información de los fármacos asociados a este gen y aprobados por la FDA (*Food and Drug Administration*). La información esta estructurada en una tabla con tres columnas: la primera muestra el nombre comercial de la droga; la segunda, el nombre farmacéutico de la droga; y la tercera, información acerca la indicación, posologia, mecanismo de acción, toxicidad, etc.



6. Red de genes por enfermedad



La asociación entre genes y términos tienen el potencial de revelar la conexión subyacente entre genotipo y fenotipo. Estas asociaciones se computan generando una matriz de documentos de términos y los genes encontrados en cada documento, y que constituirá la matriz de ocurrencias.

Posteriormente se aplica el procedimiento de la descomposición del valor singular (SVD), y de este modo la matriz original (x) es descompuesta en el producto de tres matrices (Análisis semántico latente). Las matrices resultantes contendrán “vectores singulares” y “valores singulares”, estos últimos contienen la variabilidad para cada dimensión a lo largo de los términos y documentos. Una matriz contendrá la representación de los términos (T), cuyos componentes o factores serán linealmente independientes de la relación con los documentos en la matriz original. Otra matriz contiene la representación de los documentos (D) de la misma forma que la de términos, es decir, como vectores singulares cuyos componentes son linealmente independientes

de la relación con los términos en la matriz original. Por último, una matriz diagonal (S) de valores singulares escalados (de mayor a menor aportación para agrupar) y cuya aportación es que la matriz independiente de términos multiplicada por ella y, a su vez, multiplicada por la matriz traspuesta de la matriz independiente de documentos, reconstruyen la matriz inicial.

Las filas de las matrices reducidas de vectores singulares se toman como coordenadas de los puntos que representan a los documentos y términos en un espacio de dimensión k cuyos ejes están reescalados por cantidades relacionadas con los valores de la diagonal de la matriz S. Los productos escalares (coseno del ángulo) entre los puntos nos darán las relaciones de similitud entre los distintos puntos.

El grafo resultante y que se muestra es un gráfico de nodos y conexiones que es creado a partir de la matrix de similitud de genes resultante del análisis semántico latente y después de comparar par a par las relaciones entre el conjunto de genes extraido para el término buscado.

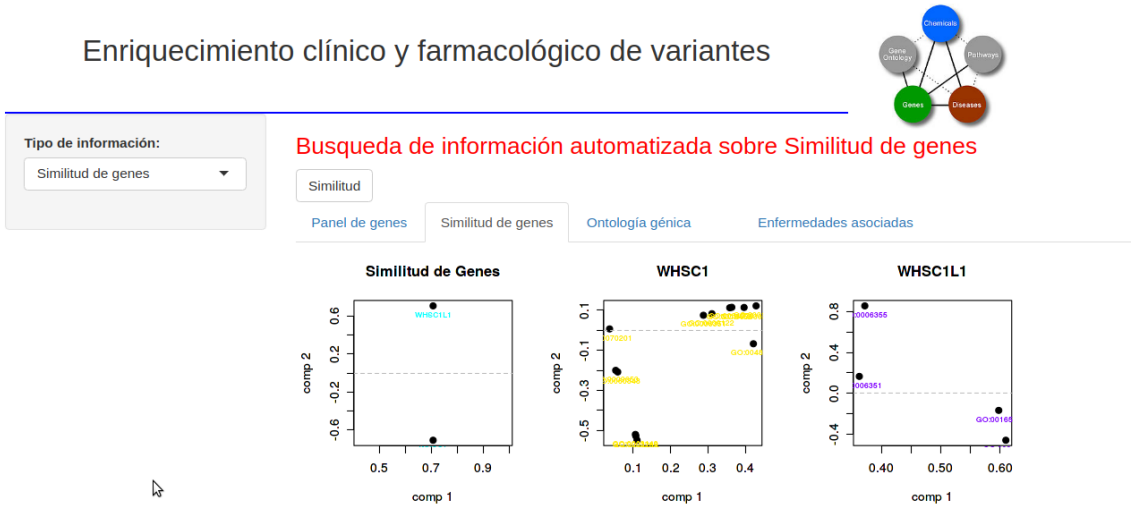
7. Similitud de genes

7.1 Panel de Genes

Plantilla para busqueda de similitud de genes

Para ejecutar la busqueda de similitud se debe seleccionar al menos 2 genes > presionar el botón “Similitud” y luego a la pestaña “Similitud de Genes”. Es importante saber que el tiempo de procesamiento del cálculo aumenta de manera logarítmica a medida que se seleccionan un número mayor de genes, por lo que lo recomendable es seleccionar como máximo 4 genes.

