UNIVERSIDAD PABLO DE OLAVIDE

TESIS DOCTORAL

Validación de modelos genéticos en bioinformática: implementación y visualización

Autor:
Juan José DÍAZ MONTAÑA

Director:
Dr. Norberto DÍAZ DÍAZ

Tesis presentada en cumplimiento de los requisitos para optar al grado de Doctor en Biotecnología, Ingeniería y Tecnología Química

dentro de

DATA^i: Intelligent Data Analysis Group (TIC-239)
Escuela de Doctorado de la Universidad Pablo de Olavide (EDUPO)

3 de febrero de 2022

A mis padres.
Por hacerme lo que soy y empujarme siempre a ser mejor.

Agradecimientos

Este trabajo ha sido largo y costoso. Mucho más de lo que debería. Combinar la investigación con otros trabajos no es fácil y muchas situaciones personales se han interpuesto en el camino a lo largo de los años. Sería mentira decir que nunca pensé en dejarlo porque si que lo hice. Y no una, sino muchas veces. Por ello, he de agradecer a todas esas personas que de una manera u otra me han empujado a seguir y han contribuido a alisarme el camino.

Gracias a todos los investigadores que de una forma u otra han contribuido al desarrollo de esta tesis. A Owen y Enrico por acogerme durante mi estancia en Inglaterra. A Fran, Carlos y a Nacho por trabajar conmigo y ayudarme a avanzar. A todos aquellos que de una forma u otra me ayudaron a avanzar. En especial, a Norberto, por engañarme para meterme en una tesis "corta y sencilla". No lo ha sido, pero no me arrepiento de nada. Gracias por saber entender las situaciones por las que he ido pasando, darme mi tiempo cuando me ha hecho falta y empujarme a seguir cuando era el momento.

Gracias a todos los amigos de España, Inglaterra, Finlandia y el resto del mundo que me han apoyado de una forma u otra durante esta aventura. A Francesco, por brindarme su apoyo durante mi tiempo en Londres y ayudarme a realizar mi primera estancia. A Pedro y Alejandro, con los que he compartido incontables saunas en Tampere discutiendo las penas y alegrías de la investigación académica junto con otros muchos temas. Hacer un doctorado mayormente en remoto es muy solitario y vosotros habéis sido mi grupo de apoyo local.

Gracias a Chelsea, que ha compartido conmigo la mayoría de las penas y alegrías que este trabajo me ha traído. Sin tú apoyo, esta tesis no hubiera sido posible.

Gracias a mi hermano Enrique, por contagiarme su pasión por la investigación y a mi hermano Antonio, capaz de aliviar hasta los momentos más duros y hacerte ver lo que de verdad importa. Y por encima de todo, gracias a mis padres, Juan José y María Teresa, por regalarme mi presente y mi futuro. Sin vosotros no sería lo que soy. Vuestro apoyo incondicional ha sido fundamental para darme fuerzas en los momentos bajos y ayudarme a continuar. Esta tesis es para vosotros.

UNIVERSIDAD PABLO DE OLAVIDE

Resumen

DATA^{i}: Intelligent Data Analysis Group (TIC-239)
Escuela de Doctorado de la Universidad Pablo de Olavide (EDUPO)

Doctor en Biotecnología, Ingeniería y Tecnología Química

Validación de modelos genéticos en bioinformática: implementación y visualización

by Juan José DÍAZ MONTAÑA

Since the human genome was completely sequenced for the first time, the great scientific and technological advances in the biotechnology industry have greatly reduced the cost of experiments while significantly improving results. This has led to an exponential growth in the biological information available and, due to this huge amount of information, researchers are faced with mountains of data with only flakes of knowledge [1].

Approaches as Knowledge Database Discovery (KDD) [2] are used to generate models that allows researcher to gain knowledge about complex biological systems.

Gene networks arose as a straightforward way of representing gene sets including their interactions. They are presented as a network structure where each node represents a gene or gene product (protein) while each edge denotes the relationship between the nodes at its ends. The concrete nature of each relationship and the meaning of its weight depend on the network architecture and the inference algorithm used. A gene network is an abstraction that facilitates the study of its underlying biological system. They are easy to visualize, and they are informative on their own.

Gene networks have been successfully used in clinical diagnosis [3] and a large number of inferred interactions have been confirmed experimentally, thus confirming their reliability [4]. The inference of gene networks has also allowed a better understanding of fundamental processes that occur in living organisms such as development or nutrition and metabolic coordination [5].

Research has focused on inferring these networks using different experimental and computational techniques [6], as well as analyzing those networks to extract knowledge [7]. Also, a significant number of methods have been developed to validate the inferred networks in order to verify their quality and reliability [8], [9].

All the methodologies of gene network inference, analysis, and validation are based on algorithms and computer tools. Given the increasing importance and popularity of these computational approaches, it becomes increasingly critical to ensure that the software is usable and accessible, as these features provide the basis for the reproducibility of published biomedical research [10].

Based on the existing need for automatic techniques of inference, analysis and validation of models for the study of interactions between genes and the deficiencies in existing techniques, this work presents different novel approaches for the inference, analysis and validation of genetic models, especially gene networks, with a special emphasis on the usability and accessibility of the proposed solutions.

Índice general

Índice de figuras

List of Abbreviations

GRN	**G**ene **R**egulatory **N**etwork
ADN	**Á**cido **D**esoxirribo**N**ucleico
ARN	**Á**cido **R**ibo**N**ucleico
ARNm	**Á**cido **R**ibo**N**ucleico **M**ensajero
ARNt	**Á**cido **R**ibo**N**ucleico de **T**ransferencia
ARNnc	**Á**cido **R**ibo**N**ucleico **N**o **C**odificante
SNP	**S**ingle **N**ucleotide **P**olymorphism (polimorfismo de un núcleo simple)
TF	**T**ranscrption **F**actor (factor de transcripción)
EST	**E**xpressed **S**equence **T**ag (secuenciación de marcador de secuencia expresada)
SAGE	**S**Serial **A**nalysis of **G**ene **E**xpression (análisis en serie de la expresión génica)
MPSS	**M**assively **P**arallel **S**ignature **S**equencing (secuencia paralela masiva de la firma)
GO	**G**ene **O**ntology
IC	**I**nformation **C**ontent (contenido de información)
MICA	**M**ost **I**nformative **C**ommon **A**ncestor (ancestro común más informativo)
LCA	**L**owest **C**ommon **A**ncestor (ancestro común más bajo)
DCA	**D**isjoint **C**ommon **A**ncestors (ancestros comúnes disjuntos)
KDD	**K**nowledge **D**iscovery in **D**atabases
IDA	**I**ntelligent **D**ata **A**nalysis
NGS	**N**ext **G**eneration **S**equencing
GGMs	**G**raphical **G**aussian **M**odels (modelos gráficos gaussianos)
BN	**B**oolean **N**etwork (red booleana)
PBNs	**P**robabilistic **B**oolean **N**etwork (red booleana probabilística)
DBN	**D**ynamic **B**ayesian **N**etwork (red bayesiana dinámica)
DAG	**d**irected **A**cyclic **G**raph (grafo acíclico no dirigido)
ODE	**O**rdinary **D**ifferential **E**quation (ecuación diferencial ordinaria)
ANN	**R**ed **N**euronal **A**rtificial
FNN	**R**ed **N**euronal **F**eedforward
RNN	**R**ed **N**euronal **R**ecurrentes
CNN	**R**ed **N**euronal con **C**onvolución
GNN	**R**ed **N**euronal de **G**rafos
GCNN	**R**ed **N**euronal de **G**rafos con **C**onvolución
AR	**R**egla de **A**sociación
WGPA	**W**eb-based **G**ene **P**athogenicity **A**nalysis
GFD-Net	**G**O-based **F**unctional **D**issimilarity of **Net**works
GNC	**G**ene **N**etwork **C**oherence
GRNCOP2	**G**ene **R**egulatory **N**etwork inference by **C**ombinatorial **OP**timization 2

Parte I

Introducción

Capítulo 1

Introducción

1.1. Planteamiento

Desde que el genoma humano fuera completamente secuenciado por primera vez y gracias a los grandes avances tanto científicos como tecnológicos en la industria biotecnológica, se ha producido durante los últimos años un crecimiento exponencial de la información biológica disponible. Dichos avances han permitido disminuir enormemente el coste de los experimentos, mejorando al mismo tiempo los resultados de manera significativa.

Ante esta avalancha de información, los investigadores se enfrentan a montañas de datos, en las que sólo una pequeña parte permite generar conocimiento [1]. En el área de análisis de expresión génica, se han desarrollado multitud de técnicas computacionales para extraer patrones y conocimiento a partir de los niveles de expresión de miles de genes tomados simultáneamente [11]-[14].

Estas técnicas se basan en el conocido flujo de trabajo Knowledge Database Discovery (KDD). El cual va desde el preprocesamiento de los datos de entrada hasta la validación de los modelos generados, a menudo realizada mediante la búsqueda de información en bases de datos y la comparación con datos experimentales anteriores.

El proceso comienza con los datos de entrada, que suelen ser conjuntos de datos de expresión génica, que pueden obtenerse ya sea de forma experimental o de bases de datos existentes como NCBI GEO [15] (Paso 1 de la Figura 2.3). Una vez seleccionado el conjunto de datos de entrada, éste se puede preprocesar mediante algún método computacional para mejorar la calidad del estudio (Paso 2 de la Figura 2.3). Estos datos preprocesados se utilizan como entrada para un algoritmo de inferencia computacional (aprendizaje automático), que infiere un modelo como resultado (Paso 3 de la Figura 2.3). Finalmente, el modelo obtenido se optimiza y valida de forma que se puedan obtener verdaderos conocimientos biológicos a partir de él, mediante una comparación con el conocimiento biológico real existente (Pasos 4 y 5 en la Figura 2.3).

Uno de los modelos que más popularidad ha adquirido en los últimos años son las redes genéticas. Este tipo de modelo simplifica la representación de procesos biológicos complejos representándolos mediante una estructura en forma de red. En dicha red, cada nodo representa elemento biológico (gen, proteína, metabolito o ARN) y cada arista denota la relacione entre los nodos a sus extremos. La naturaleza exacta de estas relaciones depende del método de inferencia utilizado y del modelo concreto que se genere. Las aristas pueden codificar detalles sobre la relación como, por ejemplo, su relevancia ponderando cada arista con un valor numérico o peso. Las redes genéticas representan una abstracción que facilita el estudio del sistema biológico subyacente, son fáciles de visualizar y son informativas por si solas.

Las redes genéticas se han utilizado con éxito en el diagnóstico clínico [3] y un gran número de las interacciones inferidas se han confirmado experimentalmente, confirmando así su fiabilidad [4]. La inferencia de redes genéticas también ha permitido un mejor entendimiento de procesos fundamentales que ocurren en los organismos vivos como son el desarrollo o la nutrición y coordinación metabólica [5].

Durante los últimos años, se han desarrollado multitud de enfoques para la inferencia de redes genéticas siguiendo el modelo de KDD visto en anteriormente y produciendo diferentes arquitecturas y modelos [6]. Estás técnicas pueden variar desde modelos sencillos basados en correlación [16], regresión [17] o teoría de la información [18], computacionalmente sencillos pero limitados, hasta modelos probabilísticos más complejos como son los basados en redes bayesianas [19] o ecuaciones diferenciales ordinarias [20], [21], capaces de representar más realísticamente la complejidad inherente a los sistemas biológicos. Teniendo en cuenta que los sistemas biológicos son complejos, que los datos pueden venir de escenarios heterogéneos y que diferentes condiciones biológicas pueden llevar a que se activen diferentes estados, esto es una suposición muy restrictiva asumir que todos los datos se pueden modelar en una sola red genética. Por ello, se han desarrollado también diversas técnicas que permiten inferir múltiples estructuras de red a partir de un solo conjunto de datos [22], [23]. Además, debido precisamente a que la mayoría de las redes genéticas son difíciles de mapear con precisión mediante un modelo matemático, los enfoques sin modelo (model-free) [24], los cuales ofrecen una forma de identificar los mecanismos regulatorios directamente a partir de los datos de entrada/salida sin ningún modelo subyacente, han ganado popularidad. Concretamente, los enfoques basados en reglas de asociación son técnicas sin modelo altamente abstractas que requieren una menor cantidad de datos que otras técnicas, son computacionalmente simples y presentan una capacidad de inferencia importante.

Una vez inferidas, las redes genéticas deben ser validadas con el fin de verificar su calidad y fiabilidad. Dados la complejidad y el alto costo de validar las redes experimentalmente [25], existen técnicas de validación internas basadas en datos estadísticos de la propia red inferida [26] o en la topología de ésta [27], [28]. Sin embargo, este tipo de validaciones no considera la relevancia biológica de la red. Por ello, es común la validación mediante el uso de conocimiento previo [8], [9]; por ejemplo una comparación directa entre la red inferida utilizando repositorios de interacción entre genes [29] como gold-standard. Este método de validación es simple, sin embargo, tiene dos limitaciones importantes: el gold-standard utilizado debe ser una red o un conjunto de interacciones entre genes y debe ser totalmente fiable y adecuado para la evaluación la red bajo estudio. Debido a ello, otro enfoque para validar redes genéticas basado en el conocimiento biológico existente consiste en el uso de anotaciones sobre las entidades biológicas, concretamente sobre los genes y los productos génicos (proteínas), las cuales se almacenan a menudo de forma controlada en forma de terminologías u ontologías [30], [31]. Siguiendo este segundo enfoque, existen principalmente dos tipos de análisis para explotar las anotaciones de genes: los análisis de enriquecimiento [32], [33] y los de similitud semántica [34]. Los análisis de enriquecimiento proporcionan una clasificación de los genes conocidos y puntúa la evidencia de sus asociaciones con una lista objetivo de genes de interés. Aunque este enfoque ha sido ampliamente utilizado para analizar la relevancia de conjuntos de genes [35], [36], sólo proporciona información sobre la distribución de las anotaciones y no da una medida cuantitativa. Las medidas de similitud semántica basadas en ontologías [37], [38], por contra, proporcionan un valor numérico que representa la coherencia funcional de un conjunto de entrada. Este tipo de metodología se basa en la comparación de los genes o productos génicos, según su similitud en función

de sus características funcionales expresadas por los términos de ontología. El principal inconveniente de este tipo de análisis es que no tienen en cuenta las relaciones entre los genes representados en la red, lo cual es la característica principal de las redes genéticas que las diferencian de simples conjuntos de genes. Por dicho motivo, los enfoques de análisis existentes no son totalmente exactos para la evaluación de redes de genes.

Todas estas metodologías de inferencia, análisis y validación de redes genéticas se basan en algoritmos y herramientas informáticas. Teniendo en cuenta la creciente importancia y popularidad de estos enfoques computacionales, se hace cada vez más crítico garantizar que el software sea usable y accesible, ya que estas características proporcionan la base para la reproducibilidad de la investigación biomédica publicada [10].

A partir de la existente necesidad de técnicas automáticas de inferencia, análisis y validación de modelos para el estudio de interacciones entre genes y las carencias en las técnicas existentes, en este trabajo se presentan diferentes aproximaciones novedosas para la inferencia, análisis y validación de modelos genéticos, especialmente redes genéticas, con un especial énfasis en la usabilidad y accesibilidad de las soluciones propuestas.

1.2. Objetivos y publicaciones

1.2.1. Objetivos

Este trabajo se centra en la creación de herramientas para la generación, validación y análisis de modelos genéticos, especialmente redes genéticas, que ofrezcan una alta accesibilidad y usabilidad. Para ello se destacan los siguientes sub-objetivos:

1. Estudiar y analizar las técnicas de inteligencia artificial existentes para la generación y validación automática de modelos a partir de datos genómicos, prestando especial atención a las aproximaciones de redes genética.

2. Estudiar los problemas de usabilidad existentes en las herramientas usadas por la comunidad investigadora en el campo de la bioinformática.

3. Crear aproximaciones integradas en las herramientas usadas por la comunidad investigadora combinando metodologías de vanguardia existentes o nuevas de forma que ofrezcan gran accesibilidad o usabilidad.

4. Proponer una metodología de inferencia y/o evaluación de modelos genéticas basadas en las características diferenciadoras de las redes genéticas, como son sus interacciones, que respondan a las necesidades actuales.

5. Demostrar la utilidad de las herramientas para el estudio de enfermedades en humanos. La bioinformática no es una ciencia puramente teórica, sino que tiene aplicaciones directas en otras ramas como la medicina o la farmacología. En este trabajo, nos centraremos en utilizar las medidas desarrolladas para el estudio de enfermedades en seres humanos.

1.2.2. Publicaciones

Como resultado de la persecución de dichos objetivos, se han realizado 4 publicaciones relevantes:

- Diaz-Montana, J.J., Rackham, O.J., Diaz-Diaz, N. and Petretto, E., 2016. Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data. Bioinformatics, 32(4), pp.635-637. IF: 7.31, Q1. [39]

- Díaz-Montaña, J.J., Díaz-Díaz, N. and Gómez-Vela, F., 2017. Gfd-net: A novel semantic similarity methodology for the analysis of gene networks. Journal of biomedical informatics, 68, pp.71-82. 71-82. IF: 2.88, Q2. [40]

- Díaz-Montaña, J.J., Gómez-Vela, F. and Díaz-Díaz, N., 2018. GNC–app: A new Cytoscape app to rate gene networks biological coherence using gene–gene indirect relationships. Biosystems, 166, pp.61-65. IF: 1.62, Q3. [41]

- Díaz–Montaña, J.J., Díaz–Díaz, N., Barranco, C.D. and Ponzoni, I., 2019. Development and use of a Cytoscape app for GRNCOP2. Computer methods and programs in biomedicine, 177, pp.211-218. IF: 3.63, Q1. [42]

1.2.3. Relación entre los objetivos y las publicaciones

Los dos primeros objetivos son un paso previo y necesario a cualquier trabajo científico realizado, y es evidenciado en las introducciones desarrolladas en los artículos, así como en la justificación de la propia propuesta. En este sentido, en primer lugar, se han abordado las técnicas de validación existentes y cómo estas hacen uso de información externa. A partir de este análisis se desarrolló la primera de las propuestas [39]. Seguidamente, se analizaron las técnicas de evaluación de redes genéticas, en donde la interacción entre genes juega un papel muy relevante al ser el elemento especialmente diferenciador de este modelo. En este sentido se destacan la segunda y tercera propuesta [40], [41]. Por último, y atendiendo a las competencias alcanzadas en los estudios anteriores, se amplia el análisis a las técnicas de inferencia, ya que se domina cómo estas pueden ser evaluadas. Sobre este análisis de desarrolla la cuarta propuesta [42].

Con respecto al tercer objetivo, las cuatro publicaciones previamente destacadas solventan gran parte de los problemas de accesibilidad y usabilidad conocidos [43], [44]. En la primera propuesta [39] se presenta una aplicación web que combina múltiples metodologías implementadas en múltiples tecnologías. WGPA simplifica el uso de metodologías que de otra manera requerirían conocimientos relativamente avanzados de informática para poder realizar análisis similares. Al ser una accesible como aplicación web, WGPA es accesible a toda la comunidad investigadora y no requiere de ningún software o hardware especial por parte del usuario. Por contra, las visualizaciones se implementaron ad-hoc e introducir visualizaciones complejas, por ejemplo, de grandes redes, hubiera sido complicado. Además, al no exponer ninguna interfaz (API), resulta difícil integrarse en flujos de trabajo más complejos que incluyan otras herramientas. Por ello, las otras tres propuestas [40]-[42] se implementaron como apps de Cytoscape [45], [46]. Cytoscape es una de las aplicaciones más usadas por la comunicad investigadora bioinformática para el estudio de redes genéticas. Parte de su éxito es la potente plataforma que ofrece para la visualización redes complejas, la gestión de sus metadatos y la integración de distintas fuentes de datos así como de herramientas de análisis populares. Además, Cytoscape ofrece un gran ecosistema de apps, las cuales extienden sus funcionalidades básicas permitiendo la realización de análisis y visualizaciones más complejas. Como ejemplo, la integración de la cuarta propuesta en Cytoscape fue clave para poder combinar la metodología original con otras aplicaciones del ecosistema y realizar un análisis

topológico de los resultados. Además, siguiendo las recomendaciones para la mejora de la archivabilidad de las herramientas bioinformáticas presentadas por varios autores [47], [48], tanto el código fuente como la documentación relacionados con las cuatro publicaciones han sido publicados en GitHub bajo licencias de permisivas.

El cuarto objetivo se alcanzó en la segunda propuesta [40], una metodología completamente novedosa que aplica el concepto de similitud semántica para validar y analizar redes genéticas, mejorando las medidas similares existentes.

Finalmente, el quinto objetivo también ha sido cubierto ampliamente. La primera propuesta [39] se validó mediante la identificación de mutaciones de novo en encefalopatías epilépticas, la segunda [40] mediante la identificación in-silico de tres enfermedades humanas a partir de redes de genes relevantes: diabetes mellitus tipo II, resistencia a la insulina y enfermedad de Alzheimer, y la cuarta [42] mediante el análisis de la intercomunicación entre las rutas metabólicas de la enfermedad de Alzheimer y la proteólisis mediada por la ubiquitina.