7.2 Similitud de genes

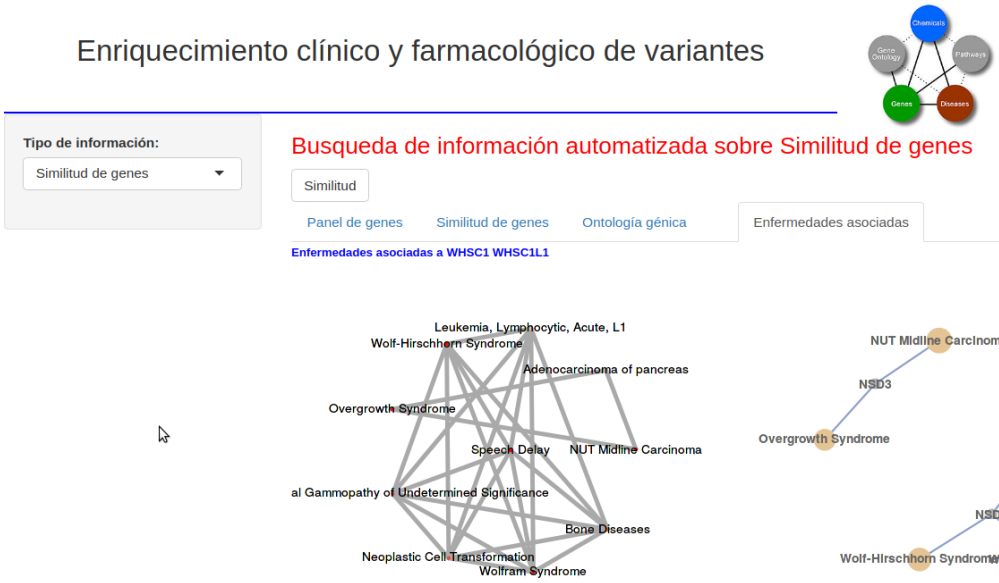


Los gráficos resultantes muestran la proximidad de genes basado en el análisis semántico de la Ontología de los mismos. Las medidas de similitud semántica permiten obtener valores numéricos en función de la cercanía del significado entre términos de una ontología, o entre los conjuntos de anotados a determinadas entidades. Estas medidas se basan en la comparación entre términos buscando el ancestro común más bajo, y siguiendo la métrica propuesta por Resnik [23].

7.3 Ontología Génica

La tabla muestra los términos GO asociados a cada gen seleccionado. Todos los términos de la GO tienen un nombre y un identificador único de la forma GO:nnnnnnn, y la mayoría con una definición textual. La Ontología de Genes agrupa realmente tres ontologías que se corresponden con tres aspectos diferentes de la biología celular: función molecular, proceso biológico y componente o localización sub-celular.