1.2.4. Otras publicaciones

Aparte de las cuatro publicaciones principales ya mencionadas, se han realizado otras tres contribuciones en la misma linea:

- Díaz-Montaña, J.J., 2014. Cytoscape: guía de iniciación al desarrollo. MoleQla: revista de Ciencias de la Universidad Pablo de Olavide, (14), pp.8-5. [49]
- Diaz-Montana, J.J. and Diaz-Diaz, N., 2014. Development and use of the Cytoscape app GFD-Net for measuring semantic dissimilarity of gene networks. F1000Research, 3. [50]
- Diaz-Montana, J.J. and Diaz-Diaz, N., 2015, October. GFD-Net: a novel apporach for analyzing the functional dissimilarity of gene networks. In 6th Argentinian Conference on Bioinformatics and Computational Biology. A2B2C. [51]

Parte II

Marco teórico

Capítulo 2

Bioinformática

2.1. Introducción

En este capítulo se desarrollan los conceptos básicos de biología y bioinformática necesarios para poder comprender el resto del documento. Primero, se describe el Dogma Central de la Biología Molecular, el cual ilustra los mecanismos de transmisión y expresión de la herencia genética, definiendo los fundamentos de la biología actual y sentando las bases de la bioinformática. A continuación, se incluye una breve introducción a la bioinformática, su historia, las distintas áreas de estudio existentes, los problemas que intenta resolver y las soluciones existentes. Para terminar, se detallan en mayor profundidad algunas de las herramientas existentes, las cuales son especialmente relevante durante el resto de este documento.

2.2. Dogma Central de la Biología Molecular

El Dogma Central de la Biología Molecular, es una explicación del flujo de la información genética en sistemas biológicos propuesta por Francis Crick en 1958 [52].

El desarrollo y funcionamiento de todos los organismos vivos e incluso algunos virus está determinado por el ácido desoxirribonucleico (ADN), un ácido nucleico que almacena a largo plazo las instrucciones genéticas para construir otros componentes de las células, como son las proteínas y las moléculas de ácido ribonucleico (ARN). Además, el ADN es el responsable de la transmisión hereditaria de ciertas características. Las secuencias de ADN están formadas por genes, que son segmentos que contienen la información genética concreta para crear una molécula. En un gen, la secuencia de nucleótidos a lo largo de una hebra de ADN se transcribe a un ARN mensajero (ARNm), el cual a su vez permite al organismo sintetizar o "expresar" una determinada proteína en uno o varios momentos de su vida (ver figura 2.1). Hasta hace poco tiempo se pensaba que el ADN no codificante, es decir, las secuencias de ADN que no codifican ninguna proteína, no tenía utilidad alguna. Sin embargo, estudios recientes indican que eso es inexacto ya que pueden tener propósitos estructurales o reguladores de la expresión génica [53].

En todos los organismos, el ADN contenido en cada una de sus células es esencialmente idéntico. Esto quiere decir que toda la información necesaria para la síntesis de todas las proteínas está contenida en todas las células. Sin embargo, no todos los genes se expresan al mismo tiempo ni en todas las células y esta diferencia en los niveles de expresión es lo que determina la funcionalidad de una célula en un determinado momento. Esta regulación de los niveles de expresión se conoce como regulación génica y depende de factores tanto internos como externos a la célula.

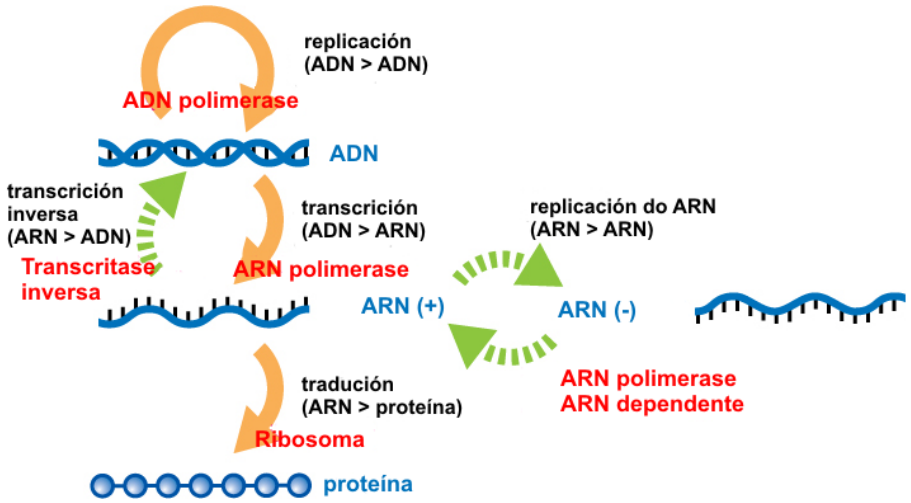


FIGURA 2.1: Dogma Central de la Biología Molecular

De esta forma, una célula puede no solo desarrollarse normalmente sino también responder a su entorno.

El primer paso de la expresión génica es la transcripción. Durante la transcripción genética, gracias a una enzima llamada ARN polimerasa, las secuencias de ADN se transforman en un ARNm que mantiene la información de la secuencia de ADN. La unidad codificadora del código genético es un grupo de tres nucleótidos (triplete o codón), representado por las tres letras iniciales de las bases nitrogenadas (por ejemplo, ACT, CAG o TTT). Cada codón codifica un aminoácido o un símbolo de puntuación (Comienzo, parada) y, durante el proceso de transcripción, estos tripletes provenientes del ADN se transcriben en sus bases complementarias en el ARNm (para el ejemplo anterior UGA, GUC o AAA, respectivamente).

Una vez completado el ARNm, éste es traducido por un ribosoma. Durante la traducción, el ARNm se decodifica para generar una cadena específica de aminoácidos, llamada polipéptido, la cual permite la formación de una proteína de acuerdo con las reglas especificadas por el código genético. El ribosoma lee los codones del ARNm e interacciona con una molécula de ARN de transferencia (ARNt) que contenga el triplete complementario, denominado anticodón. Cada ARNt porta el aminoácido correspondiente al codón de acuerdo con el código genético, el cual es agregado a la cadena polipeptídica que se está sintetizando. A medida que los aminoácidos se van uniendo a la cadena peptídica en crecimiento, la cadena comienza a plegarse de una forma determinada. Generalmente, la traducción comienza en un codón AUG (adenina-uracilo-guanina) o un iniciador de metionina corriente por debajo del sitio de unión con el ribosoma y termina con un codón de terminación que puede ser un triplete UAA, UGA o UAG.

Finalmente, el ARNm no contiene toda la información para especificar la naturaleza final de la proteína, sino que la cadena polipeptídica creada por el ribosoma suele requerir de un procesamiento adicional antes de ser completamente funcional. El plegado de una proteína en su forma final juega un papel crucial en la regulación de la actividad biológica de la proteína, así como en la localización de esta en las diferentes ubicaciones celulares [54]. Dicho proceso es complejo y suele requerir de otras proteínas chaperonas para controlar la forma final de la proteína [55]. Además, algunas proteínas extirpan segmentos internos de sus propias cadenas peptídicas, otras deben dividirse en múltiples secciones sin empalmes, otras deben estar reticuladas, y otras deben estar unidas a cofactores como el hemo [56].

Por otro lado, para la transferencia de la información genética de una generación a la siguiente, es necesario crear copias idénticas del ADN existente. Este proceso se conoce como replicación y consiste esencialmente en la separación de las dos hebras de la doble hélice del ADN, las cuales sirven de molde para la posterior síntesis de cadenas complementarias a cada una de ellas en forma de ARNm.

El dogma central de la biología molecular establecía que el flujo de actividad y de información era unidireccional: $ADN \rightarrow ARN \rightarrow$ Proteína. No obstante, en la actualidad ha quedado demostrado que algunos virus son capaces de realizar una "transcripción inversa", también llamada "retrotranscripción" [57]. Es decir, pueden producir ADN a partir de ARNm. Además, se sabe que existen secuencias de ADN que se transcriben a ARN sin llegar a traducirse nunca a una proteína como es el caso de los ARN interferentes [58].

También es conocido que muchos virus contienen una enzima llamada ARN polimerasa dependiente del ARN que permite replicar el ARN a un nuevo ARN. Un proceso similar ocurre en muchas células eucariotas donde están involucrados en el silenciamiento del ARN [59].

2.3. Bioinformática

La bioinformática, como su propio nombre indica, es un campo científico interdisciplinario que combina las ciencias de la vida y las ciencias de la información [60], [61].

El NIH (National Institute of Health, Institutos Nacionales de la Salud de los Estados Unidos), define la bioinformática como "la investigación, desarrollo o aplicación de herramientas computacionales y aproximaciones para la expansión del uso de datos biológicos, médicos, conductuales o de salud, incluyendo aquellas herramientas que sirvan para adquirir, almacenar, organizar, analizar o visualizar tales datos" [62].

2.3.1. Historia

El uso de la computación en estudios biológicos se hizo popular en los años 50 con el inicio de la secuenciación de proteínas [63]. Durante los años 60 y 70, se produjeron grandes avances en el alineamiento de secuencias [64] y comenzó la secuenciación de ADN y el desarrollo de software para analizarlo [65], [66]. En 1978 Sanger et al. publicaron la primera secuencia de genes completa de un organismo, el fago $\phi - X174$ (5.386 pares de bases que codifican 9 proteínas) [67]. Durante los años 80, se produjeron grandes avances tanto experimentales como computacionales y surgieron importantes bases de datos biológicas (GenBank en 1982 [68], Swiss-Prot en 1986 [69]) y redes para interconectarlas (EMBnet en 1988 [70]). Además, se potenciaron o crearon diferentes organismos e instituciones (EMBL se constituye en 1974 pero se desarrolla durante la década de los 80 [71], NCBI en 1988 [72]). En los años 90, con el advenimiento de internet [73] y el Proyecto Genoma Humano [74] se produjo una explosión en la cantidad de datos disponibles, en la capacidad de computación disponible y, por lo tanto, en el desarrollo de la bioinformática.

2.3.2. Áreas de estudio

La bioinformática suele utilizarse como término general para cualquier estudio biológico que utiliza la computación como parte de su metodología y engloba multitud de áreas de estudio [60].

Extracción y análisis de secuencias

Desde que el Fago $\phi - X174$ fuera secuenciado en 1977 [67], las secuencias de ADN de miles de organismos han sido decodificadas y almacenadas en diferentes bases de datos. Un punto de inflexión determinante fue el Proyecto Genoma Humano, un proyecto internacional que permitió determinar la secuencia de pares de bases químicas que componen el ADN e identificar y mapear el conjunto completo de genes de un genoma humano promedio desde un punto de vista físico y funcional, incluyendo tanto los genes que codificantes como los no codificantes.

Anotación del genoma

En el contexto de la genómica, la anotación es el proceso de marcado de los genes (lugares en la secuencia de ADN que codifican una proteína), el ARNt, y otras funciones biológicas de la secuencia de ADN, así como de la atribución de funciones a dichos elementos [75].

Genómica comparativa

El análisis de alineamiento de secuencias se basa en la biología evolutiva y en el hecho de que cuanto más similares sean dos secuencias más similares tenderán a ser las funciones de las proteínas codificadas por ellas, especialmente para especies filogenéticamente cercanas. Normalmente dos secuencias tienen una alta similitud porque son homólogas, es decir, comparten un ancestro común. A diferencia de la similitud, la homología no es un término cuantitativo, dos secuencias o son homólogas, derivan del mismo ancestro, o no lo son. Sin embargo, la similitud de dos secuencias es un valor cuantitativo que hace posible la inferencia de homología entre dos proteínas [76].

Análisis de estructura de proteínas

La secuencia de aminoácidos que forma una proteína determina como ésta se pliega y adquiere una estructura tridimensional específica, conocida como estructura primaria. Además, la estructura final de una proteína tiene un carácter jerarquizado, es decir, implica unos niveles de complejidad creciente que dan lugar a 4 tipos de estructuras (primaria, secundaria, terciaria y cuaternaria) construyéndose cada una a partir de la anterior. La estructura tridimensional de una proteína es un factor determinante en su actividad biológica. De ahí que un objetivo fundamental del estudio de las proteínas, es entender la relación entre la estructura de una proteína y su función, lo que a su vez permite predecir estructuras y funciones a partir de secuencias de aminoácidos [77]-[79].

Análisis de datos de expresión génica

Como se ha visto en la sección 2.2, el ADN contenido en cada célula de un organismo es esencialmente idéntico, y es la diferencia en los niveles de expresión de los distintos genes en un determinado momento lo que determina la funcionalidad de una célula en dicho instante. Estos niveles de expresión génica pueden determinarse mediante la medición de los niveles de ARNm. Para ello, existen multitud de técnicas entre las que cabe destacar los microarrays de ADN, la secuenciación de marcador de secuencia expresada (EST, Expressed Sequence Tag), los análisis en serie de

la expresión génica (Serial Analysis of Gene Expression - SAGE) o secuencia paralela masiva de la firma (MPSS, Massively Parallel Signature Sequencing) [11]-[14]. Todas estas técnicas producen un gran volumen de datos, pero son extremadamente propensas al ruido y están sujetas a sesgos en la medición biológica. Por ello, es necesario el desarrollo de herramientas para separar los datos válidos del ruido [80]. También se realizan análisis de la expresión de proteínas utilizando microarrays o espectrometría de masas con similares problemas a los ya mencionados [81], [82]. Los estudios de expresión génica se usan a menudo para determinar como los genes se regulan los unos a los otros, así como para detectar genes implicados en enfermedades, mediante la comparación de los niveles de expresión entre una persona sana y una enferma.

Genética de las enfermedades

Gracias a las grandes cantidades de genoma secuenciado y la información sobre cómo los genes se expresan en distintos individuos, es posible realizar estudios comparando individuos sanos con enfermos para detectar cambios genéticos que se puedan asociar con cierto fenotipo [83]. Por ejemplo, la variación de un sólo par de bases (SNP) o una deleción, repetición, etc. en una secuencia del genoma que siempre resulte en el mismo fenotipo. Ese tipo de estudios son especialmente relevantes el campo de la oncogenómica, el cual se centra en las alteraciones genómicas, epigenómicas y de transcripción en el cáncer [84].

Análisis de regulación génica

Hoy en día, sabemos que los genes no se expresan de forma aislada. La regulación génica es la compleja orquestación que resulta de la red de interacciones formadas entre los genes de un conjunto, así como de señales extracelulares y define qué genes se activan o reprimen produciendo el incremento o decremento en la actividad de una o más proteínas [85].

Biología de sistemas

Aunque los análisis de regulación génica representan un importante paso adelante para comprender los complejos sistemas biológicos, hoy en día sabemos que los niveles de expresión génica en una célula están afectados por mucho más que los propios genes. La biología de sistemas representa un enfoque holístico que tiene como objetivo comprender de forma íntegra el funcionamiento de los sistemas (procesos) biológicos y profundizar en el entendimiento de cómo sus interacciones tanto internas como con otros sistemas conllevan la aparición (emergencia) de nuevas propiedades. Para ello, la biología de sistemas se basa en la modelización matemática de los procesos bajo estudio [86].

Otros

Existen muchas otras áreas de estudio dentro de la bioinformática. Solo por nombrar algunas podemos destacar los análisis de la organización celular, análisis automáticos de la literatura existente, análisis automáticos de imágenes y vídeos, estudios de biodiversidad y biología evolutiva, u ontologías e integración de datos de múltiples fuentes.

2.3.3. Objetivos

Tras el Proyecto Genoma Humano, la necesidad de técnicas de secuenciación de ADN rápidas y de bajo coste impulsó el desarrollo de técnicas de secuenciación de alto rendimiento (inicialmente conocidas como "next-generation") [12]-[14]. Estas técnicas son capaces de paralelizar las operaciones de secuenciación y reducir significativamente los costes habituales de este tipo de experimentos. La secuenciación de alto rendimiento combinada con el desarrollo de internet ha producido un crecimiento exponencial de los datos biológicos disponibles en los últimos 20 años y facilitado el acceso a ellos. A consecuencia de ello, se ha producido también un incremento significativo en el desarrollo de nuevas técnicas computacionales que permitan la inferencia de nuevos conocimientos a mayor velocidad y menor coste de los que se podría descubrir experimentalmente.

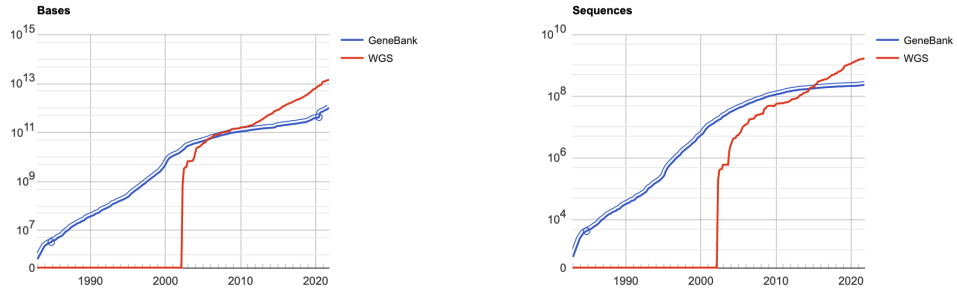


FIGURA 2.2: Crecimiento de los datos en GeneBank a lo largo de los años

Este crecimiento exponencial de los datos biológicos disponibles (ver figura 2.2) ha creado dos problemas fundamentales: por un lado, el almacenamiento y manejo eficiente de la información y, por otro, la extracción de información útil (conocimiento) a partir de dichos datos [60].

2.3.4. Minería de Datos en Bioinformática

Un enfoque para resolver dichos problemas es la minería de datos, la cual intenta descubrir y comprender patrones y modelos a partir de grandes cantidades de datos mediante la aplicación de técnicas como son el aprendizaje automático, los métodos estadísticos, la inteligencia artificial y la visualización y reconocimiento de patrones [87].

La minería de datos agrupa una serie de técnicas diferentes [88] y es un paso fundamental dentro de un proceso más complejo de análisis generalmente conocido como Knowledge Discovery in Databases (KDD) o Intelligent Data Analysis (IDA), el cual incluye el almacenamiento y procesamiento de dichos datos, la aplicación de diversos algoritmos, y la visualización e interpretación de los resultados. Por ello, se trata de un proceso iterativo, compuesto por diversas fases que deben repetirse y refinarse para proporcionar precisión y encontrar soluciones adecuadas [2] (ver figura 2.3).

El proceso de KDD parte de una comprensión de el dominio de la aplicación, el conocimiento previo existente y los objetivos del proceso. En base dicho conocimiento se crea un conjunto de datos de inicial, ya sea seleccionando un conjunto de datos completo o centrándose en un subconjunto de variables o muestras de datos.

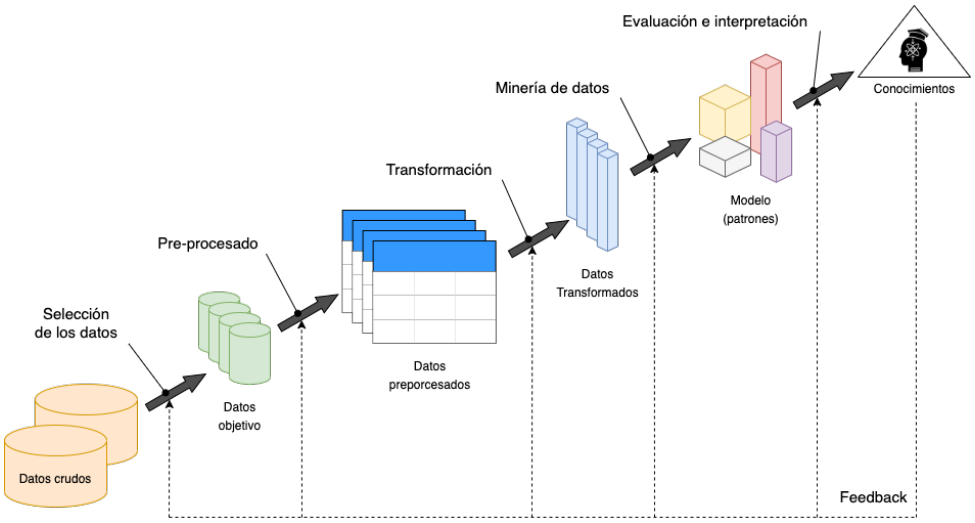


FIGURA 2.3: Extracción de conocimiento mediante el proceso de Knowledge Discovery Data (KDD)

Dicho conjunto de datos es limpiado y preprocesado para, por ejemplo, tratar valores atípicos o ruido, tratar datos faltantes, integrar datos de distintas fuentes o tratar cambios en los datos a lo largo de un periodo temporal.

Estos datos preprocesados son transformados mediante técnicas de reducción y proyección con el objetivo de encontrar las características o atributos útiles para representar los datos en función del objetivo de la tarea. Si la cantidad de datos o atributos es muy grande, se suelen utilizar métodos de reducción o transformación de dimensionalidad para reducir el número efectivo de variables bajo consideración o para encontrar representaciones invariables de los datos.

Finalmente, se llega a la fase de minería de datos, la cual suele tener dos objetivos principales: la predicción de eventos futuros y la descripción mediante la generación de modelos. Las tareas principales para extraer nuevos patrones significativos de los datos son:

- **Clasificación:** permiten el mapeo (clasificación) de un elemento dado en una de varias clases predefinidas.
- **Estimación:** dados algunos datos de entrada, permiten obtener un valor para alguna variable desconocida.
- **Predicción:** similar a la clasificación y la estimación, excepto que los elementos se clasifican/estiman de acuerdo con algún comportamiento o valor futuro estimado.
- **Reglas de asociación:** también llamado modelado de dependencia, permiten determinar qué elementos están relacionados de alguna manera.
- **Agrupación:** permiten la segmentación de una población en varios subgrupos.
- **Descripción y visualización:** permiten la representación de los datos mediante técnicas de visualización.

A finalizar esta etapa, se obtiene un modelo (o patrón) que, potencialmente, permite resolver el problema propuesto. Generalmente, el proceso de KDD es un proceso iterativo que puede conllevar la construcción de varios modelos intermedios no óptimos que van mejorando en cada iteración.

Finalmente, los modelos resultantes son validados mediante distintas técnicas para asegurar que sean robustos y precisos. Una vez validados, dichos modelos pueden utilizarse para resolver el problema propuesto.

La minería de datos se ha aplicado en múltiples industrias y ha sido particularmente exitosa en el campo de la bioinformática dados la gran cantidad de datos disponibles y la necesidad de hallazgos esenciales como son la expresión génica, el modelado de proteínas o el descubrimiento de fármacos [89].

2.3.5. Fuentes de datos biológicos

Aunque el conocimiento biológico ha alcanzado un nivel bastante representativo, está aun lejos de ser absoluto y sigue creciendo exponencialmente. Para ser útil y permitir la sostenibilidad de dicho crecimiento, estos datos deben ser procesados y almacenados de forma que sean accesibles por la comunidad científica. Durante los últimos 20 años se han creado una gran multitud de bases de datos para el almacenamiento y manejo de la información biológica existente. La mayoría de estas bases de datos se encuentran accesibles en línea en forma de sitios web que permiten a los usuarios navegar a través de los datos y realizar búsquedas. Adicionalmente, se han desarrollado múltiples formatos y estándares de forma que los datos puedan estar disponibles en distintos formatos para su descarga [90].

La revista científica Nucleic Acids Research (NAR) publica anualmente un número especial dedicado a las bases de datos biológicas. La edición de 2021 contiene 189 artículos, de los cuales 89 son nuevos y 90 son actualizaciones recientes sobre recursos que aparecieron en versiones anteriores del artículo. Toda esta información va acompañada de una base de datos llamada Online Molecular Biology Database Collection que categoriza 1641 bases de datos en línea [91]. Existen, además, otras colecciones de bases de datos como MetaBase [92] y Bioinformatics Links Collection [93].

Xiong categorizó las bases de datos biológicas en tres categorías en función de su contenido: primarias, secundarias y especializadas [94]. Las bases de datos primarias contienes datos biológicos obtenidos experimentalmente sin ningún tipo de tratamiento o aprendizaje automático. Las bases de datos secundarias se basan en la información existente en las bases de datos primarias, las cuales extiende mediante información inferida mediante experimentos computacionales o mediante la revisión manual de la literatura científica existente. Por último, las bases de datos especializas se centran en un campo de investigación concreto.

Por otro lado, debido a la reciente explosión en el número de bases de datos y repositorios de información biológica, es más práctico clasificar las bases de datos en función del tipo de datos que contiene y por lo tanto el área de estudio dentro de la bioinformática en el que son relevante.

Secuencias de nucleótidos (ADN)

Como se ha expuesto en la sección anterior, una de las áreas que ha tenido un crecimiento exponencial en los últimos años es el análisis de secuencias, el cual ha dado pie a muchos de los grandes avances que se han visto en el campo de la bioinformática.

Las bases de datos de secuencias de ácidos nucleicos están compuestas por una serie de entradas con información sobre la secuencia en sí, así como anotaciones funcionales y otra información relevante. Esta información se obtiene mediante el

procesado de la literatura científica en busca de publicaciones en las que se reportan fragmentos solapados que permitan identificar la secuencia completa.

De especial relevancia en cuanto a secuencias de ADN es la *International Nucleotide Sequence Database Collaboration (INSDC)*, un esfuerzo conjunto para recolectar secuencias de ADN y ARN que incluye: *GeneBank del National Center for Biotechnology (NCBI)* en EEUU, *European Nucleotide Archive del European Bioinformatics Institute (EBI)* en Reino Unido y *DNA Data Bank of Japan (DDBJ)* del National Institute of Genetics en Japón. Aunque las tres bases de datos presentan distintas características, se sincronizan diariamente por lo que ofrecen esencialmente la misma información.

Existen, además, un número de bases de datos secundarias sobre ADN codificante y no codificante, además de bases de datos sobre estructuras de genes, intrones y exones, así como bases de datos sobre factores de transcripción (TF) y la regulación transcripcional.

Secuencias de nucleótidos (ARN)

De manera similar, existen bases de datos especializadas en ARN. *miRBase* es una base de datos muy popular de secuencias de micro-ARN, un tipo de ARN relacionado con la regulación génica [95]. *Rfam* es una colección de familias de ARN en la que cada familia está representada por un alineamiento de secuencia múltiple, una estructura secundaria de consenso y un modelo de covarianza [96]. Además, ya que el ARN se pliega sobre se mismo formando complejas estructuras, existen diversas bases de datos dedicas específicamente a dichas estructuras como por ejemplo *BPS* [97]. Cabe destacar también *RNAcentral* una base de datos de secuencias de ARN no codificante (ARNnc) que consolida la información de múltiples bases de datos de ARNnc [98].

Genomas

Aparte de bases de datos centradas en secuencias de nucleótidos y genes individuales, existen bases de datos dedicadas a genomas completos. Destaca *Ensembl* [99] un proyecto conjunto del *EBI* y *The Wellcome Sanger Institute*, que surgió a raíz del Proyecto Genoma Humano y a día de hoy pretende ofrecer la información necesaria para genetistas, biólogos moleculares y otros investigadores que estudian los genomas de nuestra propia especie y otros vertebrados y organismos modelo [100]. *Ensembl* se nutre de datos de secuencias, los cuales son introducidos en un sistema automático de anotación de genes para crear un conjunto de predicciones y almacenarlas en la base de datos. El concepto central de *Ensembl* es su interfaz gráfica, la cual permite al usuario desplazarse por un genoma y observar la ubicación relativa de determinadas características como las distintas anotaciones (genes, loci SNP), patrones de secuencia (repeticiones) y datos experimentales (secuencias y características de secuencia externa mapeadas en el genoma). Además, *Ensembl* permite generar automáticamente vistas gráficas de la alineación de genes y otros datos genómicos con respecto a un genoma de referencia.

Existen además un gran numero de bases de datos dedicadas al genoma completo de un organismo concreto como son *Saccharomyces Genome Database* [101], *FlyBase* [102], *Mouse Genome Database (MGD)* [103] *Rat Genome Database (RGD)* [104], *WormBase* [105], *The Arabidopsis Information Resource (TAIR)* [106], *Xenbase Information Network* [107] o *Zebrafish Information Network* [108].

Secuencias de aminoácidos (Proteínas)

También existen bases de datos especializadas en secuencias de aminoácidos que forman proteínas. Las bases de datos más representativas en este campo son *The Protein Information Resource (PIR)* [109] del *National Biomedical Research Foundation de la Universidad Médica de Georgetown* en Washington; y *SWISS–PROT* y *TrEMBL* [110], del *Swiss Institute of Bioinformatics (SIP)* en Ginebra y del *European Bioinformatics Institute (EBI)*. Aunque inicialmente estas bases de datos coexistían y tenían distintas prioridades y cobertura, en 2003, las tres instituciones crearon un consocio y lanzaron *UniProt* [111], una base de datos centrada en la secuenciación de proteínas y en anotaciones funcionales. *UniProt* está formada por cuatro bases de datos: *UniProtKB* (con dos subpartes: *Swiss-Prot* y *TrEMBL*), *UniParc* [112] y *UniRef* [113].

UniProtKB/TrEMBL contiene entradas generadas computacionalmente y anotadas de forma automática. Una vez estás una proteína ha sido analizada en detalle por un bio-curador utilizando múltiples herramientas de análisis de secuencias así como la literatura científica existente y ha pasado un control de calidad, ésta se añade a *UniProtKB/Swiss-Prot* para obtener una base de datos de alta calidad y no redundante (todas las variantes de una proteína se combinan en una sola entrada). Las anotaciones que se suelen añadir a las proteínas son, entre otras, los nombres de la proteína y genes asociados, su función, información específica sobre enzimas (como actividad catalítica, cofactores y residuos catalíticos o ubicación subcelular), sus interacciones con otras proteína, sus patrones de expresión, sus ubicaciones y roles significativos, sitios de unión de iones, sustratos y cofactores o formas variantes (producidas por variación genética natural, edición de ARN u empalme alternativo, procesamiento proteolítico o modificación postraduccional). La separación entre estas dos bases de datos se debe a que debido a la explosión en la cantidad de datos generados por experimentos genómicos no todas las secuencias disponibles pueden ser manualmente revisadas y anotadas, lo cual requiere una gran cantidad de tiempo y esfuerzo.

UniProt Archive (UniParc) es una base de datos completa y no redundante, que contiene todas las secuencias de proteínas de las principales bases de datos de secuencias de proteínas disponibles públicamente incluyendo *INSDC EMBL-Bank / DDBJ / GenBank nucleotide sequence databases*, *Ensembl*, *European Patent Office (EPO)*, *FlyBase*, *H-Invitational Database (H-Inv)*, *International Protein Index (IPI)*, *Japan Patent Office (JPO)*, *Protein Information Resource (PIR-PSD)*, *Protein Data Bank (PDB)*, *Protein Research Foundation (PRF)*, *RefSeq*, *Saccharomyces Genome Database (SGD)*, T*he Arabidopsis Information Resource (TAIR)*, *TROME*, *US Patent Office (USPTO)*, *UniProtKB / Swiss-Prot*, *UniProtKB/Swiss-Prot protein isoforms*, *UniProtKB/TrEMBL*, *Vertebrate and Genome Annotation Database (VEGA)* y *WormBase*.

Por último, *UniProt Reference Clusters (UniRef)* consta de tres bases de datos de conjuntos de secuencias de proteínas de *UniProtKB* y registros *UniParc* seleccionados agrupados según su similitud. *UniRef100* combina secuencias y subfragmentos idénticos de cualquier organismo en una única entrada *UniRef*. *UniRef90* y *UniRef50* se crean agrupando las secuencias de *UniRef100* en los niveles de identidad de secuencia de 90 % y 50 % respectivamente.

Análisis de estructura de proteínas

Como se vio anteriormente, la estructura final de una proteína es determinante para su funcionalidad por lo que se capaces de entender y ser capaz de predecir dichas estructuras a partir de secuencias y otra información es esencial. Para ello,

existen varias bases de datos que almacenan y anotan la estructura de proteínas típicamente obtenidos por cristalografía de rayos X, Espectroscopia de resonancia magnética nuclear de proteínas (NMR) o, cada vez más, por microscopía crioelectrónica,.

La base de datos más representativa es el *Protein Data Bank (PDB)* [69], una base de datos sobre la estructura de macromoléculas mantenida por el consorcio internacional *Worldwide Protein Data Bank (wwPDB)* [114] y que incluye el *Protein Data Bank of Europe (PDBe)* [115], el *Protein Data Bank in Japan (PDBj)* y el *Research Collaboratory for Structural Bioinformatics (RCSB)* [116].

Además, existen multitud de bases de datos que se nutren de los datos contenidos en *PDB*. Por ejemplo, *Structural Classification of Proteins (SCOP)* [117] o *Class Architecture Topology Homology (CATH)* [118] ofrecen una estructura jerarquizada de las proteínas en *PDB*. Recientement, ha sido publicada *SCOPe*, como una extensión de SCOP mejorando ciertos aspectos [119].

Familias de proteínas y motivos estructurales en proteínas

En biología molecular se considera que una proteína pertenece a la misma familia que otra cuando ambas están relacionadas mediante un ancestro en común, es decir, son homólogas. Dichas proteínas típicamente tienen secuencias, estructuras tridimensionales y funciones significativamente similares, aunque el indicador más importante de homología es sin duda la alineación de sus secuencias. Estas familias, además, se agrupan en grupos llamados superfamilias basados en la similitud estructural y mecanicista, incluso si no hay una homología identificable en sus secuencias (ver figura 2.4).

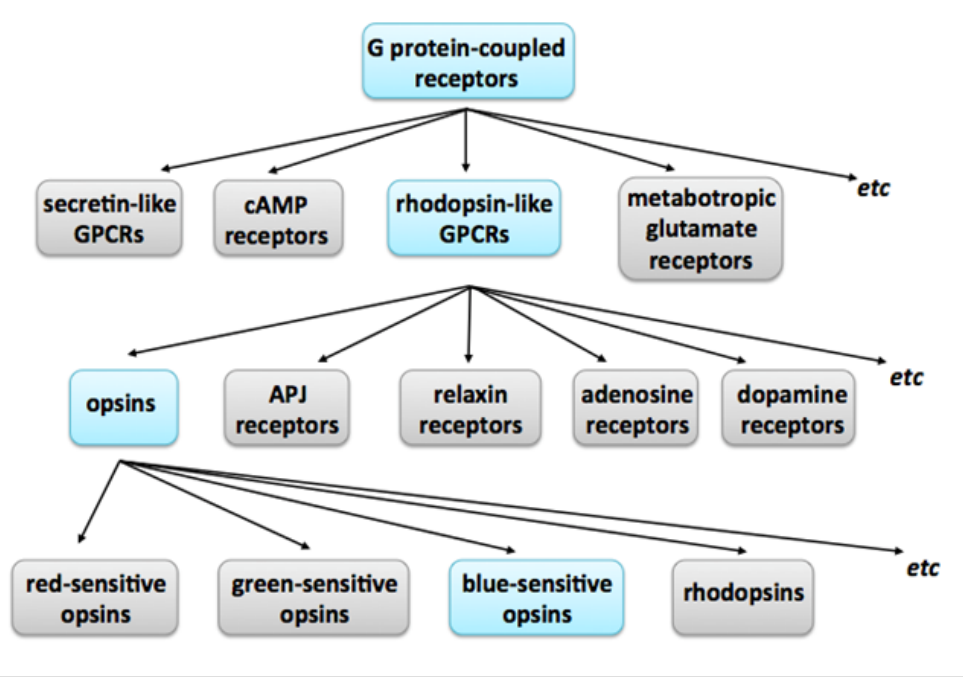


FIGURA 2.4: Ejemplo de familias y superfamilias a las que pertenece la opsina 1 (en azul) desde las opsinas sensibles a la luz azul hasta los receptores acoplados a proteínas G (GPCRs)

El concepto de familia de proteínas se concibió en una etapa muy temprana y se centraba principalmente en proteínas pequeñas con un solo dominio, como la mioglobina o la hemoglobina. Sin embargo, en los últimos años se ha descubierto que muchas proteínas están formadas por múltiples unidades o dominios estructurales y funcionales, los cuales, han evolucionado y evolucionan independientemente. Esto ha cambiado el enfoque, en los últimos años, hacia las familias de dominios de proteínas.

La identificación correcta de las familias de proteínas es crítica para el análisis filogenético, la anotación funcional y la exploración de la diversidad de las funciones proteicas en una rama filogenética dada.

Para ello, existen multitud de bases de datos centradas en las familias de proteínas, las cuales permiten a los usuarios identificar si una proteína identificada recientemente pertenece a una familia conocida. Caben destacar *Pfam* [120], centrada en el alineamientos de familias estructurales; *PROSITE* [121], centrada en dominios proteicos, familias y sitios funcionales; *PIRSF* [122], un sistema de clasificación de superfamilias o las ya mencionadas *SCOP* y *CATH*.

Anotaciones funcionales

Como se ha indicado anteriormente, aparte de identificar las secuencias que forman entidades biológicas (ADN, ARN, proteínas, etc.), es muy relevante anotarlas con información adicional que ayude a comprender su funcionamiento.

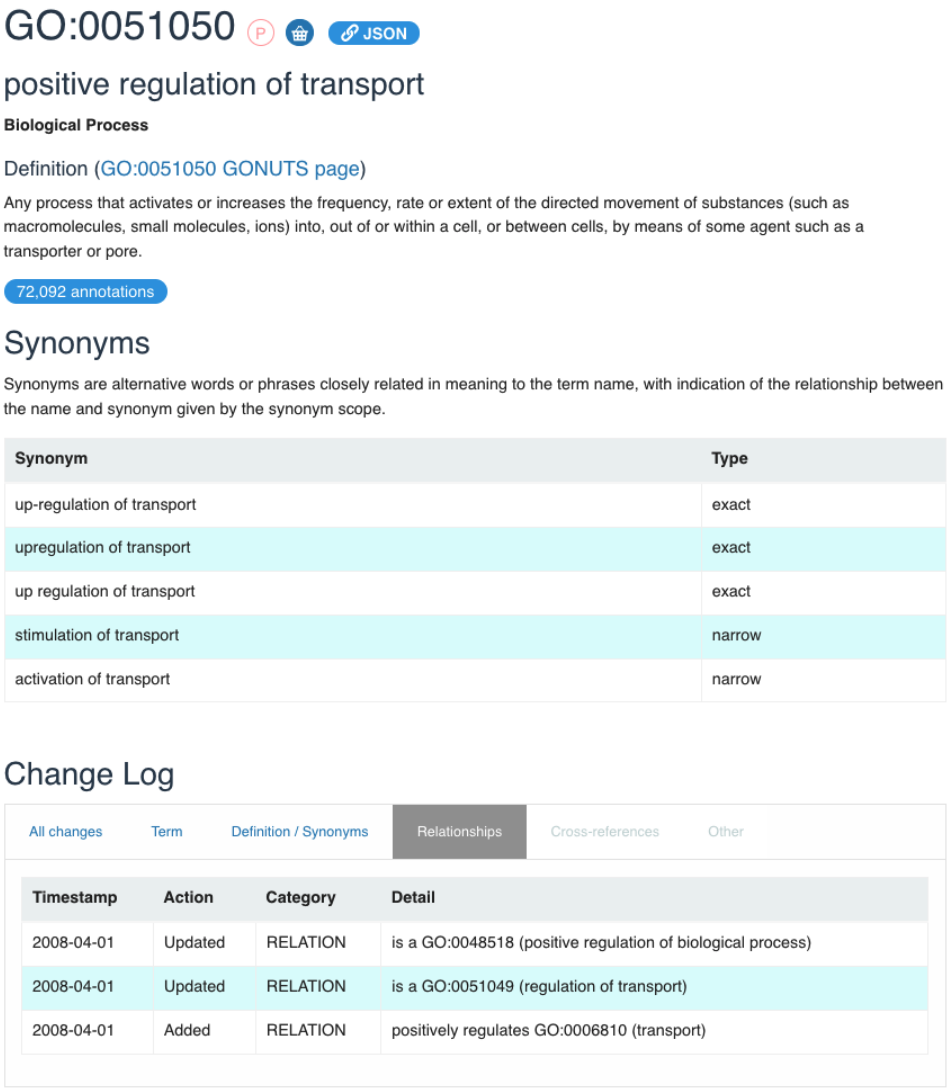
En este contexto, surge en 1998 *Gene Ontology (GO)* [123], [124] como una colaboración entre *FlyBase*, *Saccharomyces Genome Database (SGD)* y *Mouse Genome Database (MGD)* con el objetivo de ofrecer anotaciones sobre genes y productos génicos de forma consistente. Desde su creación, multitud de organismos modelo han sido añadidos y prácticamente todas las bases de datos principales contribuyen a GO hoy en día. Actualmente, GO contiene casi 8 millones anotaciones sobre más de 1 millón y medio de productos génicos pertenecientes a más de 5 mil especies, respaldados por más de 166000 artículos científicos publicados.

La principal diferencia entre Gene Ontology y otras bases de datos de anotaciones funcionales es que, como su propio nombre indica, almacena dichas anotaciones en forma de ontología. Una ontología es una representación formal de un cuerpo de conocimiento dentro de un dominio dado. Las ontologías suelen consistir en un conjunto de clases o términos con relaciones que operan entre ellos.

Términos en GO Cada término en GO contiene un identificador único de siete dígitos precedido por "GO", un nombre, una descripción y su relación con otros términos. Además, cada término GO puede contener identificadores secundarios, sinónimos (los cuales pueden ser exactos, más genéricos, más concretos o solo relacionados), referencias cruzadas a otras bases de datos y otros comentarios. Por último, a veces se agrega un término a GO que está fuera de contexto, es erróneo o debería representarse de otra manera. En estos casos, el término se etiqueta como obsoleto, se eliminan todas las relaciones con otros términos y se agrega un comentario detallando el motivo de la obsolescencia y etiquetas que especifican los términos de reemplazo.

En la figura 2.5 se puede ver la información existente en GO para el término GO:0051050 que indica un regulador positivo del transporte. Este término es aplicable cualquier proceso que activa o aumenta la frecuencia, velocidad o extensión del movimiento dirigido de sustancias (como macromoléculas, moléculas pequeñas, iones) hacía dentro, hacia fuera, en el interior de una célula, o entre células, por medio

de algún agente como un transportador o poro. Este término, tiene varios sinónimos como son la regulación hacia arriba del transporte, la estimulación del transporte o la activación del transporte y está relacionado con varios otros términos.



GO:0051050 JSON

positive regulation of transport

Biological Process

Definition (GO:0051050 GONUTS page)

Any process that activates or increases the frequency, rate or extent of the directed movement of substances (such as macromolecules, small molecules, ions) into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore.

72,092 annotations

Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

Synonym	Type
up-regulation of transport	exact
upregulation of transport	exact
up regulation of transport	exact
stimulation of transport	narrow
activation of transport	narrow

Change Log

All changes | Term | Definition / Synonyms | Relationships | Cross-references | Other

Timestamp	Action	Category	Detail
2008-04-01	Updated	RELATION	is a GO:0048518 (positive regulation of biological process)
2008-04-01	Updated	RELATION	is a GO:0051049 (regulation of transport)
2008-04-01	Added	RELATION	positively regulates GO:0006810 (transport)

FIGURA 2.5: Ejemplo de información del término en GO:0051050, regulador positivo del transporte.