7.4 Enfermedades asociadas



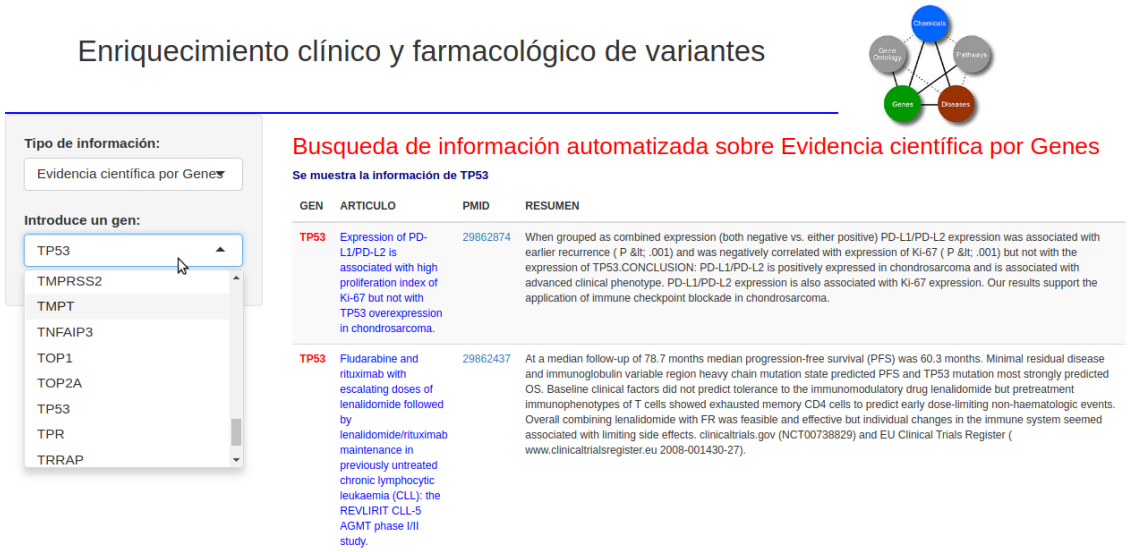
Los gráficos mostrados proceden de un análisis semántico latente en el que la matrix de documentos-términos proceden de la busqueda en PubMed de todas aquellas enfermedades asociadas a un conjunto de genes seleccionados. La nodos constituyen cada una de las entidades que formarán las filas de la matriz y los arcos se calculan tras el cálculo del coseno entre los vectores que forman cada una de las entidades y que servirán de base para establecer la similitud.

8. Ejemplo: secuenciación del exoma completo (WES) de un cáncer de mama metastásico. Informe realizado utilizando Enrich_Gen.

Variantes somáticas en el tumor metastático - Ganglio

ARID5B	chr10:63851403	Exón 7 c.C1452G p.D484E	Sustitución no sinónima	
NOTCH2	chr1:120458633	Exón 34 c.6711insT p.2238fs	Inserción con desplazamiento del marco de lectura	
DOT1L	chr19:2216683	Exón 20 c.C2327T p.P776L	Substitución no sinónima	
HNF1A	chr12:121431484	Exón 3 c.G688T p.E230X	Codón de parada	
TP53	chr17:7574005	Exón 6 c.T626A p.F209Y		

Tabla evidencia científica



Esta tabla muestra la evidencia científica para cada uno de los genes y publicada en PubMed. A partir del resumen de esta tabla podemos extraer los siguientes datos:

1. Un mecanismo crítico que induce la trasnformación epitelio-mesenquimal es la presencia de la mutación en el gen NOTCH2 mediada por STAT5, JAG-1 y DLL4. Por otra parte, NOTCH2 puede inhibir la angiogénesis y el crecimiento del cáncer de mama mediada por el gen MINAR1 que se expresa ampliamente en diversos tejidos, incluidas las células epiteliales de la mama y las células endoteliales de los vasos sanguíneos (1)

2. las mutaciones en el gen DOT1L están involucrado en la regulación de la transición epitelio-mesenquima en el cáncer de mama. DOT1L induce la transformación

neoplásica de las células epiteliales de mama, a través de la activación transcripcional dependiente de DOT1L de genes promotores, tales como Snail, ZEB1 y ZEB2. Por lo tanto, DOT1L puede determinar la agresividad del cáncer de mama mediante la activación de factores promotores de la transformación epitelio-mesenquimal (2).