Relaciones entre términos y la estructura de GO Como se ha mencionado, una de las grandes ventajas de GO es como sus términos están relacionados. La estructura de GO se puede describir en términos de un grafo acíclico dirigido (DAG, Directed Acyclic Graph), donde cada término de la ontología es un nodo y las relaciones entre los términos son las aristas. GO es vagamente jerárquico, en el sentido de que parte de una raíz y los términos se vuelven más especializados a medida que se alejan de ella. Sin embargo, a diferencia de los árboles u otras estructuras jerárquicas más estrictas, cada nodo puede tener múltiples padres.

El DAG de GO tiene 3 nodos raíz, es decir, que en realidad no se trata de una ontología sino de tres: funciones moleculares, componentes celulares y procesos biológicos.

Los términos que desciende de función molecular denotan una acción o actividad que se realiza en el contexto de un proceso molecular. Estas acciones se describen desde dos perspectivas distintas pero relacionadas: actividad bioquímica, y papel como un componente en un sistema o proceso mayor.

Los términos que desciende de componente celular describen una ubicación, relativa a los compartimentos y estructuras celulares, ocupada por un conjunto de moléculas cuando llevan a cabo una función molecular. Hay dos formas en que los biólogos describen las ubicaciones de los productos génicos: en relación con estructuras celulares (Lado citoplásmico de la membrana plasmática) o compartimientos (mitocondria) y complejos macromoleculares estables de los que son parte (ribosoma). A diferencia de los otros términos de GO, los términos relativos a componentes celulares no se refieren a procesos sino a una anatomía celular.

Los términos que descienden de proceso biológico denotan un objetivo específico que el organismo está genéticamente programado para lograr. Los procesos biológicos a menudo se describen por su resultado o estado final, por ejemplo, el proceso biológico de división celular da como resultado la creación de dos células hijas a partir de una única célula parental. Un proceso biológico se logra mediante un conjunto particular de funciones moleculares llevadas a cabo por productos génicos específicos (o complejos macromoleculares), a menudo de una manera altamente regulada y en una secuencia temporal particular.

Dentro del DAG, los términos GO pueden estar relacionados de distintas maneras:

- *is_a* **(es un/una)** es la relación fundamental de GO. Si decimos que A es un B , queremos decir que el nodo A es un subtipo (y no una instancia) del nodo B . Por ejemplo, el ciclo celular mitótico es un ciclo celular, o la actividad liasa es una actividad catalítica. Todos los elementos de GO tienen garantizado tener un padre al cual están relacionados mediante "*is_a*", es decir, que el el grafo de GO considerando solo dichas relaciones es conexos. Además, la relación "*is_a*" es transitiva (si A es B y B es C , podemos inferir que A es C) y agrupable (si A participa en B y B es C , A participa en C). "*is_a*" es el único tipo de relación que no cruza las ontologías.

- *part_of* **(parte de)** indica que A es necesariamente parte de B , es decir, dondequiera que exista A , es como parte de B , y la presencia de B implica la presencia de A . Sin embargo, dada la ocurrencia de B , no podemos decir con certeza que A existe. "*part_of*" es transitiva y agrupable al igual que "*is_a*". Dicha transitividad también se puede aplicar a las relaciones "*is_a*" (si A es B y B es parte de C , A es parte de C ; o si A es parte de B y B es C , A es parte de C).

- *has_part* **(tiene como parte)** es la relación complementaria a "*part_of*". En GO, A tiene B como una parte si y solo si A necesariamente tiene la parte B , es decir, si A existe, B siempre existirá. Sin embargo, si B existe, no podemos decir con certeza que A exista. "*has_part*" es transitiva y agrupable de la misma manera que "*part_of*".

- *regulates* **(regula)** representa aquellas relaciones en la que un proceso afecta directamente la manifestación de otro proceso o cualidad, es decir, el primero regula el último. La relación "*regulates*" tiene dos sub-relaciones, "*positively_regulates*" (regula positivamente) y "*negatively_regulates*" (regula negativamente), para precisar mejor el tipo de regulación. "*regulates*" es transitiva y agrupable de

la misma manera que "*part_of*" y "*has_part*". Además, si *A* regula (positiva o negativamente) a *B* y *B* es parte de *C* , podemos inferir que *A* regula a *C* (pero no sabemos si positiva o negativamente).

- **Extensiones** GO también incluye una serie de relaciones consideradas como adicionales como son "*acts_on_population_of*", "*has_agent*", "*causally_upstream_of*", "*exists_during*", "*happens_during*", "*has_participant*", "*has_input*", "*stabilizes*", "*has_output*", "*results_in_formation_of*", "*results_in_acquisition_of_features_of*", "*results_in_development_of*", "*results_in_movement_of*", "*results_in_morphogenesis_of*", "*results_in_maturation_of*", "*results_in_division_of*", "*regulates_o_has_participant*", "*has_regulation_target*", "*regulates_o_has_input*", "*regulates_expression_of*", "*regulates_transcription_of*", "*regulates_translation_of*", "*occurs_at*", "*occurs_in*", "*part_of*", "*activated_by*", "*inhibited_by*", "*results_in_commitment_to*", "*results_in_determination_of*", "*results_in_specification_of*", "*has_part*" o "*during*".

Siguiendo con el ejemplo anterior y mirando la figura 2.6, se pueden ver como la "regulación positiva del transporte" (GO:0051050) es una "regulación del transporte" (GO:0051049) y regula positivamente el "transporte" (GO:0006810). Además, si continuáramos el GO DAG podríamos ver que la "regulación positiva del transporte de agua renal" (GO:2001153), "regulación positiva del transporte de pentasacáridos" (GO:1900362), "regulación positiva del transporte intracelular" (GO:0032388), "regulación positiva de la fagocitosis" (GO:0050766), "regulación positiva de la captación de epinefrina" (GO 0051628), "regulación positiva del transporte de laminaritriosa" (GO:1900305), "regulación positiva del transporte de toxinas" (GO:1902009), "regulación positiva del transporte de compuestos que contienen nucleobase" (GO:0032241), "regulación positiva del transporte de heptasacáridos" (GO:1900296), "regulación positiva de la galactotriosa tran" (GO:1900293) son todos "regulación positiva del transporte" (GO:0051050). Y si siguiéramos inspeccionado veríamos, por ejemplo, que la "regulación positiva del transporte de vesículas a lo largo de los microtúbulos" (GO: 1901610) es "regulación positiva del transporte intracelular" (GO:0032388).

Anotaciones en GO Una anotación en GO es una declaración sobre la función de un producto génico particular. Cada anotación GO consiste en la asociación entre un producto génico y un término GO y representa una "instantánea" del conocimiento biológico actual. Estas anotaciones son siempre respaldadas por la literatura científica, directa o indirectamente. GO utiliza diferentes niveles para describir el grado de evidencia científica del conocimiento descrito por una determinada anotación. Para ello, todas las anotaciones contienen un código de evidencia y la referencia a una publicación científica o la descripción de la metodología utilizada para generar la anotación. GO incluye los siguientes códigos de evidencia:

- **Anotaciones apoyadas experimentalmente (EXP, EXPerimental)** Anotaciones para las que existe evidencia experimental. Estas anotaciones son creadas por expertos revisores que analizan la literatura científica en busca de evidencias. Existen varios subtipos de este tipo de evidencia: inferidas directamente a partir de un articulo (IDA, Inferred from Direct Assay), inferidas a partir interacción física (IPI), inferidas a partir de un fenotipo mutante (IMP), inferidas a partir de interacciones genética (IGI) y inferidas a partir del patrón de expresión (IEP).

- **Anotaciones inferidas filogenéticamente (IBA, Inferred from Biological Ancestry)** Anotaciones inferidas mediante la reconstrucción de los eventos evolutivos. GO tiene su propio software llamado PAINT (Phylogenetic Annotation

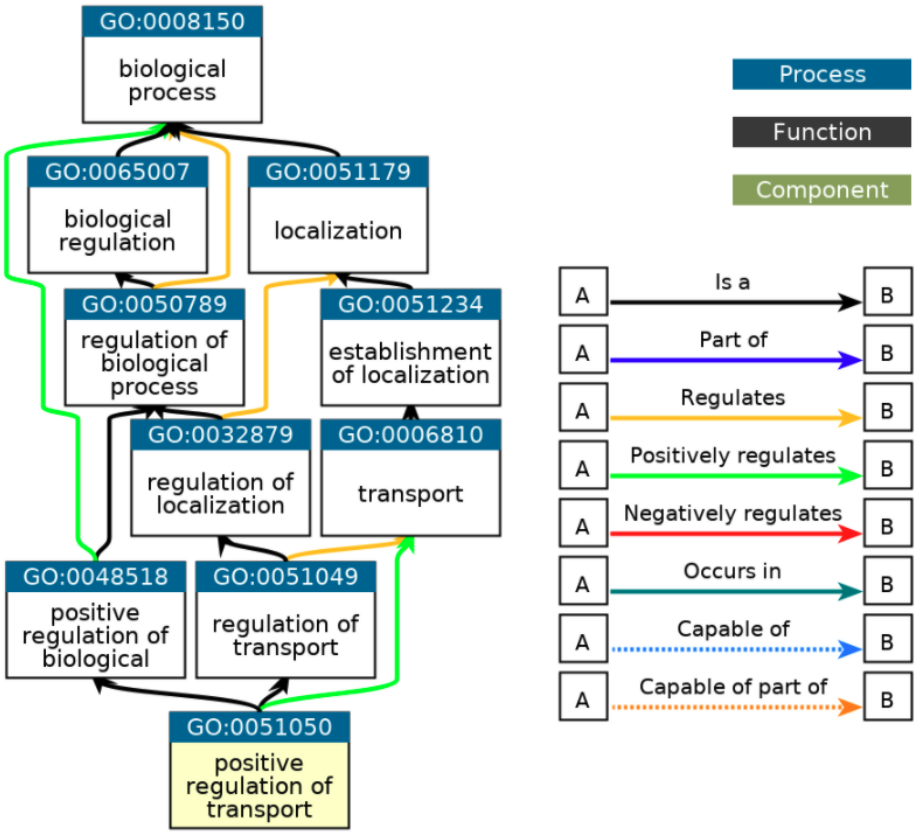


FIGURA 2.6: Ejemplo del GO DAG para GO:0051050 mostrando solo los ancestros.

Inference Tool) con el que un revisor experto puede ver las anotaciones experimentales para genes en una determinada familia, e inferir anotaciones para miembros no caracterizados de la familia [125]. Existen varios subtipos de este tipo de anotaciones: inferidas a partir del aspecto biológico de un ancestro (IBA), inferidas a partir del aspecto biológico de un descendiente (EII), inferidas a partir de residuos clave (IKR) o inferidas por divergencia rápida (IRD).

- **Anotaciones inferidas computacionalmente** Anotaciones basadas en análisis in-silico de la secuencia del gen y/u otros datos. Existen varios tipos de anotaciones en esta categoría indicando un grado variable de aportación curatorial: inferidas a partir de Secuencia o Similitud estructural (ISS), inferidas a partir de Sequence Orthology (ISO), inferidas a partir de la alineación de secuencia (ISA), inferidas a partir de modelo de secuencia (ISM), inferidas a partir de contexto genómico (IGC), inferidas a partir de Análisis Computacional Revisado (RCA).

- **Anotaciones basadas en las declaraciones de un autor** Anotaciones realizadas en base a una declaración hecha por el autor o autores en la referencia citada. Los códigos de evidencia de la declaración del autor son: declaración de autor rastreable (TAS) y declaración de autor no rastreable (NAS).

- **Anotaciones basadas en las declaraciones del curador** Anotaciones realizadas en base a un juicio curatorial que no encaja en una de las otras clasificaciones.

Los códigos de la declaración curatorial son: inferida por el curador (IC) y no hay datos biológicos disponibles (ND).

- **Anotaciones inferidas electrónicamente (IEA)** Anotaciones que no se revisan individualmente (aunque generalmente realiza una revisión manual de una muestra). Suelen estar basada en homología y/u otra información experimental o de secuencia, pero generalmente no se puede rastrear una evidencia experimental.
- **Extensiones** Las anotaciones de GO están muy restringidas por el formato inicial diseñado para favorecer la simplicidad y la adopción de GO por la comunidad científica. Para superar esta limitación, GO ofrece extensiones sobre las anotaciones, las cuales ofrecen información adicional [126].

Siguiendo con el ejemplo anterior, como se puede ver en la figura 2.5, la "regulación positiva del transporte" (GO:0051050) está presente en más de 70 mil anotaciones. En la figura 2.7 se pueden ver algunas anotaciones para homo sapiens. Mirando la primera anotación, se observa que el gen "TCAF2", cuando produce la proteína "TRPM8 channel-associated factor 2", regula positivamente las proteínas dirigidas a la membrana (GO:0090314) y por herencia "regulación positiva del transporte" (GO:0051050). La notación ha sido "confirmada manualmente mediante fenotipo mutante" (ECO:0000315) que indica que ha sido "confirmada manualmente mediante evidencia fenotípica experimental" (ECO:0000315). Dicha experimentación queda evidenciada en Gkika et al. [127].

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon	Assigned By	Annotation Extension
UniProtKB:A6NFQ2	TCAF2	involved_in	GO:0090314 positive regulation of protein targeting to membrane	ECO:0000315 IMP	PMID:25559186		9606 Homo sapiens	UniProt	
UniProtKB:A6NI73	LILRA5	involved_in	GO:0051928 positive regulation of calcium ion transport	ECO:0000314 IDA	PMID:12393390		9606 Homo sapiens	UniProt	regulates_o_occurs_in (CL:0000576)
UniProtKB:O00159	MYO1C	involved_in	GO:0090314 positive regulation of protein targeting to membrane	ECO:0000315 IMP	PMID:23262137		9606 Homo sapiens	UniProt	has_input (UniProtKB:P35968)
UniProtKB:O00187:PRO_0000027599	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:24174618		9606 Homo sapiens	ComplexPortal	
UniProtKB:O00187:PRO_0000027599	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:22966085		9606 Homo sapiens	ComplexPortal	
UniProtKB:O00187:PRO_0000027599	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:16116205		9606 Homo sapiens	ComplexPortal	
UniProtKB:O00187:PRO_0000027599	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:11907111		9606 Homo sapiens	ComplexPortal	
UniProtKB:O00187:PRO_0000027599	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:10679061		9606 Homo sapiens	ComplexPortal	
UniProtKB:O00187:PRO_0000027600	MASP2	involved_in	GO:1903028 positive regulation of opsonization	ECO:0000314 IDA	PMID:24174618		9606 Homo sapiens	ComplexPortal	

FIGURA 2.7: Información sobre anotaciones en GO en las que está presente GO:0051050

Interacciones entre entidades biológicas

Como ya se vio al principio de este capítulo, hoy en día sabemos que los genes no se expresan de forma aislada, si no en el contexto de un complejo sistema biológico que depende tanto de las interacciones entre los elementos internos al sistema como con otros factores externos. Por ello, cada vez son más los estudios centrados en entender las distintas relaciones entre genes, productos genéticos (proteínas) y otras moléculas, así como las bases de datos centradas en dichas interacciones.

Database of Interacting Proteins (DIP) [128] fue la primera base de datos creada para almacenar interacciones física proteína–proteína (PPI) experimentalmente probadas. *DIP* enriquece la información sobre las interacciones con detalles sobre el método experimental utilizado así como referencias a artículos publicados que evidencien dicha interacción. Un gran número de bases de datos han sido creadas desde entonces y el enfoque se ha expandido al estudio de interacciones entre genes, proteínas y otras moléculas en lugar de solo entre proteínas. Por ejemplo, *Biomolecular Interaction Network Database (BIND)* [129], *Human Protein Reference Database (HPRD)* [130], *IntAct* [131], *Molecular Interaction Database (MINT)* [132] o *MIPS* [133].

Cabe destacar particularmente *BioGRID* [134], una base de datos dedicada a la anotación y almacenamiento de interacciones proteicas, genéticas y químicas para todas las especies de organismos modelo y humanos. *BioGRID* se nutre de interacciones evidenciadas por la literatura, las cuales son añadidas a la base de dato mediante un proceso manual de revisión por parte de expertos. Según los últimos datos publicados, en diciembre de 2021, *BioGRID* contenía más de 2 millones de interacciones genéticas y de proteínas, y más de 1 millón de modificaciones postraduccionales, anotadas manualmente a partir de casi 80 mil publicaciones. Además, para facilitar los enfoques basados en redes para el descubrimiento de fármacos, *BioGRID* ha incorporado un gran número de interacciones químico-proteína extraídos de *DrugBank* [135] para ayudar en el desarrollo de fármacos (ver figura 2.8).

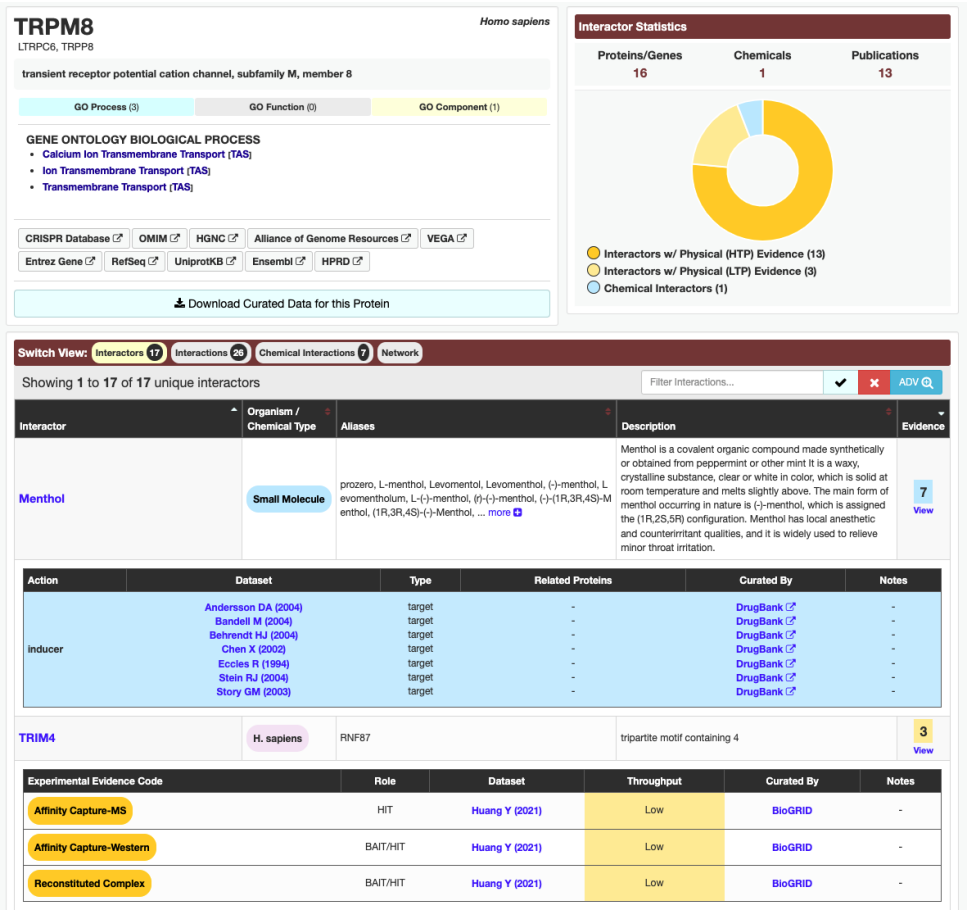


FIGURA 2.8: Información sobre interacciones de la proteína TRPM8 en BioGRID

En Diciembre de 2011 se estimaba que había más de 1000 bases de datos sobre

interacciones proteína-proteína, todas ellas con su propia financiación y sus propios objetivos. Para mitigar los problemas derivados de dicha dispersión, se creó un formato de archivo común para representar datos de interacción de proteínas y se publicaron las directrices sobre *Información mínima sobre experimentos de interacción molecular (MIMIX)* [136] que definen una lista de verificación de la información que se debe proporcionar al describir los datos de interacción molecular experimental en un artículo de revista. Finalmente, en 2012 se crea el *International Molecular Exchange Consortium (IMEx*) [137], una colaboración internacional entre los principales proveedores de datos de interacción para compartir el esfuerzo de curación de los datos y crear un conjunto no redundante de 'roteínas, actualmente centralizado en la base de datos *IntAct*. A día de hoy *IMEx* esta formado por *DIP*, *HPIDB*, *IntAct*, *MBInfo*, *MINT*, *MatrixDB*, *Molecular Connections*, *I2D*, *InnateDB*, *UCL-BHF group*, *UCL London, UniProt group y Swiss-Prot group (SIB)*, *EMBL-EBI* como socios activos, *BioGRID* y *PrimesDB* como observadores, y *MPact*, *BIND* y *MPIDB* como socios inactivos.

Existen, además, bases de datos de interacciones que no solo se centran en interacciones experimentalmente probadas si no que también incluyen interacciones predichas mediante técnicas computacionales. Por ejemplo, *STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)* [138] no solo almacena las interacciones físicas entre proteína si no que también almacena las relaciones funcionales siempre y cuando sean significativas. Además, no solo almacena derivadas de la literatura y demostradas experimentalmente si no que también almacena interacciones computacionalmente inferidas mediante la minería de texto de textos científicos, calculadas a partir de características genómicas, o transferidas desde organismos modelo basándose en ortología. También cabe destacase *GeneMANIA* [139] creada más recientemente y que contiene interacciones inferidas por coexpresión génica, interacción genética, interacción física, dominios compartidos entre proteínas, colocalización o predicciones computacionales. *GeneMANIA* permite combinar la información de distintas fuentes en una sola red utilizando distintos algoritmos de prioritización.

Existen además un gran número de bases de datos especializadas en el interactoma de un organismo concreto como son *YeastNet* [140], *PPIM* [141], *Drosophila Interactions Database (DroID)* [142], *Flynet server* [143], *SPiD* [144] o *AtPID* [145].

Como se ha mencionado anteriormente, aparte de estudiar las interacciones individualmente, es tremendamente útil ser capaces de establecer todas las relaciones existentes en un proceso biológico concreto. En este sentido cobran especial relevancia las bases de datos de rutas metabólicas y de transducción de señal. Una ruta metabólica es una serie de reacciones químicas relacionadas que ocurren dentro de una célula, generalmente en una posición determinada dentro de la célula.

El metabolismo de una célula consiste en una elaborada red de vías interconectadas que permiten la síntesis y la descomposición de las moléculas (anabolismo y catabolismo). El flujo de metabolitos a través de una ruta se regula dependiendo de las necesidades de la célula y la disponibilidad del sustrato, contribuyendo al mantenimiento de la homeostasis dentro del organismo. El producto final de una ruta puede usarse de inmediato, iniciar otra ruta metabólica o almacenarse para un uso posterior.

Algunas bases de datos destacadas sobre rutas metabólicas son *KEGG* [146], *BioCyc* [147], *Reactome* [148] y *NCI-Nature Pathway Interaction Database* [149]. Estas bases de datos difieren en diversos aspectos y todas tienen sus propias fortalezas y debilidades. Por ejemplo, ofrecen soluciones diferentes para problemas técnicos tales como cómo presentar los datos al usuario, cómo consultar la base de datos o el formato en el que exportar dichos datos. Sorprendentemente, existe una falta de consenso importante en este tipo de bases de datos por lo que solo un pequeño

porcentaje de las interacciones coincide en todas ellas. A pesar de ello, análisis de enriquecimiento GO sobre los genes de consenso muestra una evidencia de que existe un núcleo de procesos metabólicos acordes en todas las bases de datos.

Finalmente, como una forma de corregir la falta de consenso y la cobertura desigual de las distintas bases de datos sobre el conocimiento existente, existen metabases de datos, las cuales agregan datos de otras bases de datos existentes. Destacan en cuanto a interacciones proteína-proteína: Agile Protein Interactomes Dataserver (APID) [150], The Microbial Protein Interaction Database (MPIDB) [151], Protein Interaction Network Analysis (PINA) platform, [152], and Wiki-Pi [153]. De forma similar, ConsensusPathDB [154] y Pathways Commons [155] agregan data de las bases de datos de rutas metabólicas.

2.3.6. Herramientas bioinformáticas

Como se ha mencionado anteriormente, la generación de esta enorme cantidad de datos requiere de algoritmos y metodologías que permitan procesarlos y extraer conocimiento a partir de ellos.

Uno de los primeros tipos algoritmos en desarrollarse para la investigación bioinformática fueron los algoritmos de alineamiento de secuencias de nucleótidos (ADN, ARN) o aminoácidos (proteínas). Este tipo de algoritmo permite comparar secuencias en busca de zonas comunes, las cuales podrían indicar relaciones funcionales o evolutivas o de zonas no coincidentes en proteínas relacionadas, las cuales podrían indicar mutaciones puntuales (sustituciones, inserciones o deleciones). La alineación de secuencias puede ser global, intenta alinear cada residuo de cada secuencia, o local, más útil para secuencias diferenciadas en las que se sospecha que existen regiones muy similares. Como ejemplo de búsqueda global destaca el algoritmo de Needleman–Wunsch [64] y como ejemplo de búsqueda local el algoritmo de Smith-Waterman [156]. El rendimiento de estos algoritmos puede no ser suficiente para la gran cantidad de datos que un experimento bioinformático necesita analizar. Para ello, la comunica científica ha desarrollado métodos heurísticos como FASTA [157] o BLAST [158]. Existen también algoritmos que en lugar de alinear las secuencias por pares intentan alinear múltiples secuencias, entre los que destacan Clustal [159] y T-Coffee [160].

Existe, además, una gran variedad de software bioinformático con otros objetivos, como, por ejemplo, alineamiento estructural de proteínas, predicción de funcionalidad genética, predicción de estructura de proteínas, predicción de acoplamiento proteína-proteína, inferencia de regulación génica a partir de análisis de expresión genética, la identificación de mutaciones causantes de enfermedades o modelado de sistemas biológicos. Los algoritmos de inferencia de regulación génica así como de validación de redes genéticas son particularmente relevantes para esta tesis y se discuten en profundidad en el las secciones 3.3.1 y 3.3.2.

Para facilitar el uso algunos de estos algoritmos, el acceso a bases de datos y el desarrollo de nuevas aplicaciones, existen librerías de código abierto para realizar aplicaciones bioinformáticas en multitud de lenguajes informáticos (Bioconductor [161], BioPerl [162], Biopython [163], BioJava [164], BioJS [165], BioRuby [166], Bioclipse [167], .NET Bio). Bioconductor, por ejemplo, contiene más de 2000 paquetes de software, está disponible de diversas maneras (código, maquina virtual en la nube o contenedor Docker) y tiene una importante comunidad respaldándolo. Sin embargo, ya que la bioinformática es un campo multidisciplinar, no todos los usuarios de métodos bioinformáticos tienen los suficientes conocimiento de programación para poder utilizar este tipo de software.

En este sentido, para facilitar el acceso a técnicas bioinformáticas sin la necesidad de conocimiento informáticos, existe una gran oferta de servicios web que permiten realizar diversos análisis bioinformáticos, entre los que se pueden destacar: EMBL FASTA [168], NCBI BLAST [169], ClustalW [170] o T-Coffee [160] para el alineamiento de proteínas, o EMBOSS [171] o Gene Ontology [123] para otros tipos de análisis.

Cytoscape

Ya que este documento se centra principalmente en la inferencia, análisis y validación de redes genéticas, cobra especial relevancia Cytoscape [45], [172]. Cytoscape es una plataforma de código abierto para la visualización y el análisis de redes así como la integración de esta con distintos datos en forma de atributos (ver figure 2.9. Aunque Cytoscape es agnóstico en términos de uso, es decir, se puede utilizar para visualizar y analizar cualquier tipo de red o grafo (por ejemplo, redes sociales), fue diseñado y se utiliza mayoritariamente para analizar redes biológicas (por ejemplo, visualizar redes de interacción molecular y rutas biológicas e integrar estas redes con anotaciones o perfiles de expresión génica).

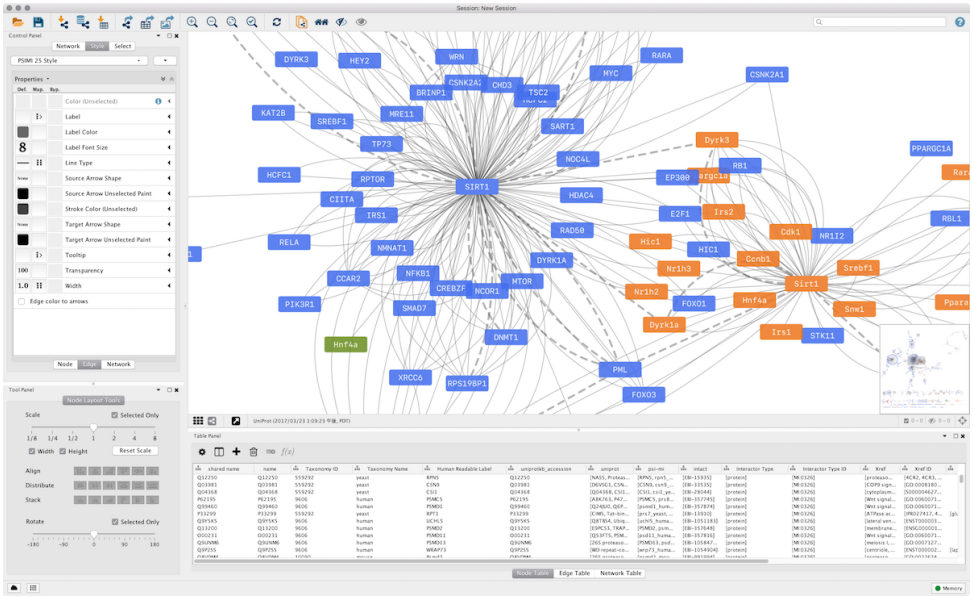


FIGURA 2.9: Cytoscape UI

Núcleo proporciona un plataforma central que contiene algunas funcionalidades típicas para la integración, análisis y visualización de datos.

Esta plataforma integra multitud de estándares para la importación/exportación de datos como son SIF (Simple Interaction Format), GML, XGMML, BioPAX, PSI-MI, GraphML, KGML (KEGG XML), SBML, OBO, and Gene Association. Además, permite importar datos en formato CSV, TSV o xlsx y añadirlos como atributos a los nodos, aristas o las propias redes. Cytoscape funciona como cliente a algunas de las bases de datos mencionadas anteriormente como por ejemplo Pathway Commons, IntAct, BioMart y NCBI Entrez Gene. Para poder manejar todos estos datos, incluye herramientas de búsqueda y filtrado para los datos así como algunas herramientas de análisis de grafos básicas.

Cytoscape es una potente herramienta de visualización capaz de manejar redes de más de 100000 nodos y aristas gracias un eficiente motor de renderizado. Cytoscape incluye diversos diseños, así como herramientas para navegar por una red, organizar varias redes, tener varias vistas de la misma red, etc. Cytoscape además permite crear y compartir estilos visuales que para poder conectar la red y como esta se ve con sus datos y atributos.

Además, Cytoscape también puede funcionar en modo "headless", es decir, sin interfaz grafica y ser accedida programáticamente a través de la línea de comandos o de su interfaz web RESTful [173]. Esto favorece la integración de Cytoscape en procesos de análisis más complejos que integren varías herramientas y estén altamente automatizados, y ofrece la posibilidad de ofrecer nuestras herramientas, análisis o resultados de forma online, gracias a herramientas como Cytoscape.js [174].

Apps Come se ha visto, el núcleo de Cytoscape es potente y muy flexible pero limitado en cuanto a su capacidad de análisis. Para ello, Cytoscape confía en su rico ecosistema de Apps (inicialmente llamadas Plugins [46]).

Una App es un componente de software que puede ser desarrollado por cualquier persona y se ejecuta en el contexto de Cytoscape. Todas las App tienen acceso inmediato a todas las funcionalidades centrales de Cytoscape, las cuales pueden extender de forma casi ilimitada. De esta forma, investigadores y desarrolladores pueden ofrecer algoritmos y herramientas de análisis ofreciendo visualizaciones, manejo de datos e interacciones de forma familiar y amigable. Además, las distintas Apps pueden depender la una de la otra para ofrecer mayores capacidades de análisis y reducir las redundancias. Estas Apps son validadas por el equipo que mantiene Cytoscape y publicadas para el resto de los usuarios a través de la App Store de Cytoscape [175].

A día de hoy existen más de 300 Apps en la App Store, las cuales permiten, por ejemplo, acceder distintas fuentes de información [176], realizar análisis de enriquecimiento [177], realizar análisis de agrupamiento [178], anotar los genes de una red [179], inferir relaciones entre ellos [180], etc.

Usabilidad y accesibilidad

Con la creciente importancia y popularidad de los enfoques computacionales y basados en datos, se hace cada vez más crítico garantizar que el software sea usable y accesible, ya que estas características proporcionan la base para la reproducibilidad de la investigación biomédica publicada [10], [44].

La aplicación de factores humanos e ingeniería de usabilidad para optimizar la usabilidad de las herramientas bioinformáticas en línea permitiría a los investigadores buscar, interactuar, compartir, sintetizar, visualizar y manipular datos de manera más efectiva y eficientemente [181].

Actualmente, las herramientas de software desarrolladas en el ámbito académico suelen presentar una peor usabilidad en comparación con las desarrolladas en entornos industriales [10]. Desarrollar software siguiendo buenas prácticas conlleva tiempo y recursos. Sin embargo, cuanta más calidad tenga el software más tiempo de vida tendrás y más será utilizado [182]. Por ello, es importante que la comunidad científica mejore significativamente las herramientas que produce utilizando las prácticas probadas en usabilidad y en otras ramas de la ingeniería [183] así como siguiendo las distintas guías y recomendaciones existentes [43], [47], [48].

Herramientas como Cytoscape [45] o Bioconductor [161] representan la nueva ola de herramientas bioinformáticas basadas en software con un fuerte enfoque en la usabilidad y accesibilidad.

2.4. Conclusiones

En este capítulo se han repasado algunos conocimientos básicos de biología y se ha realizado una introducción al campo de la bioinformática, incluyendo su historia, objetivos y herramientas principales.

Este documento se centra principalmente en los análisis de regulación génica y la biología de sistemas. Aunque también se tocan aspectos de análisis de datos de expresión génica y genética de las enfermedades. Por ello, se ha hecho especial hincapié en Gene Ontology y Cytoscape, dos herramientas bioinformáticas muy relevantes para el estudio de interacciones genéticas.

Capítulo 3

Estado del arte

3.1. Introducción

Como se ha visto en el capítulo anterior, la explosión en la cantidad de información biológica, así como de la capacidad de computación disponible, que se ha producido en los últimos años ha permitido a la comunidad investigadora realizar grandes avances en diversas áreas de estudio de la bioinformática desde la simple secuenciación de ADN y los estudios de expresión genética, los cuales trabajan sobre genes individuales o como conjunto, hasta estudios más complejos, que consideran la relaciones específicas entre los distintos genes y otras entidades biológicas, como los de regulación génica.

Esto ha propiciado la aparición de una nueva rama de la biología denominada biología de sistemas, la cual considera los organismos como sistemas complejos en los que una propiedad emergente puede deberse a la interacción entre partes más simples del sistema, ya sean factores internos o externos. Para ello, la biología de sistemas biológicos busca crear modelos precisos en tiempo real de la respuesta de un sistema a estímulos tanto ambientales como internos mediante el uso de métodos reductivos en los que se reúnen grandes cantidades de datos y se analizan computacionalmente.

3.2. Conjuntos de genes

3.2.1. Técnicas de clustering

Las primeras aproximaciones para extraer conocimiento a partir de los estudios de microarrays y NGS se basaron en el agrupamiento de genes mediante técnicas de clustering. Este tipo de técnicas permite separar las instancias que componen los datos de entrada (para el caso de los datos de expresión genética, los genes) en conjuntos disjuntos, llamados clusters, de forma que los genes que se hayan situado dentro de en un cluster sean muy similares en función de las medidas establecidas por el algoritmo en cuestión, mientras que aquellos que se encuentren en clusteres distintos sean muy diferentes [184], [185].

Las técnicas de clustering son extensamente cubiertas por la literatura académica [185], [187], [188] y se pueden clasificar como:

- **Clustering por particiones -** dividen el conjunto de datos de entrada de forma que cada partición representa un grupo o cluster. Como algoritmo más representativo de este tipo de técnicas destaca el algoritmo K-medias [189], [190].
- **Clustering jerárquico -** descompone jerárquicamente el conjunto de datos de entrada de forma que la solución es representada por un dendograma [191], [192] (ver figura 3.1).

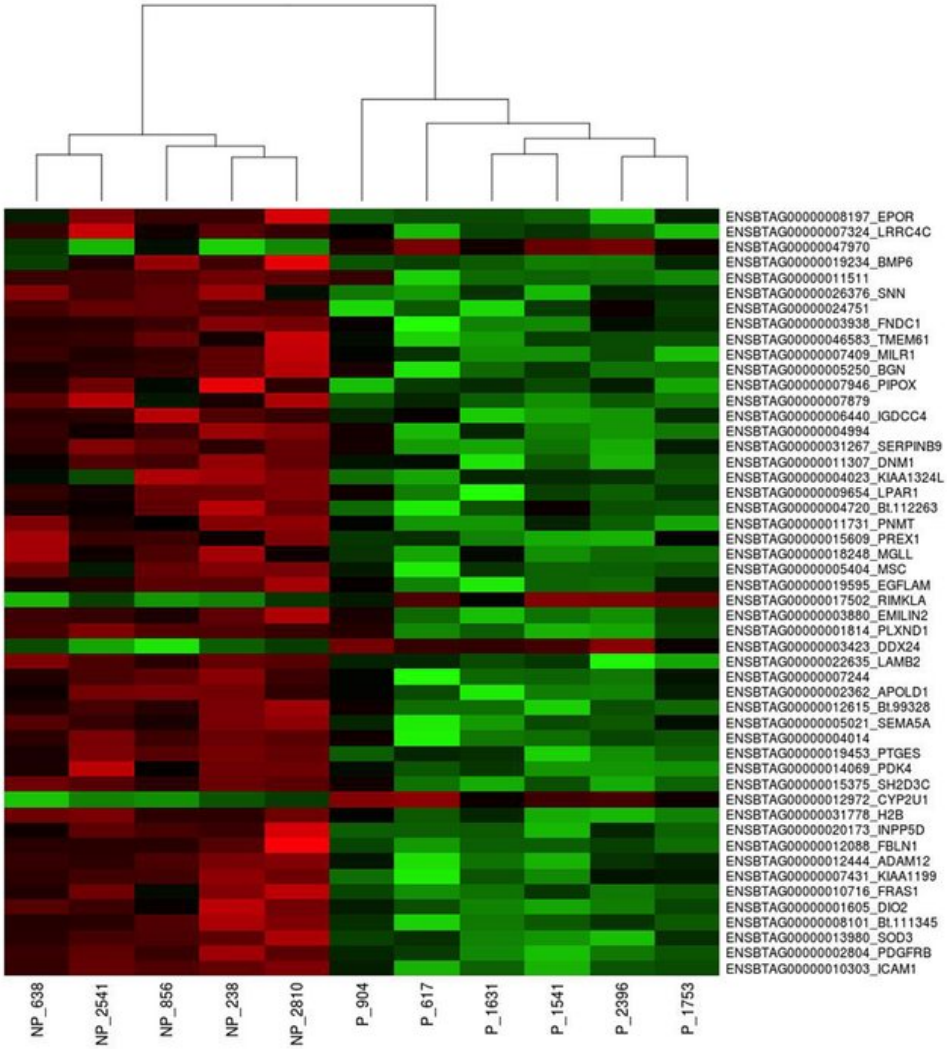


FIGURA 3.1: Ejemplo de algoritmo de clustering jerárquico presentado por Binelli et al. [186] en el que se pueden ver los diferentes niveles de expresión de un grupo de genes para vacas gestantes y vacas que no.

- **Clustering basado en mapas de modelos auto-organizanizados (SOM) -** también llamados redes de Kohonen, son un tipo de red neuronal competitiva no supervisada. Este tipo de red suele estar distribuida de forma regular en una matriz de dos dimensiones y tiene como fin el de descubrir la estructura subyacente de los datos introducidos en ella. El fin de esta matriz es ser empleada para la agrupación de las diferentes instancias en base al concepto de "puntos vecinos" que terminan siendo mapeados en un espacio k-dimensional. [193], [194]

- **Clustering basado en teorías de grafos -** representa los datos explícitamente en forma de grafo, de modo que el problema de clustering se traduce en un problema de teoría de grafos basado en encontrar el corte mínimo o "clique" (grafo conexo) máximo en el grafo de proximidad[195].

Las técnicas de clustering permiten establecer un modelo global mediante la búsqueda exclusiva y exhaustiva de grupos de genes que presenten el mismo patrón de comportamiento teniendo en cuenta todas las condiciones experimentales medidas.

Esto representa una gran limitación ya que hoy en día sabemos que grupos de genes generalmente independientes pueden ser coregulados o coexpresados bajo ciertas condiciones experimentales [196]. Otra limitación severa de las técnicas de clustering es el hecho de que cada gen debe ser asociado a un único conjunto. Sin embargo, los últimos avances científicos demuestran que un gen puede pertenecer a varios procesos biológicos [197].

3.2.2. Técnicas de biclustering

Las técnicas de biclustering son una forma de superar las limitaciones de las técnicas de clustering anteriormente mencionadas, ya que permiten realizar el agrupamiento en dos dimensiones (genes y condiciones) de manera simultanea. Este tipo de técnicas, por lo tanto, permiten general un modelo local de forma que cada gen es estudiado utilizando un subconjunto de las condiciones elegido a partir de un subconjunto de los genes (ver figura 3.2). Otra ventaja de las técnicas de bicluster es que permiten el solapamiento, es decir, que cada gen o condición bajo estudio puede pertenecer a más de un bicluster o a ninguno [198].

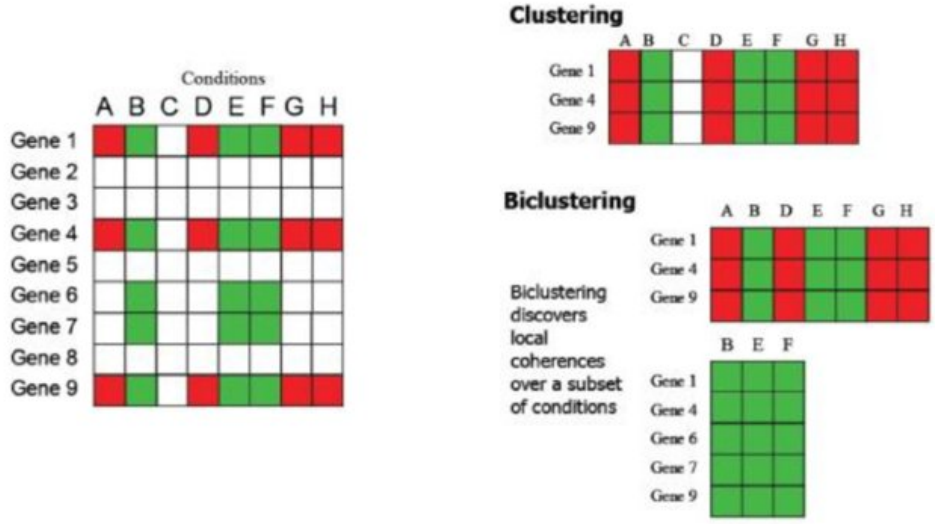


FIGURA 3.2: Ejemplo ilustrativo de como los algoritmos de biclustering son capaces de encontrar modelos locales.