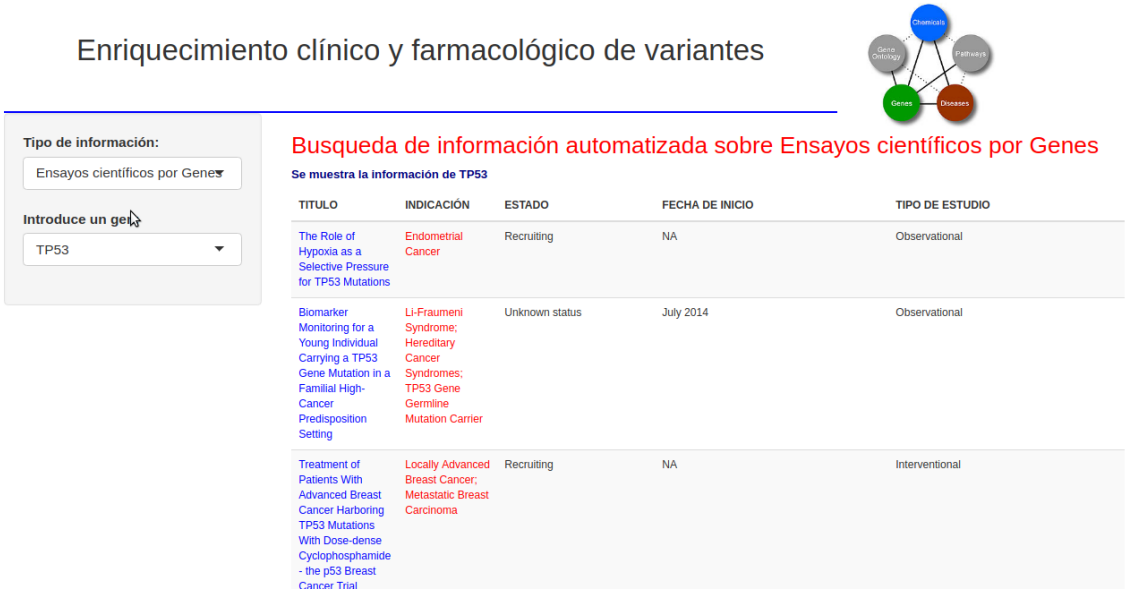
3. El gen ARID5B tiene la capacidad de regular la transcripción y participan en la desdiferenciación y proliferación celular, por lo que contribuirá a mantener y perpetuar la transición epitelio-mesenquima. De todas las enfermedades malignas, la leucemia es la más estrechamente relacionada con ARID5B. La creciente evidencia muestra que las mutaciones y los polimorfismos de un solo nucleótido de ARID5B se asocian con el desarrollo de leucemia linfoblástica aguda y el resultado del tratamiento, de la leucemia mieloblástica aguda en la infancia, pero no hay estudios que permitan establecer su valor pronóstico y predictivo en el cáncer de mama.

4. Las mutaciones en TP53 se asocian con tumores altamente aneuploides posiblemente porque estos tumores presentan también conjuntamente sobreexpresión de los activadores transcripcionales mitóticos que aumentan la tasa de cromosomas de anafase rezagados y contribuyen a la mala respuesta a la quimioterapia en el cáncer de mama.

Tabla fármacos por genes

Esta tabla muestra el conjunto de fármacos experimentales dirigidos a cada gen. El número de códigos PMID apoya la evidencia científica de cada fármaco.

Tabla ensayos clínicos por genes



Esta tabla muestra los ensayos clínicos disponibles para cada gen, el tipo tumoral elegible para el ensayo, la fase del ensayo, la fecha de inicio y el tipo de estudio. En esta tabla podemos encontrar los siguientes datos en relación al caso que presentamos:

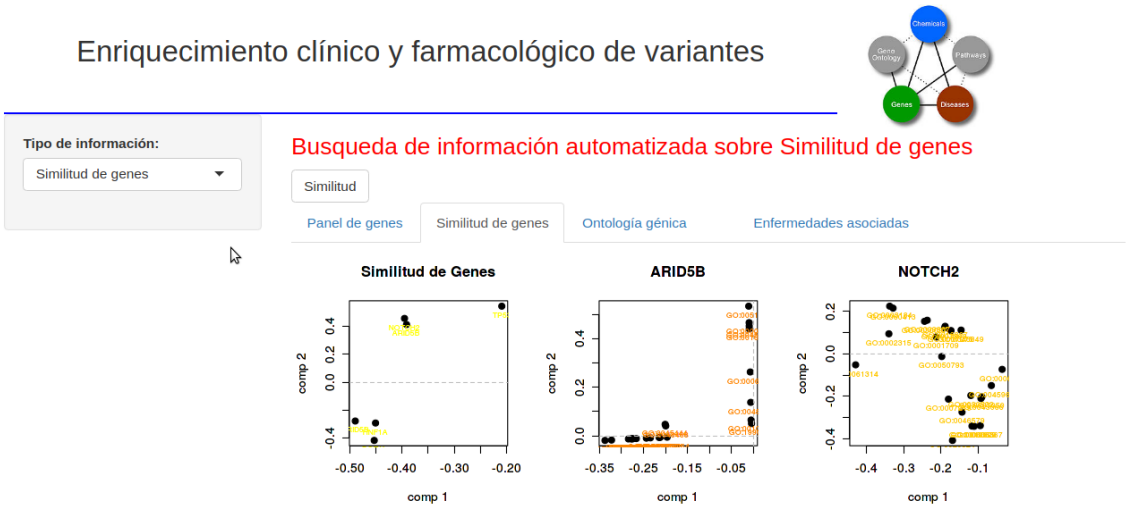
1. El bloqueo NOTCH2 mediante la combinación de MK-0752 y Tocilizumab inhibe significativamente el crecimiento tumoral y, por lo tanto, podría servir como una estrategia terapéutica novedosa para tratar mujeres con mama que expresa mutaciones en NOTCH2 / NOTCH3.