Los biclusters se pueden clasificar según los patrones que presentan los genes:

- **Valores constantes -** indica subconjuntos de genes con valores de expresión similares dentro de un subconjunto de condiciones.
- **Valores constantes en filas o columnas -** indica un subconjunto de genes/condiciones con niveles de expresión similares en un subconjunto de condiciones/genes. Por lo tanto, los niveles de expresión pueden variar de un gen a otro o de una condición a otra.
- **Valores coherentes en filas y columnas -** indica relaciones más complejas entre genes y condiciones.
- **Evoluciones coherentes -** indica que un subconjunto de genes está regulado hacia arriba o hacia abajo en un subconjunto de condiciones sin tener en cuenta sus valores de expresión reales. En esta situación, los datos del bicluster no siguen ningún modelo matemático.

O en función de su estructura:

- **Exhaustivo en las filas -** cada gen pertenezca al menos a un bicluster.
- **Exhaustivo en las columnas -** cada condición pertenezca a al menos un bicluster.
- **No exhaustivo -** los genes y las condiciones podrían no asignarse a ningún bicluster.
- **Exclusivo en las filas -** cada gen puede formar parte de un bicluster como máximo.
- **Exclusivo en las columnas -** cada condición puede formar parte de un bicluster como máximo.
- **No exclusivo -** este aspecto representa la posibilidad de obtener biclusters superpuestos, es decir, varios biclusters pueden compartir genes y/o condiciones.

Por su parte, los algoritmos de biclustering pueden clasificarse de la siguiente manera:

- **Búsqueda voraz iterativa -** siguen la estrategia de hacer una elección óptima local en cada paso, con el fin de encontrar un óptimo global. Este tipo de heurística no asegura la obtención de una solución global óptima, pero la aproxima en un tiempo razonable [199]-[201].
- **Búsqueda voraz iterativa estocástica -** agregan un componente aleatorio a la búsqueda voraz iterativa, haciendo que el algoritmo sea no determinista [202]-[205].
- **Metaheurísticas inspiradas en la naturaleza -** reproducen comportamientos eficientes observados en la naturaleza como los sistemas evolutivos, sistemas inmunes artificiales, optimización de colonias de hormigas u optimización de enjambres, entre otros [206]-[210].
- **Enfoques basados en clustering -** se basan su búsqueda en el uso de un algoritmo de clustering para el agrupamiento de una dimensión tradicional, junto con una estrategia adicional que proporciona el análisis de segunda dimensión [211]-[215].
- **Enfoques basados en grafos-** utilizan la teoría de grafos utilizando los nodos para representar elementos, ya sean genes, muestras o genes y muestras, o incluso biclusters completos [216]-[219].
- **Modelos probabilísticos -** realizan un análisis estadístico para describir datos basados en el uso de la teoría de la probabilidad [220]-[223].
- **Álgebra lineal -** utilizan de espacios vectoriales y mapeos lineales entre dichos espacios para describir y encontrar las submatrices más correlacionadas del conjunto de datos de entrada.
- **Reordenamiento óptimo de filas y columnas -** realizan permutaciones de las filas y columnas originales en la matriz de datos, lo que lleva a una mejor disposición de los elementos previamente a la búsqueda [224]-[227].

Los métodos de biclustering son ideales cuando en los datos de entrada sólo un pequeño grupo de genes participa en un mismo proceso celular, una actividad celular puede producirse sólo bajo un subconjunto de condiciones experimentales o un sólo gen puede participar en múltiples procesos biológicos que no se produce en todas las condiciones experimentales estudiadas.

Por contra, al ser más flexibles, los algoritmos de biclustering pueden ser más vulnerable al sobreajuste. Además, este tipo de algoritmos son NP-duros [196] por lo que una búsqueda exhaustiva de todo el espacio de soluciones puede ser inviable. Por ello, es fundamental el uso de heurísticas y métodos de optimización que garanticen que los biclusters obtenidos son significativos.

Es importante considerar que tanto las técnicas de clustering como de biclustering pueden proporcionar conjuntos de genes relacionados y las condiciones experimentales bajo las cuales se relacionan. Sin embargo, este tipo de modelos no provee ninguna información sobre las relaciones existentes entre los genes de cada grupo entre los propios grupos. Esto hace que no sean lo suficientemente precisos para estudiar la complejidad de los sistemas biológicos.

3.3. Redes genéticas

Las redes de genes, son una forma directa de representar conjuntos de genes incluyendo sus interacciones [228], [229] por lo que son más precisas que los conjuntos a la hora de estudiar la complejidad de los sistemas biológicos. Este tipo de modelo se presenta como una estructura de red donde cada nodo representa un gen o producto genético (proteína) mientras que cada arista denota la relación entre los nodos a sus extremos. De esta forma, una red genética modela la influencia que un gen concreto puede tener sobre el resto y como es afectados por los demás.

Existen multitud de arquitecturas de redes genéticas y lo que cada arista de la red representa depende en gran medida del tipo de red y del algoritmo de inferencia utilizado. De esta forma, las redes genéticas son lo suficientemente flexibles como para modelar los procesos biológicos a distintos niveles de abstracción o poniendo el foco en un aspecto concreto de dicho proceso.

Al ser una representación abstracta, una red genética no tiene una interpretación semántica que permita extraer comportamiento de la red en si. Sin embargo, la estructura o topología de la red provee un modelo diagramático sencillo de visualizar y de razonar, el cual suele ser informativo por sí solo y reduce la complejidad subyacente en los datos (ver figura 3.3).

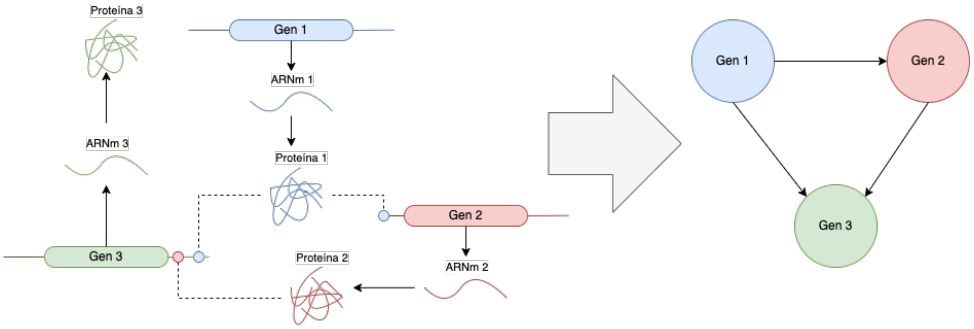


FIGURA 3.3: Ejemplo de como una red genética permite abstraer a un modelo más sencillo y diagramático sistemas biológicos complejos como la regulación genética.

Una de las cualidades más importantes a este respecto es el grado de un nodo, es decir, el número de aristas que están conectadas al nodo, así como la distribución de grados de la red, es decir, la distribución empírica de grados en todos los nodos de la red. Las distribuciones de grados a menudo codifican propiedades de las redes que se pueden interpretar intuitivamente, como la presencia de hubs o la capacidad de llegar rápidamente a cualquier nodo desde cualquier nodo inicial, y en muchos casos pueden estar relacionadas con distintos mecanismos estocásticos mediante los cuales puede surgir la red. En el caso de las redes dirigidas, se puede distinguir además entre el grado de entrada, el número de aristas que terminan en un nodo, y el grado de salida, el número de aristas comenzando en un nodo.

Las redes genéticas son conocidas por seguir una arquitectura scale-free. Esto significa que no hay un número típico de conexiones por nodo; más bien, la distribución del número de conexiones (k) por nodo (N) sigue una ley de potencia ($N(k) \sim k^{-}\gamma$). En otras palabras, hay muchos nodos con pocas conexiones y un número pequeño pero significativo de nodos con muchas interacciones. Además, estas redes tienen una arquitectura small-world, lo cual implica que cuando un nodo está conectado a otros dos nodos, estos dos últimos también tienden a tener una conexión directa entre sí. Por otro lado, la longitud promedio de la ruta más corta en la red (L, el número mínimo de conexiones que uno necesita para obtener de un nodo a cualquier otro nodo) es casi tan baja como la de las redes aleatorias [230]. Las arquitecturas scale-free y small-world son comunes de las redes intracelulares en las que los nodos están conectados cuando están involucrados en el mismo proceso biológico. Mientras que este tipo de redes son poco habituales en general [231]. Estas dos características de las redes genéticas llevan a la hipótesis de centralidad-letalidad, que establece que nodos con un grado mayor tiene más posibilidades de producir un fenotipo letal si son alterados comparados con el resto de los nodos [232].

Por otro lado, se pueden considerar redes ponderadas, donde cada arista esté asociada con un número, su peso, que defina la importancia de dicha relación, por ejemplo, cuantificando el soporte que ofrecen los datos para que exista. Las redes ponderadas a menudo se visualizan como redes con bordes de diferente grosor, que retienen la inmediatez visual de la abstracción de la red, pero transmiten de manera efectiva más información. En la Figura 3.4 se muestra un ejemplo esquemático de una representación gráfica estándar para redes dirigidas, no dirigidas y ponderadas.

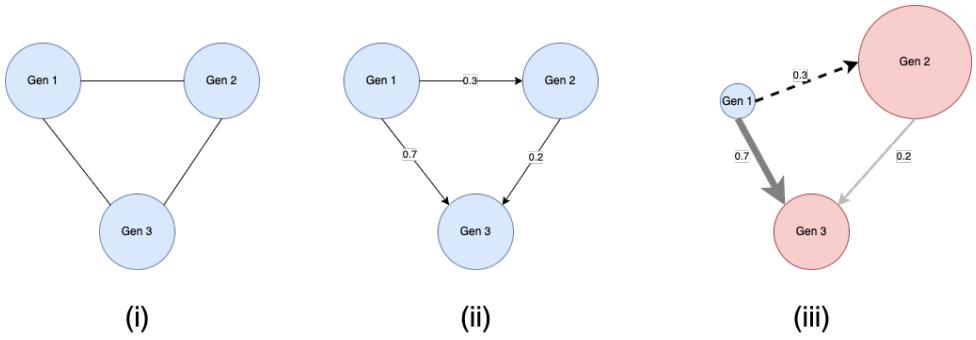


FIGURA 3.4: Ejemplos de tipos de redes genéticas: (i) de asociación, (ii) dirigida y ponderada, y (iii) dirigida, ponderada y enriquecida con otro atributos. Por ejemplo, el tamaño del nodo puede representar su nivel de enriquecimiento, el color del nodo indica clústeres, el grosor de las aristas la importancia de la relación, el color de las aristas la evidencia científica, y el tipo de línea el tipo de relación.

Al ser las redes genéticas un simple grafo, se pueden analizar matemáticamente mediante el uso de teoría de grafos. Utilizando métricas como grado, coeficiente de agrupamiento, caminos más cortos, centralidad, densidad. Podemos identificar elementos clave (hubs) y subredes relevantes, que pueden ayudarnos a dilucidar los mecanismos de interacción. Por ello, cada vez más, las redes genéticas están avanzando de ser una herramienta de visualización y didáctica a ser una herramienta de análisis y formulación de hipótesis.

Las redes genéticas han sido probadas como una herramienta muy útil para la adquisición de conocimiento biológico. Se utilizan para la predicción de la función génica ya que examinar genes (proteínas) en un contexto de red muestra conexiones con conjuntos de genes/proteínas involucrados en el mismo proceso biológico que probablemente funcionen en ese proceso. También, se utilizan para la detección de complejos de proteínas y otras estructuras modulares ya que, aunque las redes de interacción se basan en interacciones por pares, existe una clara evidencia de modularidad y organización de orden superior (motivos, ciclos de retroalimentación). Por último, las redes genéticas también se utilizan para la predicción de nuevas interacciones y asociaciones funcionales utilizando distintos métodos de inferencia de redes y aprendizaje automático.

Además, las redes genéticas se utilizan también en el estudio de enfermedades y en el desarrollo de fármacos. Dada una red genética, es posible identificar subredes que son transcripcionalmente activas si hay una enfermedad. Esto permite identificar componentes clave en rutas metabólicas durante la progresión de la enfermedad y proporciona pistas para estudios adicionales y posibles dianas terapéuticas. Las subredes también proporcionan una rica fuente de biomarcadores para la clasificación de enfermedades, basada en perfiles de ARNm integrados con redes de proteínas para identificar biomarcadores de subredes (genes interconectados cuyos niveles de expresión agregados predicen el estado de la enfermedad). Además, las redes moleculares proporcionan un marco poderoso para mapear los mecanismos de rutas metabólicas comunes afectados por distintos genotipos.

3.3.1. Inferencia de redes genéticas

Como se ha mencionado en la sección anterior, una red genética puede representar distintos modelos. Esto se conoce como la arquitectura del modelo, viene dada por el método de inferencia utilizado y define el significado y los atributos de las relaciones entre nodos de la red.

Existen multitud de técnicas experimentales y computacionales para la inferencia de redes genéticas [6], [7], [233]-[237]. A continuación, se clasifican y resumen los distintos tipos.

Modelos basados en correlación

El método más simple para inferir redes genéticas consiste en considerar los genes como red completamente conectada a partir de los genes que se desea estudiar, asignar un peso a cada relación a partir de la información de partida y filtrar las aristas con un peso que no pase de un umbral establecido. De forma que, cuanto más alto sea el umbral más dispersa será la red resultante [238].

La principal medida para determinar las dependencias entre genes es el coeficiente de correlación, en especial los coeficientes de Pearson, Spearman o Kendall.

La suposición de que los genes que interactúan deberían tener una expresión correlacionada es biológicamente plausible, y métodos como el WGCNA (análisis de

redes de coexpresión de genes ponderados [16]) han demostrado ser consistentemente fiables y ampliamente adoptados.

Estas redes, también llamadas redes de coexpresión, son computacionalmente simples de inferir ya que escalan con el posible número de relaciones que es el cuadrado del numero de genes. De esta forma permiten inferir redes genéticas de grandes dimensiones. Por contra, no consideran que dos genes puedan parecer correlacionados no porque interactúen entre si, sino por el efecto de un tercer gen (o varios). Por ejemplo, dos genes correlacionados pueden simplemente compartir el mismo gen regulador. Además, al ser redes no dirigidas solo representan la asociación entre pares de genes de forma que si el gen i regula al gen j y este regula al gen k, es posible obtener una correlación entre i y k, resultando en un falso positivo. Otro problema con este tipo de aproximación es que la red inferida no es predictiva, es decir, dados los niveles de expresión de un subconjunto de los genes, es imposible estimar los niveles de expresión de los genes restantes.

Modelos basados en la teoría de la información

La linealidad de las medidas de correlación puede limitar su idoneidad para capturar relaciones regulatorias complejas. Para evitar este problema, existen sistemas de ranking alternativos basados en la teoría de la información como es la información mutua.

Ejemplos de este tipo de aproximaciones son REVEAL (The REVerse Engineering ALgorithm) [18]; RELEVANCE [239] ARACNE (Algorithm for the Reverse engineering of Accurate Cellular Network) [240] or ARACNE-based algorithms [241], [242]; CLR (Context Likelihood of Relatedness) [243] o MRNET (based on the maximum relevance/minimum redundancy) [244].

Estas redes, también llamadas redes de información mutua, siguen siendo computacionalmente simples de inferir ya que escalan con el cuadrado del posible número de relaciones. Por lo que aun permiten inferir redes genéticas de grandes dimensiones. Por contra, continúan siendo redes no dirigidas que sólo representan la asociación entre genes y siguen siendo modelos estáticos que no tienen en cuenta que múltiples genes puedan participar en la misma regulación. Además, pueden ser muy sensibles al ruido en los datos si el tamaño de la muestra es pequeño o medio.

Modelos basados en regresión

Un enfoque alternativo para cuantificar la dependencia de dos variables consiste en predecir una de la otra. En el caso más simple, utilizando regresión lineal de forma que la pendiente de la línea de regresión cuantifique la dependencia. En el contexto de las redes genéticas, esto equivaldría a hacer una regresión de cada gen contra todos los demás genes para obtener los pesos de la red.

Ejemplos de este tipo de aproximaciones son: TIGRESS [17] y GENIE3 [245] (y sus derivados [246], [247]). Además, este enfoque se puede utilizar para analizar datos de series temporales, con la simple modificación de que la regresión se calcula entre la expresión del gen g en el tiempo t frente a la expresión de los otros genes en el punto de tiempo anterior $t-1$ (modelo autorregresivo) [248].

Los métodos basados en regresión son muy útiles para inferir redes dirigidas. Son computacionalmente más complejos que los métodos vistos hasta ahora, pero añaden capacidad predictiva, es decir, que, dados los niveles de expresión de un subconjunto de los genes, podemos calcular los niveles de expresión de los genes restantes. Además, este tipo de métodos pueden identificar relaciones condicionales

superiores entre genes y no solo relaciones por parejas. Por contra, en conjuntos de datos pequeños hay una alta probabilidad de que los genes presenten una correlación alta resultando en falsos positivos.

Modelos gráficos gaussianos

Lo modelos gráficos gaussianos (GGMs) se basan en considerar las mediciones de expresión génica como un vector aleatorio normal multivariado (cada entrada del vector representa la expresión de un gen) y estimar la matriz de precisión a partir de múltiples condiciones utilizando la estimación de máxima verosimilitud [249]. Dado que esto requiere estimar un número de parámetros que es proporcional al cuadrado del número de genes, se necesitan técnicas de regularización.

Este tipo de modelos han sido usado con éxito para la inferencia de redes genéticas [250], [251]. Sin embargo, la estimación de una matriz de precisión de alta dimensión a partir de los datos es difícil y la precisión de la reconstrucción para muestras finitas es más difícil de cuantificar a priori. Además, los modelos gráficos gaussianos asumen la normalidad de los datos, lo que implica linealidad en la relación entre los diversos genes.

Modelos basados en redes booleanas

Las redes booleanas (BNs) fueron el primer modelo utilizado para inferir redes genéticas y se basa en la premisa de que un gen actúa como un interruptor y siempre está o activado o reprimido. Los datos de expresión se discretizan en valores binarios mediante técnicas de clustering y umbrales de corte: 0 para genes reprimidos y 1 para genes activados. Mediante el uso de operadores lógicos (and, or, not) a cada gen se le asigna una función que permite calcular la relaciones del nodo con los demás nodos de la red en un punto dado del tiempo.

Este tipo de modelos han sido utilizados con éxito por diversos autores [252], [253], son simples de interpretar y son capaces de describir fenómenos biológicos tales como oscilaciones, eventos multiestacionarios, correlaciones de largo alcance, así como estabilidad e histéresis de comportamientos similar a un interruptor [234]. Sin embargo, la expresión génica rara vez es una cuestión de activación o silenciamiento totales por lo que es posible que se pierdan detalles importantes del comportamiento del sistema en estados intermedios. Las redes booleanas también deben hacer frente a problemas de datos ruidosos [254], ya que la precisión del umbral determina la topología de la red [255]. Además, las redes booleanas son deterministas y la única incertidumbre que puede existir es el estado inicial. Esto se representa mediante la distribución probabilística de todos los genes.

Modelos basados en redes booleanas probabilísticas

Para tener un modelo probabilístico capaz de capturar la incertidumbre y el dinamismo existente en los sistemas biológicos, es necesario considerar las probabilidades conjuntas de todas las funciones booleanas correspondiente a todos los nodos. De esta manera, en las redes booleanas probabilísticas (PBNs), para cada momento en el tiempo, el estado de un gen es dado por una de las posibles funciones asociadas a él elegida en función de su probabilidad [256].

Este tipo de modelos mejoran las redes booleanas normales en cuanto a que pueden modelar los comportamientos dinámicos presentes en los sistemas biológicos. Por contra, incrementa la complejidad computacional de éstas.

Modelos basados en redes bayesianas

Hasta ahora, todos los métodos presentados parten de una red completamente conectada, puntúan cada par de genes (o estiman conjuntamente una matriz de precisión en el caso de los modelos gráficos gaussianos), y establecen un umbral para obtener una estructura de red dispersa. Las redes bayesianas hacen todo lo contrario, construyendo un modelo probabilístico conjunto a partir de términos condicionales locales.

En una red bayesiana, se asume que los datos de expresión génica se pueden expresar como variables aleatorias que siguen una distribución probabilística y se representan las relaciones entre los genes como relaciones probabilísticas. De esta forma, el ruido y la incertidumbre se consideran características inherentes al proceso de regulación. Además, este tipo de redes permiten integrar conocimiento previo y otros datos en el proceso de inferencia [19].

En Larjo et al. [257], se detallan los pasos necesarios para el aprendizaje de las redes bayesianas:

1. Selección del modelo, definición de grafos acíclicos no dirigidos (DAGs) como redes de relaciones candidatas

2. Ajuste de parámetros: búsqueda de las mejores probabilidades condicionales para cada nodo dadas las redes y conjuntos de datos experimentales

3. Calificación de aptitud: puntuación de cada modelo candidato de modo que cuanto mayor sea la puntuación, mejor se ajustará el modelo a los datos de forma que el modelo con la puntuación más alta representa la red inferida

Como se observa, el paso más crítico es la "selección del modelo". El enfoque naif consiste en simplemente enumerar todos los DAG posibles para el número dado de nodos (búsqueda por fuerza bruta). Esta aproximación no es viable, sin embargo, ya que la cantidad de DAG crece de manera súper exponencial al número de nodos. Por lo tanto, se necesitan heurísticas para poder utilizar este tipo de métodos de manera eficiente.