2. La utilización de dosis densas de ciclofosfamida disponibles para pacientes con cáncer de mama avanzado y portadores de mutaciones en TP53, basado en que los carcinomas de mama localmente avanzados no inflamatorios con mutaciones enTP53 tenían una alta tasa de respuesta patológica completa a la quimioterapia con dosis densas de doxorrubicina-ciclofosfamida (4).

Gráfico 'Red de genes por enfermedad': cáncer de mama metastásico

Este gráfico muestra el conjunto de genes descritos en PubMed en documentos que incorporan en el título los términos «breast cancer». Se observan dos redes de genes que no están interconectadas, lo cual indica que los genes que aparecen en el gráfico azul aparecen en unos documentos, y los genes que aparecen en naranja aparecen en otros documentos. Esto puede ser debido a que los genes en el gráfico azul son genes descubiertos recientemente, y sobre los cuales el conocimiento es mucho más pequeño por lo que las implicaciones con el cáncer de mama también son menores en el momento actual. La red de genes en el gráfico naranja corresponden a los genes más conocidos en el cáncer de mama y sobre los que el conocimiento y las relaciones génicas son más notables.

Gráfico 'Similitud de genes':



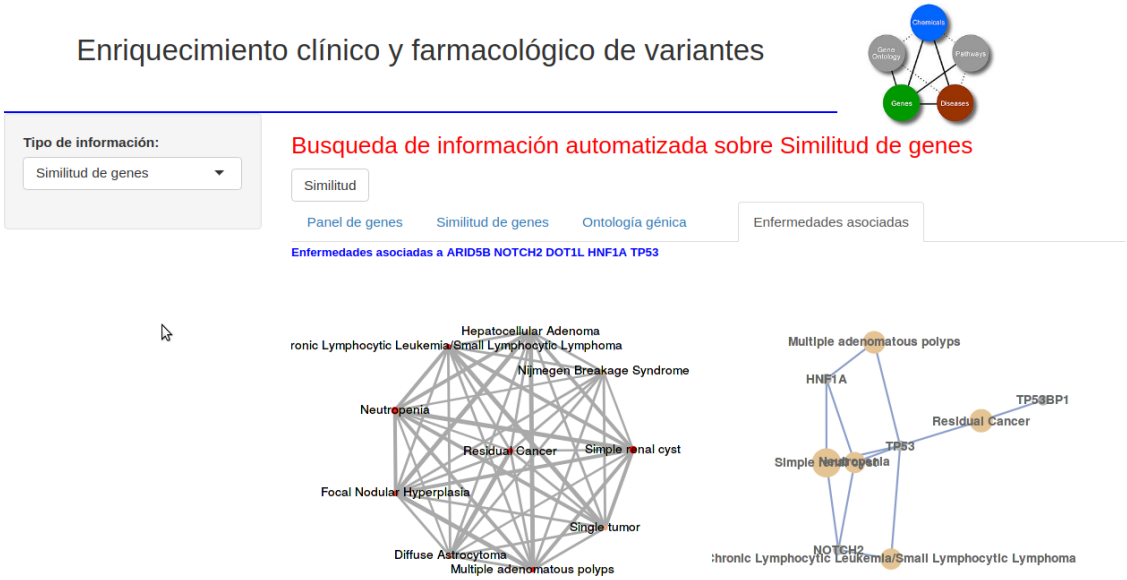
Este gráfico muestra la relación geométrica de los genes que han aparecido mutados en el estudio y catalogados como variantes patológicas, basado en la similitud semántica de los términos GO de cada gen. De este modo, con estos gráficos podemos llegar en el ejemplo que presentamos a las siguientes conclusiones:

El estudio genómico del tumor ha mostrado cinco genes mutados que podemos agrupar en tres grupos basado en su similitud funcional: uno estaría formado por los genes NOTCH2 y ARID5B que intervienen en la regulación negativa de la transcripción del promotor de la ARN polimerasa II; otro por DOT1L y HNF1A que actúan como puntos de control del daño del ADN e intervienen en la reparación no homóloga de las roturas de doble cadena, asi como regulan positivamente la transcripción del promotor de ARN polimerasa II; y el último por TP53 que presenta una mayor similitud funcional a los genes NOTCH2 y ARID5B, y además interviene en la reparación de escisión de base y escisión de nucleótidos del ADN.

Ontología Génica

En esta tabla se muestra los términos GO asociados a cada gen. Esta información es utilizada para encontrar el ancestro común más próximo para cada conjunto de genes, y de este modo poder medir la distancia de proximidad o similitud.

Enfermedades asociadas



Siguiendo el mismo planteamiento que en la gráfica anterior de red de genes, este gráfico muestra las enfermedades más asociadas a cada gen basado en la descomposición de los valores singulares de la matriz que contiene el corpus documental de los artículos con el nombre del gen en el abstract. Esta tabla sirve para saber el grado de asociación del gen mutado con respecto al tumor que presenta el paciente.

Interpretación y conclusiones

El estudio genómico del tumor ha mostrado cinco genes mutados que podemos agrupar en tres grupos basado en su similitud funcional: uno estaría formado por los genes NOTCH2 y ARID5B que intervienen en la regulación negativa de la transcripción del promotor de la ARN polimerasa II; otro por DOT1L y HNF1A que actúan como puntos de control del daño del ADN e intervienen en la reparación no homóloga de las roturas de doble cadena, asi como regulan positivamente la transcripción del promotor de ARN polimerasa II; y el último por TP53 que presenta una mayor similitud funcional a los genes NOTCH2 y ARID5B, y además interviene en la reparación de escisión de base y escisión de nucleótidos del ADN.

Los genes prioritarios en este estudio son, como veremos más adelante, NOTCH2 y DOT1L por inducir una transformación epitelio-mesenquimal de las células basales del cáncer de mama, y TP53 como epifenómeno genómico que conlleva una alta aneuploidía celular y resistencia a los tratamientos.

El estudio genómico del tumor muestra una desdiferenciación de las células tumorales y transformación hacía un fenotipo epitelio-mesenquimal. La adquisición de este fenotipo se ha asociado con la resistencia terapéutica, y conduce a una pérdida de adhesión celular, cambios en la polarización de la célula y el citoesqueleto, la migración, la intravasación, la supervivencia en el sistema vascular, la extravasación y la metástasis. Un mecanismo crítico que induce la trasnformación epitelio-mesenquimal es la presencia de la mutación en el gen NOTCH2 mediada por STAT5, JAG-1 y DLL4. Por otra parte, NOTCH2 puede inhibir la angiogénesis y el crecimiento del cáncer de mama mediada por el gen MINAR1 que se expresa ampliamente en diversos tejidos, incluidas las células epiteliales de la mama y las células endoteliales de los vasos sanguíneos. El bloqueo NOTCH2 mediante la combinación de MK-0752 y Tocilizumab inhibe significativamente el crecimiento tumoral y, por lo tanto, podría servir como una estrategia terapéutica novedosa para tratar mujeres con mama que expresa mutaciones en NOTCH2 / NOTCH3 (1).

De igual modo, las mutaciones en el gen DOT1L están involucrado en la regulación de la transición epitelio-mesenquima en el cáncer de mama. DOT1L induce la transformación neoplásica de las células epiteliales de mama, a través de la activación transcripcional dependiente de DOT1L de genes promotores, tales como Snail, ZEB1 y ZEB2. Por lo tanto, DOT1L puede determinar la agresividad del cáncer de mama mediante la activación de factores promotores de la transformación epitelio-mesenquimal (2).