Las redes bayesianas se han usado con éxito para identificar mecanismos de control conocidos así como para identificar nuevos biomarcadores y otros conocimientos noveles [258]-[260]. Este enfoque tiene ventajas considerables en la facilidad con la que se puede codificar la información previa y en la forma en que se representa la incertidumbre intrínseca en el sistema: típicamente, tales métodos devuelven un conjunto de estructuras de red plausibles, ponderadas por su probabilidad posterior. Sin embargo, es computacionalmente complejo y, a pesar de los avances recientes [261], la escalabilidad de este tipo de métodos para grandes conjuntos de datos sigue siendo un desafío. Además, un requisito fundamental en la estructura de una red bayesiana es la ausencia de bucles (condición DAG). Esto es una gran limitación considerando que los sistemas biológicos a menudo exhiben ciclos de retroalimentación como un mecanismo de robustez y estabilidad.

Modelos basados redes bayesianas dinámicas

Las redes dinámicas bayesianas (DBNs) expanden el conjunto de variables aleatorias en consideración, de modo que los nodos de la red representen la expresión de genes en un momento específico. Las aristas de la red solo pueden conectar nodos pertenecientes a diferentes puntos en el tiempo, de modo que un gen solo puede influir en la expresión de otro gen (o incluso de si mismo) en un momento posterior.

De esta manera, la condición DAG se satisface automáticamente, mientras que al mismo tiempo se pueden incorporar fácilmente características biológicamente plausibles como los mecanismos de retroalimentación.

Las DBNs han sido ampliamente utilizadas [262]-[264]. Este enfoque simplifica la inferencia con respecto a las redes bayesianas estándar ya que la condición DAG se satisface automáticamente. A pesar de ello, sigue siendo computacionalmente complejo. Además, la mayoría de las técnicas basadas en DBNs asumen un modelo dinámico lineal y un modelo en el que cada punto temporal en el modelo corresponde a un tiempo de observación. Existen extensiones que incluyen asignaciones no lineales entre puntos de tiempo [265], [266] o que relajan el supuesto de homogeneidad de tiempo [267], sin embargo, estas incurren en costos computacionales aun más altos y/o imponen restricciones en la clase de funciones no lineales permitidas.

Modelos basados ecuaciones diferenciales ordinarias

Las ecuaciones diferenciales ordinarias (ODE) utilizan variables continuas en lugar de discretas. Esto permite crear un modelo más preciso que incluye el modelado dinámico de la regulación genética. Las ecuaciones diferenciales representan cambios en la expresión génica como una función sobre la expresión de otros genes, y tiene en cuenta los factores ambientales. Esto permite realizar un modelado cuantitativo más cercano al comportamiento real del sistema biológico [20], [21], [234]. Emplear una semántica de tiempo continuo tiene la ventaja adicional de limitar la influencia de las decisiones de diseño experimental como la elección de puntos de tiempo/frecuencias de muestreo en el resultado final.

Los métodos basados en ODE han dado grandes resultados para la inferencia de redes genéticas [246], [268]. A pesar de estas ventajas, los modelos basados en ODE solo consideran funciones lineales o tipos específicos de funciones no lineales [235], [269], mientras que los procesos regulatorios a menudo se caracterizan por dinámicas complejas y no lineales. Además, los modelos ODE sufren de los mismos problemas de escalabilidad debido a su complejidad computacional que las técnicas mostradas anteriormente.

Modelos basados redes neuronales

Otro modelo, especialmente bueno para modelar las complejidades de los sistemas biológicos son las redes neuronales [270]. Inspiradas por las redes neuronales de los organismos vivos, este tipo de modelo se compone de una serie unidades (las neuronas) conectadas entre si para transmitirse señales. Cada neurona ejecuta una función sobre su entrada y su resultado puede ser ponderado para incrementar o inhibir el estado de activación de las neuronas adyacentes.

Este tipo de modelo aprende automáticamente en una fase de entrenamiento en la cual los pesos de las distintas neuronas se van modificando hasta encontrar la configuración óptima y una vez entrenada la red cualquier información de entrada atraviesa la red produciendo unos valores de salida.

Existen multitud de tipos de redes neuronales como son las redes neuronales feedforward (FNN), redes neuronales recurrentes (RNN) o redes neuronales con convolución (CNN). Además, existen redes neuronales especialmente diseñadas para tratar con grafos (o redes) como las redes neuronales de grafos (GNN) o redes neuronales de grafos con convolución (GCNN). Este tipo de redes neuronales operan directamente sobre la estructura del grafo.

Las redes neuronales han sido utilizadas con éxito para el estudio de interacciones entre genes y la identificación de biomarcadores llegando a mejorar muchas de las medidas existentes [271]-[273].

Este tipo de redes permiten reconocer automáticamente patrones en los datos y modelar cualquier tipo de relación por compleja que sea incluyendo relaciones no lineares y comportamientos dinámicos. Por ello son especialmente adecuadas para problemas difíciles de resolver mediante programación convencional. Por contra, son computacionalmente muy complejas y difíciles de entrenar [274].

Modelos multi-redes

Todos los métodos vistos hasta ahora se basan en la premisa de que todos los datos se pueden modelar en una sola red genética. Esto puede ser razonable cuando los datos provienen de un contexto común bajo condiciones similares. Sin embargo, esto es una suposición muy restrictiva cuando se intenta modelar conjuntamente datos de escenarios heterogéneos, ya que diferentes condiciones biológicas pueden llevar a que se activen diferentes estados. En este caso utilizar múltiples estructuras de red pueden ser más apropiadas.

Existen escenarios en los que los datos (por ejemplo, series temporales) están disponibles en condiciones diferentes pero relacionadas. Por lo tanto, se pueden asumir algunos puntos en común entre las estructuras de red subyacentes, por lo que se necesitan métodos que puedan transferir información a través de las condiciones. Esta transferencia se puede lograr mediante la introducción de una penalización por diversidad compartida dentro de diferentes problemas de optimización [22], [23]. De manera equivalente pero más flexible, la reconstrucción conjunta de las diferentes redes se puede lograr adoptando un enfoque bayesiano jerárquico [19], [275].

Por otro lado, existe la idea de redes que varían en el tiempo de forma que la estructura de la red en sí puede reconectarse a lo largo del tiempo, por ejemplo, para tener en cuenta los puntos de control durante el desarrollo o la evolución del cáncer. La solución generalmente se compone de dos pasos: la identificación de los puntos de cambio y un aprendizaje conjunto de redes relacionadas en los tramos homogéneos de la serie temporal. Esta idea ha sido explorada tanto en el contexto de los enfoques de optimización [276], [277] como bayesianos [278], [279].

Inferencia de sin modelo (model-free)

Todos los enfoques vistos hasta ahora requieren conocer a priori un modelo de la dinámica del sistema (a menudo de alta dimensionalidad). Sin embargo, es posible inferir interacciones directas únicamente a partir del análisis de la dinámica colectiva no lineal. Este tipo de enfoque no requiere, en general, ninguna restricción o conocimiento previo sobre la estructura de la red de relaciones, ni hace supuestos relacionados con los principios fisicoquímicos que gobiernan las interacciones entre genes. Estos métodos solo necesitan la información de expresión génica como fuente de datos para el proceso de inferencia.

Un ejemplo de este tipo de métodos son las técnicas de extracción de reglas de asociación (AR). Una AR establece un vínculo causal entre dos o más variables, donde la semántica y la interpretación de la regla dependen de los datos de entrada y de los mecanismos empleados para inferir la asociación. De esta forma, una red genética se puede representar simplemente como un conjunto de ARs. Las ARs se han utilizado ampliamente para descubrir interesantes relaciones entre variables en

grandes conjuntos de datos [280] y, en bioinformática, este tipo de métodos son utilizados para revelar asociaciones biológicamente relevantes entre genes, en diversos entornos condiciones u observaciones puntuales, a partir de muestras de microarrays [281]-[283].

Los enfoques sin modelo presentan la ventaja de ser capaces de proporcionar predicciones precisas incluso si solo se registra una fracción de la red.

3.3.2. Validación de redes genéticas

Las redes genéticas inferidas o reconstruidas a partir de datos experimentales deben ser validadas [284], [285] para confirmar su calidad y fiabilidad. Según Qian y Dougherty [286] existen dos problemas a la hora de validad una red genética:

1. Dada una red de genes inferida, ¿provee buenas predicciones con respecto a los hechos observados experimentalmente?

2. Dado un algoritmo de inferencia, ¿produce redes que son correctas de acuerdo con algún criterio de bondad?

El primer problema se refiere a la validación científica de la calidad y fiabilidad de la red [285], [287] ya que, desde el punto de vista del conocimiento biológico, sin ser capaces de responder la pregunta ¿en qué medida un modelo de red está de acuerdo con los fenómenos observados?, la red carece de validez científica.

El segundo problema se refiere a la validación del algoritmo en sí y generalmente requiere de un conocimiento al menos parcial acerca de las verdaderas interacciones, el cual es generalmente incompleto, incierto o difícil de obtener en la práctica (especialmente en el campo del modelado una red de genes).

En la práctica, sin embargo, la línea que separa a ambos problemas se diluye ya que la validación de un método de inferencia requiere la validación de los modelos inferidos y la validación de los modelos sirve para mejorar el método de inferencia [286].

Validación experimental

El enfoque más obvio para la validación de la red es confiar en la validación experimental para corroborar los modelos inferidos [25], [288], [289].

Este tipo de experimentos de laboratorio (“wet lab”) generalmente requieren mucho tiempo y son costosos, especialmente si se tiene en cuenta que las redes inferidas pueden ser de gran tamaño [290]. Además, en muchos casos, la validación experimental sin ambigüedades no es posible para todas las redes inferidas.

Validación mediante datos sintéticos

Una alternativa para dichas dificultades a la hora de encontrar datos es utilizar datos sintéticos [291], [292]. En esté tipo de aproximación, datos sintéticos son generados representado la red y los datos de expresión génica. De esta forma, se puede medir la eficacia de un algoritmo de inferencia.

El principal inconveniente de este tipo de validación es que no permite validar una red por sí misma, sino que valida el algoritmo de inferencia en sí.

Validación interna/estadística

La validación interna hace uso de la propia información empleada en la generación del modelo para llevar a cabo la validación del mismo. En estadística, existen diferentes técnicas de muestreo para evaluar el rendimiento de generalización o robustez de un modelo como son el submuestreo, validación cruzada, bootstrapping o perturbación. Las técnicas de submuestreo, como la validación cruzada y el bootstrapping, se basan en la división de los datos disponibles en los conjuntos de datos de entrenamiento y prueba. En validación cruzada con k muestras, el conjunto de datos se divide en k submuestras. Una sola submuestra se mantiene como el conjunto de datos de prueba, y los $k-1$ grupos restantes se utilizan para el entrenamiento.

Este tipo de técnicas no son muy adecuadas para trabajar con redes obtenidas mediante datos de series temporales, ya que, la división de dichos datos de entrada afecta al proceso de inferencia haciendo que éste pierda mucha información relevante. Además, este tipo de medidas presentan una evaluación demasiado optimista que no se puede generalizar al contextos biológicos nuevos [26].

Análisis basados en topología de la red

Como se vio en el la sección 3.3, generalmente, las redes biológicas poseen determinadas características topológicas. Concretamente, se ha demostrado que la mayoría de ellas siguen una topología scale–free donde pocos nodos se encuentran altamente conectados (hubs), mientras que el resto presenta una conectividad más limitada. Trabajos recientes han considerado la importancia de que las redes inferidas posean topología de red biológica [293]. Por este motivo, se incorporan estudios topológicos de las redes para comprobar la utilidad biológica de los resultados, como por ejemplo los trabajos [27], [28].

Este tipo de validación es simple pero no tiene en cuenta la relevancia biológica de la red, sino que solo considera su estructura.

Comparación directa

Para superar estos problemas, la comunidad científica ha desarrollado métodos computacionales para validar redes genéticas en base al conocimiento existente.

Una primera aproximación consiste en la comparación directa de la red inferida mediante con una red conocida o con repositorios de interacción gen-gen utilizados como red de referencia (gold-standard). Los resultados de este tipo de validación son relativos a alguna característica de red y cuantificados según la distancia entre la característica de la red inferida y la característica de la red de referencia [8], [9], [29], [294], [295]. Para ello, se basan en técnicas de simulación estadística combinadas con conceptos como tasa de verdaderos positivos (TPR; tasa de aristas verdaderas que son detectados por el algoritmo), tasa de detección falsa (FDR; tasa de aristas detectadas que no existen en el gold-standard) y algoritmos como q-power (probabilidad de detectar al menos un porcentaje de aristas en el gold-standard) [296].

Ya que estos métodos se basan en la comparación de la topología de ambas redes, requieren un gold-standard que tenga forma de red y sea totalmente fiable y adecuado para la red que se desea evaluar. Por ejemplo, una red de coexpresión no puede utilizarse para validar una red de similitud proteína-proteína [286]. Un esfuerzo considerable ha sido dedicado a crear redes de referencia y bases de datos [140], [297]-[299], así como distintos métodos para la generación de redes a partir de dichas bases de datos [300]. A pesar de ello, no siempre es posible encontrar una red

de referencia adecuada ya que no solo los datos de diferentes tecnologías no se superponen significativamente, sino que también los datos de diferentes laboratorios que utilizan la misma tecnología difieren sustancialmente. Esto sugiere que los datos actuales están lejos de saturarse y los datos de diferentes recursos son complementarios entre sí [301]. El problema es aun más notable cuando se quiere validar una red de-novo y las relaciones entre sus genes han sido poco estudiadas [302]. Esto, supone una gran limitación a la hora de validar interacciones noveles que aun no estén presentes en los modelos de referencia.

Es importante tener en cuenta que este tipo de validación se centra exclusivamente en la red en sí misma, y no tiene en cuenta el contexto en el que se está estudiando la red. Aunque el éxito de la red inferida dependerá en cierta medida de la cercanía entre ésta y la red de referencia, tener en cuenta el contexto del estudio permite una validación más precisa, especialmente si la red de referencia está sesgada [303].

Por otro lado, este tipo de medidas asume que las topologías de las redes son estáticas y no cambian con el tiempo o en respuesta a perturbaciones externas, lo cual rara vez se cumple en el caso de las redes de regulación génica [24], [283].

Medidas basadas en anotaciones

Para solventar los problemas derivados de la comparación directa entre redes, se han desarrollado métodos basados en anotaciones sobre entidades biológicas. Este tipo de aproximaciones no requiere que el conocimiento existente tenga forma de red de genes, sino que se basa en anotaciones que a menudo se almacenan de forma controlada a través de terminologías u ontologías [30], [31]. En biología, la ontología más ampliamente utilizada es la Ontología de los genes (GO) [123] que es un vocabulario controlado de los genes y los productos génicos de multitud de especies (ver sección 2.3).

Métodos de enriquecimiento Una aproximación muy popular para la validación de redes genéticas son los análisis de enriquecimiento [32], [33]. Este tipo de aproximación trata de encontrar grupos de genes que se encuentran estadísticamente sobrerrepresentados (enriquecidos) o subrepresentados (reducidos) con respecto a una distribución predefinida. En el contexto de la validación de redes genéticas, la distribución de las anotaciones funcionales provenientes de varias bases de datos biológicas como [146], [133] o [123] y se considera que la sobre- o subrepresentación de un conjunto de genes puede estar relacionado con diferencias fenotípicas.

Existe una gran variedad de métodos de enriquecimiento, los cuales generalmente se diferencian por el tipo de prueba estadística aplicada, siendo la más común la prueba exacta de Fisher o la prueba hipergeométrica. Además, los métodos también pueden varíar en su aporte: algunos toman conjuntos de genes sin clasificar, otros clasifican conjuntos de genes, con métodos más sofisticados que permiten que cada gen se asocie con una magnitud (por ejemplo, nivel de expresión), evitando los límites arbitrarios.

Aunque este enfoque ha sido ampliamente utilizado para analizar la importancia de las redes genéticas [304], [305], es un enfoque diseñado para analizar conjuntos de genes por lo que obvia las relaciones entre los genes de la red, es decir, las aristas de la red, las cuales son la principal ventaja que éstas presentan sobre los conjuntos como modelo.

Debido a ello, se han desarrollado múltiples técnicas que combinan el análisis de enriquecimiento con otro tipo de análisis como el topológico para poder analizar redes genéticas teniendo en cuentas las relaciones entre los genes que la componen

[35], [306]-[308]. En cualquier caso, este tipo de análisis solo proporciona información sobre la distribución de anotaciones y no proporciona una medida cuantitativa.

Métodos basados en similitud semántica Para superar la carencia de una medida cuantitativa en los análisis de enriquecimiento, las medidas de similitud semántica han sido ampliamente utilizadas. Este tipo de medidas evalúa el grado de relación entre dos entidades en base a la semejanza de su significado, el cual viene dado por sus anotaciones. Todos los términos en GO se pueden representar como un grafo acíclico dirigido (DAG), de forma que las diferentes propiedades del grafo permiten medir la similitud entre dos nodos (términos) de dicho grafo. Los enfoques de similitud semántica dan como resultado una medida cuantitativa que puede utilizarse para determinar la validez de una red [34], han sido ampliamente estudiados [38], [309], [310] y multitud de herramientas existen basadas en este tipo de enfoque [37].

El primer reto de este tipo de aproximación consiste en calcular la similitud semántica entre dos términos de GO. Existen tres categorías principales las características del DAG que utiliza: nodos (información), aristas (rutas) o ambas (híbrido).

Las medidas basadas en nodos utilizan medidas basadas en el contenido de información (IC) de los nodos, el cual da una medida de cuán específico e informativo es un término. Este último tipo de medidas suele combinar la profundidad del término y el IC para evaluar la similitud de los términos. Mientras que la profundidad denota especificidad a los términos, el IC la popularidad del término (y sus descendientes) en un corpus de anotaciones. Las medidas basadas en IC más comunes han sido las de Resnik [311], Lin [312] y Jiang y Conrath [313] que se desarrollaron originalmente para WordNet y luego se aplicaron a GO [314], [315]. Resnik propuso utilizar el IC del de ancestro común más informativo (MICA) a los nodos bajo estudio para determinar su similitud. Si bien esta medida es eficaz para determinar la información compartida por dos términos, no considera qué la distancia de los términos y su MICA. Para tener en cuenta esa distancia, las medidas de Lin y Jiang y Conrath relacionan el IC del MICA con el IC de los términos que se comparan de forma que estas medidas son proporcionales a las diferencias de IC entre los términos y su antepasado común, independientemente del IC absoluto del ancestro. Para superar esta limitación, Schlicker et al. [316] propuso una medida basada en la medida de Lin, pero utilizando la probabilidad de anotación del MICA como factor de ponderación para proporcionar la ubicación del gráfico. Todas estas medidas presentan la restricción de no considerar que un término puede tener varios ancestros comunes disjuntos (DCA). Para superar esta limitación, Couto et al. [317] propuso el método GraSM, en el que el IC del MICA fue reemplazado por el IC promedio de todos los DCA. Bodenreider et al. [318] desarrolló una medida basada en nodos que también usaba datos de anotaciones, pero no se basaba en la teoría de la información.

En los enfoques basados en aristas, la puntuación de similitud semántica es una función del número de aristas (o nodos) en las rutas que conectan dos términos [319]-[322]. Normalmente se consideran cuatro factores principales en los métodos basados en la distancia de la siguiente manera:

1. densidad en el gráfico de ontología: cuanto mayor es la densidad, más cercana es la distancia entre los nodos;

2. profundidades de los nodos: cuanto más profundos son los nodos ubicados, más obvia es la diferencia entre los nodos;

3. tipos de enlaces: el tipo normal es es-una relación, y otras relaciones como parte-de y sustancia-de están asociadas con el peso de los bordes;

4. pesos de los enlaces: los bordes que conectan un cierto nodo con todos sus nodos secundarios pueden variar entre diferentes pesos semánticos.

Pekar y Staab propusieron una medida basada en la longitud del camino más largo entre el ancestro común más bajo (LCA) de dos términos y la raíz (profundidad máxima del ancestro común), y en la longitud del camino más largo entre cada uno de los términos y ese ancestro común [323]. Cheng et al. también presentaron una medida basada en la máxima profundidad del LCA, pero ponderando cada arista para reflejar la profundidad [324]. Wu et al. propusieron una medida basada en la máxima de profundidad del LCA no ponderada [325]. Wu et al. Propusieron un ajuste de esta medida, introduciendo la distancia al nodo de la hoja más cercano y la distancia al ancestro común más bajo para tener en cuenta la especificidad del término [326].

Para evitar las limitaciones de ambos tipos de aproximaciones, los métodos híbridos consideran varias características como la similitud de atributos, la jerarquía de ontologías, el contenido de información y la profundidad del LCA simultáneamente [327]-[331].

Para mejorar aún más la precisión de las medidas de similitud semántica, las investigaciones recientes sugieren mejorar las medidas de similitud semántica existentes combinándolas con otras fuentes de datos relevantes como redes de cofundación [332], [333].

Hasta ahora, se han visto técnicas para calcular la similitud semántica entre dos términos. Sin embargo, los productos génicos (proteínas) pueden estar anotados con varios términos GO. La función del producto génico a menudo se describe mediante varios términos de función molecular, y los productos génicos a menudo participan en múltiples procesos biológicos y están ubicados en varios componentes celulares. Por lo tanto, para evaluar la similitud funcional entre productos génicos es necesario comparar conjuntos de términos en lugar de términos individuales.