La mutación en el gen ARID5B debe ser interpretado como un epifenómeno secundario a un tumor altamente mutado y que se está desdiferenciando como consecuencia de la adquisición de un fenotipo epitelio-mesenquimal. Este gen tiene la capacidad de regular la transcripción y participan en la desdiferenciación y proliferación celular, por lo que contribuirá a mantener y perpetuar la transición epitelio-mesenquima. De todas las enfermedades malignas, la leucemia es la más estrechamente relacionada con ARID5B. La creciente evidencia muestra que las mutaciones y los polimorfismos de un solo nucleótido de ARID5B se asocian con el desarrollo de leucemia linfoblástica aguda y el resultado del tratamiento, de la leucemia mieloblástica aguda en la infancia, pero no hay estudios que permitan establecer su valor pronóstico y predictivo en el cáncer de mama.

Otra epifenómeno genómico observado en este estudio es la mutación en TP53 que está intimamente relacionado con el subtipo de tumor, llegando al 50% en los carcinomas de tipo basal, y probablemente después de las mutaciones en los genes NOTCH2 y DOT1L. Aunque varios estudios retrospectivos han investigado un potencial pronóstico

y un papel predictivo de la terapia para el TP53 mutante en el cáncer de mama, los resultados hasta la fecha no puede recomendarse como un biomarcador predictivo en el cáncer de mama. En la última década, sin embargo, varios compuestos han llegado a estar disponibles para reactivar la proteína TP53 mutante y convertirla a una conformación con propiedades de tipo salvaje. Se ha encontrado que algunos de estos compuestos, especialmente PRIMA-1, APR-246 PK11007 y COTI-2, exhiben actividad anticancerígena en modelos preclínicos de cáncer de mama (3).

Las mutaciones en TP53 se asocian con tumores altamente aneuploides posiblemente porque estos tumores presentan también conjuntamente sobreexpresión de los activadores transcripcionales mitóticos que aumentan la tasa de cromosomas de anafase rezagados y contribuyen a la mala respuesta a la quimioterapia en el cáncer de mama. Actualmente las drogas que están en fase de desarrollo implican a las vías PI3K/Akt y ERK1/2, potencial de las membrana mitocondrial, lo que provocan la necrosis celular asociada con la producción de especies de oxígeno reactivas y la disminución de la migración celular. Otro posible tratamiento potencial que está disponible como ensayo clínico, y actualmente en fase de reclutamiento, es la utilización de dosis densas de ciclofosfamida disponibles para pacientes con cáncer de mama avanzado y portadores de mutaciones en TP53, basado en que los carcinomas de mama localmente avanzados no inflamatorios con mutaciones enTP53 tenían una alta tasa de respuesta patológica completa a la quimioterapia con dosis densas de doxorrubicina-ciclofosfamida (4).

En conclusión, el perfil genómico del tumor analizado muestra la transformación de las células basales del cáncer de mama hacia un fenotipo epitelio-mesenquimal, lo que se asocia a la progresión, invasión, desarrollo de metástasis y la resistencia al tratamiento convencional. La utilización de dosis densas de ciclofosfamida debería considerarse como una eficaz opción terapéutica, pero debería investigarse otras drogas que se dirigen a la mutación TP53, especialmente PRIMA-1, APR-246 PK11007 y COTI-2. La combinación de MK-0752 y Tocilizumab dirigida frente a NOTCH2 / NOTCH3 podría valorarse como una estrategia terapéutica novedosa.

Referencias:

1. Sugandha Bhatia, James Monkman, Alan Kie Leong Toh, Shivashankar H. Nagaraj. Targeting epithelial–mesenchymal plasticity in cancer: clinical and preclinical advances in therapy and monitoring. Biochemical Journal 2017,474(19): 3269-3306.
2. Lee J-Y, Kong G. DOT1L: a new therapeutic target for aggressive breast cancer. Oncotarget. 2015;6(31):30451-30452.
3. Michael J. Duffy, Naoise C. Synnott, John Crown. Mutant p53 as a target for cancer treatment. European Journal of Cancer, 2017,;83: 258-265.
4. Treatment of Patients With Advanced Breast Cancer Harboring TP53 Mutations With Dose-dense Cyclophosphamide. https://clinicaltrials.gov/ct2/show/NCT02965950

4. Conclusiones

El cáncer provoca una desregulación génica a través de múltiples mecanismos, por lo que son numerosas las variaciones genómicas que presenta cualquier tumor y a distintos niveles en un sistema biológico. El conocimiento de las relaciones entre los genes, así como su asociación con el fenotipo, procesos biológicos y sensibilidad a fármacos a partir de la información disponible en la literatura permite definir la utilidad clínica de un estudio genético y genómico, y hoy día es fundamental para poder implantar la medicina personalizada en la práctica clínica.

La plataforma Enrich Gen permite la integración de la información disponible de múltiples bases de datos, y la visualización de esta información en tablas y gráficos, lo que agiliza la extracción del conocimiento y facilita la interpretación de un estudio genómico.

El futuro de la plataforma pasa por incorporar un módulo para la interpretación de las variantes en el número de copias que son muy frecuentes en el cáncer. Por otra parte, se deberá mejorar el modo de realización del resumen, en la tabla de evidencia científica, con el fin de que pueda mostrar información más relevante para el tipo de tumor considerado. Por último, la plataforma debe generar un informe sobre la información facilitada, con el fin de agilizar la interpretación posterior.

5. Glosario

FDA: Food and Drugs Administration
WES: whole exome sequencing
LSA:Latent semantic analysis
DVS o SVD: Descomposición en valores singulares
GO: Gene Ontology
CI: Contenido de Información
DAG: Directed acyclic graph
HUGO: Comité de Nomenclatura de Genes
UMLS: Unified Medical Language System
NCBI: National Center for Biotechnology Information
BMA: Best Match Average
PMID:PubMed Identification

6. Bibliografía

1. Davies K. The $1,000,000 genome interpretation. Bio-IT World. 2010;8:50–54.

2. Biesecker L.G. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: Lessons from the ClinSeq project. Genet. Med. 2012;14:393–398. doi: 10.1038/gim.2011.78.

3. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012;13(12):829–39.

4. Dai HJ, Chang YC, Tsai RT, Hsu WL. New challenges for biological text-mining in the next decade. J Comput Sci Technol. 2010;25(1):169–79.

5. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011:

6. Gene Ontology Consortium; The Gene Ontology (GO) database and informatics resource, Nucleic Acids Research, Volume 32, Issue suppl_1, 1 2004; D258–D261,

7. Spampinato C, Kavasidis I, Aldinucci M, Pino C, Giordano D, Faro A. Discovering biological knowledge by integrating high-throughput data and scientific literature on the cloud. Pract Exp. 2013;26(10):1771–86.

8. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. Bioinformatics. 2010;26(22):2924–6.

9. Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Sci Transl Med 2011; 3: 114ra27.

10. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000; 24: 236–244.

11. Ellwanger DC, Leonhardt JF, Mewes HW. Large-scale modeling of condition-specific gene regulatory networks by information integration and inference. Nucleic Acids Res 2014; 42: 0.
12. Bethany Percha, Russ B Altman; A global network of biomedical relationships derived from text, Bioinformatics, , bty114, https://doi.org/10.1093/bioinformatics/bty114

13. M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald,

G.M. Rubin, G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25 (2000), pp. 25-29

14. D. Barrell, E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, R. Apweiler. The GOA database in 2009 — an integrated Gene Ontology Annotation resource Nucleic Acids Res., 37 (2009), pp. D396-D403

15. F.M. Couto, M.J. Silva, P.M. Coutinho. Measuring semantic similarity between Gene Ontology terms Data Knowl. Eng., 61 (2007), pp. 137-152

16. P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19 (2003), pp. 1275-1283

17. P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. Proceedings of the Pacific Symposium on Biocomputing (2003), pp. 601-612

18. T. Xu, L. Du, Y. Zhou. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics, 9 (2008), p. 472

19. Lee SG, et al. A graph-theoretic modeling on GO space for biological interpretation of gene clusters, Bioinformatics , 20 (2004), pp. 381-388

20. P.H. Lee, D. Lee. Modularized learning of genetic interaction networks from biological annotations and mARN expression data. Bioinformatics, 21 (2005), pp. 2739-2747

21. S.-L. Cao, L. Qin, W.-Z. He, Y. Zhong, Y.-Y. Zhu, Y.-X. Li. Semantic search among heterogeneous biological databases based on gene ontology. Acta Biochim. Biophys. Sin., 36 (2004), pp. 365-370

22. X. Guo, R. Liu, C.D. Shriver, H. Hu, M.N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics, 22 (2006), pp. 967-973

23. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. Cmp-lg/9511007 (1995)

7. Anexos

Anexo 1. Código fuente de la plataforma Enrich Gen