Para ello, las medidas por pares vistan anteriormente deben transformarse en un solo valor representativo [334]. Existen distintas técnicas que pueden ir desde el conteo de términos superpuestos [335] hasta técnicas derivadas de las vistas anteriormente pero que en lugar de considerar una pareja de términos consideran dos conjuntos, uno por cada producto génico, y combinan las posibles combinaciones de alguna manera determinada como haciendo la media o escogiendo el mejor valor.

Aunque las medidas de similitud semántica resuelven los problemas de los enfoques anteriores, al igual que sucedía con las técnicas anteriores, no consideran cómo los genes bajo estudio interactúan entre sí, que es la principal característica de las redes genéticas.

3.4. Conclusiones

En este capítulo se han revisado el estado del arte en cuanto al modelado de sistemas biológicos complejos. Se ha analizado la evolución de los estudios de conjuntos de genes hacía el estudio de redes genéticas y se ha hecho una revisión de los distintos métodos de inferencia y validación de estas últimas. En este sentido, cabe destacar que tanto la inferencia como la validación de redes genéticas siguen siendo problemas no resueltos y la elección de un método u otro dependen de múltiples factores [336].

Las propuestas presentadas en este documento construyen sobre las carencias de algunas técnicas existentes para mejorarlas, concretamente técnicas de inferencia

basadas en el aprendizaje automático y técnicas de validación basada en la comparación directa. Además, se presenta la primera metodología de validación de redes genéticas basada en el concepto de similitud semántica.

Parte III

Publicaciones

Capítulo 4

Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data

Bioinformatics, 32(4), 2016, 635–637
doi: 10.1093/bioinformatics/btv598
Advance Access Publication Date: 21 October 2015
Applications Note

OXFORD

Data and text mining

Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data

Juan J. Diaz-Montana[1,†], Owen J.L. Rackham[2,†], Norberto Diaz-Diaz[1] and Enrico Petretto[2,*]

[1]School of Engineering, Pablo de Olavide University, Seville, 41013 Spain and [2]Duke-NUS Graduate Medical School Singapore, Singapore, 169857 Singapore

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Igor Jurisica

Received on July 9, 2015; revised on October 2, 2015; accepted on October 9, 2015

Abstract

Summary: As the volume of patient-specific genome sequences increases the focus of biomedical research is switching from the detection of disease-mutations to their interpretation. To this end a number of techniques have been developed that use mutation data collected within a population to predict whether individual genes are likely to be disease-causing or not. As both sequence data and associated analysis tools proliferate, it becomes increasingly difficult for the community to make sense of these data and their implications. Moreover, no single analysis tool is likely to capture all relevant genomic features that contribute to the gene's pathogenicity. Here, we introduce Web-based Gene Pathogenicity Analysis (WGPA), a web-based tool to analyze genes impacted by mutations and rank them through the integration of existing prioritization tools, which assess different aspects of gene pathogenicity using population-level sequence data. Additionally, to explore the polygenic contribution of mutations to disease, WGPA implements gene set enrichment analysis to prioritize disease-causing genes and gene interaction networks, therefore providing a comprehensive annotation of personal genomes data in disease.
Availability and implementation: wgpa.systems-genetics.net
Contact: enrico.petretto@duke-nus.edu.sg
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Motivation

With the growing volume of patient-specific sequences that is being generated there is an increasing need to annotate these data and distinguish possible disease causing mutations from benign mutations. To this end, a number of approaches have been developed to prioritize genes based on their predicted pathogenicity using whole-exome and whole-genome data. A recently introduced class of approaches use the pattern of functional sequence variation (i.e. rare and common mutations) observed in the human population (Petrovski *et al.*, 2013), the likelihood of observed mutations according to evolution (Rackham *et al.*, 2014) or statistical modelling of genes under selective constraint (Samocha *et al.*, 2014) to prioritize (rank) disease-causing genes from sets of genes impacted by mutations. Differently from sequence variant-level analysis (e.g. PolyPhen2 (Adzhubei *et al.*, 2013)), these methods specifically allow a *gene-level analysis* of pathogenicity, providing elegant, yet distinct schemes to evaluate the significance for individual genes in disease (Enns *et al.*, 2014; Shashi *et al.*, 2014). Here we provide an easy to use web-based tool (Web-based Gene Pathogenicity Analysis or WGPA) that integrates these methods for *gene-level* pathogenicity analysis (Petrovski *et al.*, 2013; Rackham *et al.*, 2014; Samocha *et al.*, 2014) as well as any future scoring system, therefore facilitating the assessment of the evidence

supporting a role for a gene or variant in disease pathogenesis. Beyond single-gene analyses, WGPA provides a means to assess and test pathogenicity (using gene set enrichment analysis (Subramanian *et al.*, 2005)) for groups of genes of interest, look for mutations in the so called hot-zone using the gene level scores in conjunction with PolyPhen-2 (Adzhubei *et al.*, 2013) or FATHMM (Shihab *et al.*, 2013) and also to incorporate information from known gene interaction networks all within the same web based framework. Our platform will allow the scientific community to critically evaluate and interpret the large sets of mutation data from sequencing studies, aiding in the identification of genes and networks that play a critical role in disease aetiology.

2 Methods and implementation

2.1 Measures of genic intolerance

To date, only a few methods to predict pathogenicity at the gene level using sequence or population information alone are available: Residual variance intolerance score (RVIS) (Petrovski *et al.*, 2013), Evolutionary intolerance score (EvoTol) (Rackham *et al.*, 2014) and gene constraint scores (GCS) (Samocha *et al.*, 2014). The combination of these techniques with other analysis tools can provide a means to assess pathogenicity for sets of genes that have been found to be mutated in a disease, such as those identified by whole-exome and whole-genome sequencing. Here we provide a web-based tool that integrates in a single framework of analysis the following genic intolerance measures:

- RVIS identifies an intolerant gene as a gene containing a higher number of rare mutations than would be expected compared to other genes with a similar number of mutations.
- EvoTol identifies an intolerant gene as a gene containing an excess of mutations that, on the protein space, are not favoured by evolution as compared with other genes with the same number of mutations.
- GCS identifies excessively constrained genes using a statistical model which allow to rank genes based on their relative deficiency of functional variation.

2.2 Gene set enrichment analysis of gene pathogenicity

The methods described above provide gene-level scores for the identification of variants and genes that have a critical role in disease; these scores can be used to create ranked gene lists where individual highly intolerant (or constrained) genes can be prioritized. In order to integrate these scores over sets of genes, we provide a gene set enrichment analysis (GSEA) implementation (Subramanian *et al.*, 2005) that can be used with RVIS, EvoTol or GCS. Briefly, given a ranked list of genes (calculated genome-wide for each method described above) the GSEA tool tests if the genic intolerance scores of a subset of genes (provided by the user) occupy higher (or lower) positions in the ranked gene list than what it would be expected by chance. Gene set enrichment scores and significance level of the enrichment (*P*-value, False Discovery Rate (FDR), FWER *P*-value) are provided, using the GSEA output format developed by Broad Institute of MIT and Harvard (Subramanian *et al.*, 2005).

2.3 Interactome data

Genes that are mutated in disease do not operate in isolation, but as part of highly complex cellular and regulatory systems. A number of sources of gene interaction data are available, and here we use the STRING database (von Mering *et al.*, 2003), which provides several types of gene-gene interaction data. In order to remove less reliable interactions, we have filtered the STRING network to include only those interactions that have a STRING confidence score greater than 500 and are experimentally supported (Rackham *et al.*, 2014). The interaction data is used to display the pathogenicity scores for a set of genes on a network which, for instance, can be used to identify genes that are both intolerant to mutation and network hubs.

2.4 Tools for annotating individual SNPs

In the development of RVIS the authors also defined the 'hot-zone' of mutation. This is a set of mutations that are both predicted to be damaging and also lie within genes that are predicted to be intolerant to mutation. In order to generalize this concept we have integrated both PolyPhen-2 and FATHMM, allowing for the hot-zone to be created as a combination these with of any of the three measures of intolerance.

2.5 Web interface

In order to facilitate the annotation of personal genomes data with respect to disease pathogenesis, we have developed a unified web-based tool for pathogenicity analysis of individual genes, gene sets and gene interacting networks. To this aim, we developed an intuitive graphical user interface that will make the available prioritization methods (RVIS, EvoTol, GCS) and integrated analysis tools (GSEA, cell-type specificity, gene interacting networks) easy to access and use by the general scientific community. The type of input data, integrative analyses components and outputs are schematically summarized in Figure 1, and include the following inputs, analyses and outputs:

- **Inputs –** *Gene-Level:* manual data entry; gene list (*.txt); GRP, gene set (*.grp); GMX, gene matrix (*.gmx); GMT, gene matrix transposed (*.gmt); WGCNA, weighted gene co-expression network analysis output (*.wgcna); *Variant-Level:* manual data entry; list of protein substitutions (*.txt); list dbSNP identifiers (*.txt); *Network-Level:* manual data entry; list of gene identifiers for STRING (*.txt); list of gene pairs (*.txt)
- **Analyses –** RVIS, EvoTol (can be stratified by gene expression), GCS (user-selected); RVIS, EvoTol, GCS combined with variant-level consequence predictions (PolyPhen2 (Adzhubei *et al.*, 2013)) or FATHMM (Shihab *et al.*, 2013)); gene set enrichment analysis (for *Gene-Level* inputs)
- **Outputs –** genes ranked by their genic intolerance or constraint scores (graphical and table formats); GSEA results for gene sets (graphical and table formats); gene pathogenicity annotation using both the predicted 'functionally damaging' mutations and genic intolerance (or constraint) scores (to identify the so-called *hot-zone*, i.e. predicted both highly-intolerant and 'functionally damaging') (graphical and table formats); gene interaction network annotated according to RVIS, EvoTol or GCS allowing

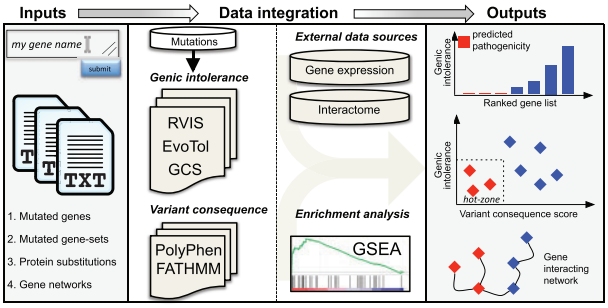


Fig. 1. Schematic representation of the inputs, integrative data analyses component and associated outputs available through WGPA

zooming out of a particular gene and visualizing its connections to other genes (graphical format).

3 Example

An example of where WGPA will be useful is to prioritize the set of genes with *de novo* mutations from trio sequencing projects. For instance in the Epi4K project Allen *et al.* (2013), trio sequencing was performed on epilepsy patients resulting in the identification of 329 *de novo* mutations impacting 176 different genes. By cross matching the RVIS, GCS and EvoTol scores and focusing on the genes from the top 25 percentile, we identify a set of 17 genes of interest (*ATP2B4, CHD4, DNM1, FLNA, FLNC, GABRA1, GABRB3, GNAO1, GRIN1, KCNQ2, MLL, MLL4, MYH6, SCN1A, SCN2A, SCN8A, WHSC1L1*, Supplementary Table S1). Using WGPA it was also possible to perform a GSEA of each of the measures of intolerance using the Epi4K mutated genes as the gene set of interest, and show that in each case the Epi4K mutated gene set is significantly enriched for predicted pathogenic genes (Supplementary Figure S1).

Funding

Supported by The Duke-NUS Graduate Medical School Signatures Research Program (Program in Cardiovascular and Metabolic Disorders).

Conflict of Interest: none declared.

References

Adzhubei,I. *et al.* (2013) *Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2*. Chapter 7. Current protocols in human genetics.

Allen,A.S. *et al.* (2013) De novo mutations in epileptic encephalopathies. *Nature*, **501**, 217–221.

Enns,G.M. *et al.* (2014) Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet. Med. Off. J. Am. Coll. Med. Genet.*, **16**, 751–758.

Petrovski,S. *et al.* (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**, e1003709.

Rackham,O.J.L. *et al.* (2014) EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res.*, **43**, e33.

Samocha,K.E. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.

Shashi,V. *et al.* (2014) The RBMX gene as a candidate for the Shashi X-linked intellectual disability syndrome. *Clin. Genet.*, **88**, 386–390.

Shihab,H.A. *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

von Mering,C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Capítulo 5

GFD-Net: A novel semantic similarity methodology for the analysis of gene networks

Journal of Biomedical Informatics 68 (2017) 71–82

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

GFD-Net: A novel semantic similarity methodology for the analysis of gene networks

Juan J. Díaz-Montaña *, Norberto Díaz-Díaz, Francisco Gómez-Vela

Intelligent Data Analysis (DATAi), Division of Computer Science, Pablo de Olavide University, ES-41013 Seville, Spain

ARTICLE INFO

Article history:
Received 8 August 2016
Revised 8 February 2017
Accepted 22 February 2017
Available online 6 March 2017

Keywords:
Gene network validation
Gene network analysis
Gene network
Semantic similarity
Gene Ontology

ABSTRACT

Since the popularization of biological network inference methods, it has become crucial to create methods to validate the resulting models.

Here we present GFD-Net, the first methodology that applies the concept of semantic similarity to gene network analysis. GFD-Net combines the concept of semantic similarity with the use of gene network topology to analyze the functional dissimilarity of gene networks based on Gene Ontology (GO). The main innovation of GFD-Net lies in the way that semantic similarity is used to analyze gene networks taking into account the network topology. GFD-Net selects a functionality for each gene (specified by a GO term), weights each edge according to the dissimilarity between the nodes at its ends and calculates a quantitative measure of the network functional dissimilarity, i.e. a quantitative value of the degree of dissimilarity between the connected genes.

The robustness of GFD-Net as a gene network validation tool was demonstrated by performing a ROC analysis on several network repositories. Furthermore, a well-known network was analyzed showing that GFD-Net can also be used to infer knowledge.

The relevance of GFD-Net becomes more evident in Section "GFD-Net applied to the study of human diseases" where an example of how GFD-Net can be applied to the study of human diseases is presented.

GFD-Net is available as an open-source Cytoscape app which offers a user-friendly interface to configure and execute the algorithm as well as the ability to visualize and interact with the results(http://apps.cytoscape.org/apps/gfdnet).

1. Introduction

During the past ten years, great scientific and technological advances in the biotechnology industry have led to the decrease of experimentation costs and the increase of available raw data. Faced with this large amount of information, biological scientists have to deal with piles of data but only flakes of knowledge [27].

Gene networks arose as a straightforward way of representing gene sets including their interactions [89,44]. They are shown as a network structure where each node represents a gene or gene product (protein) while each edge denotes the relationship between the nodes at its ends. Research has focused on inferring these networks using different experimental and computational techniques [46,13,62,31] as well as analyzing those networks to extract knowledge [66,65]. However, inferred networks must be validated in order to verify their quality and reliability [42,104].

The most obvious approach to network validation is to rely on experimental validation to corroborate the inferred models [51,110,80]. This type of wet lab experiments are typically time-consuming and expensive, especially considering that the inferred networks can be large [108]. Moreover, in many cases, unambiguous experimental validation is not possible for all inferred networks.

To overcome these issues, computational methods to validate gene networks have been developed by the scientific community. These existing methods can be divided into methods that require direct comparison between the inferred network and known networks and methods that validate the network based on existing knowledge, such as gene annotations of biological entities.

Methods based on direct comparison evaluate the inferred network by comparing it with a known network or gene–gene interaction repositories used as gold standard [81,40,96,33]. A considerable effort has been put into creating databases or datasets that can be used as gold standard [60,100,34,59] as well as developing techniques to evaluate the similarity between the inferred network and the gold standard. However, these methods are based

* Corresponding author.
E-mail addresses: jjdiamon@alumno.upo.es (J.J. Díaz-Montaña), ndiaz@upo.es (N. Díaz-Díaz), fgomez@upo.es (F. Gómez-Vela).

http://dx.doi.org/10.1016/j.jbi.2017.02.013
1532-0464/

on the comparison of the network topology which requires the gold standard to be in the form of a network and be fully reliable and suitable for the network being evaluated i.e. a co-expression network cannot be used to validate a protein similarity network [26].

Approaches that rely on gene annotations of biological entities are an alternative that does not require existing knowledge to be in the form of a gene network. Such annotations are often stored in a controlled form through terminologies or ontologies [8,97]. In biology, the ontology most widely used is the Gene Ontology (GO) [4] which is a cross-species, controlled vocabulary of genes' and gene products' attributes. GO is divided into three ontologies: (1) Cellular Component contains the parts of a cell or its extracellular environment, (2) Molecular Function contains the elemental activities of a gene product at the molecular level, such as binding or catalysis and (3) Biological Process contains the operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units (cells, tissues, organs, and organisms). GO allows the comparison of ontology terms as well as gene or gene product groups, according to their similarity based on their functional characteristics expressed by ontology terms.

Included in this category are enrichment analyses [99,50] which use statistical approaches to identify significantly over-represented (enriched) or under-represented (depleted) groups of gene annotations based on a predefined ranking of gene annotations. Although originally designed to analyze gene sets, several methods have been developed to apply enrichment analysis to gene networks [38,1,72,82]. This approach has been widely used to analyze the significance of genes networks [17,25]. However, it only provides information about the annotation distribution and does not give a quantitative measure.

To overcome this issue, semantic similarity measures were proposed and have been the focus of many studies. This type of approach assesses the degree of relatedness between two entities by the likeness of their meaning given by their annotations. All the terms in GO are typically represented as a directed acyclic graph (DAG), and the different properties of the graph are used to measure the similarity between two nodes (terms). Semantic similarity approaches result in a quantitative measure which can be used to determinate the validity of a network [10].

Several surveys have been published reviewing the existing semantic similarity measures and classifying them in different ways [69,86,35,41]. One type of semantic similarity measure uses the nodes of the DAG to assess the similarity of two terms. This type of measure uses metrics ranging from a basic term overlapping count [75] to more complex ones such as metrics based on information content (IC) and other approaches [71,113,85]. There are also measures that rely on the DAG topology given by its edges [111,70,2,49]. All these different approaches have their own limitations which has led to the development of methods that attempt to benefit as much as possible from all the information contained in GO by integrating several approaches together such as path length and IC-content [77,83,114,112,109,9]. To further improve the accuracy of semantic similarity measures recent research suggests improving existing semantic similarity measures by combining them with other relevant data sources as co-function networks [84,63,54].

Most of these semantic similarity measures are designed to assess the similarity between a pair of GO terms. In order to quantify the likeness between two genes or gene sets, these pairwise measures need to be transformed into a single representative value [101]. This transformation can be performed using several approaches such as the average, maximum, best-match average, functional similarity or information theory-based semantic similarity.

Although semantic similarity measures solve the issues of previous approaches, they do not consider how the genes under study interact among each other, which is the principal feature of gene networks over gene sets.

Here we present GFD-Net, a novel methodology that extends the concept of semantic similarity with the use of the gene network topology to analyze the functional dissimilarity of gene networks based on Gene Ontology (GO). It uses the term depth and path length concepts to select a functionality for each gene (specified by a GO term), weight each edge in the network according to the dissimilarity between the nodes at its ends and calculate a quantitative measure of the network functional dissimilarity, i.e. a quantitative value of the degree of dissimilarity between the connected genes. GFD-Net is based on the idea of "most cohesive (common and specific) function" presented in GFD [23] and its main innovation lies in its capability for taking into account the network topology. GFD-Net is available as a Cytoscape [95] app which streamlines its configuration and execution. This app was presented in Diaz-Montana et al. [24], which was strictly about the app architecture and its integration in Cytoscape and did not include any information regarding the methodology. The robustness of GFD-Net as a gene network validation tool was demonstrated by performing a ROC analysis on several network repositories. Furthermore, a well-known network was analyzed showing that GFD-Net can also be used to infer knowledge, and an example of how GFD-Net can be applied to the study of human diseases is presented.

2. Method

A gene network can be represented as a graph $G = (V, E)$, which is composed by a pair of disjoint and finite sets V and E. V denotes the set of vertices (genes). E denotes the set of edges between the elements in V and can be represented as a matrix where each column and row represent a gene and each cell contains the weight of the edge between them, or 0 if there is no edge. As mentioned before, existing approaches evaluate gene networks considering the set V without taking into account the network structure, i.e. the set E. GFD-Net is the first approach to measure the similarity of a gene network considering both sets V and E.

The methodology is outlined in Fig. 1, which shows the execution of GFD-Net on a sub-net of the Peroxisome pathway from KEGG. According to KEGG, Peroxisomes are essential organelles that play a key role in redox signaling and lipid homeostasis. They contribute to many crucial metabolic processes such as fatty acid oxidation, biosynthesis of ether lipids and free radical detoxification. Matrix proteins in the cytosol are recognized by peroxisomal targeting signals (PTS) and transported to the docking complex at the peroxisomal membrane.

A gene network may encompass one or more biological functions. GFD-Net assumes that each gene in V is involved in the same or a similar biological function as the genes that it is connected to so all the connected genes in the network are somehow related in a biological sense. Based on this premise, GFD-Net selects the most cohesive (common and specific) function in the context of the network for each gene in V, thus minimizing the semantic dissimilarity between all the connected genes in the gene network. The gene functions are given by an annotation in GO (GO term).

2.1. Inputs and outputs

GFD-Net inputs include the network being analyzed, the organism that the network belongs to and the ontology being used for the analysis. Although it is possible to work with different network types, GFD-Net considers its input network to be an association

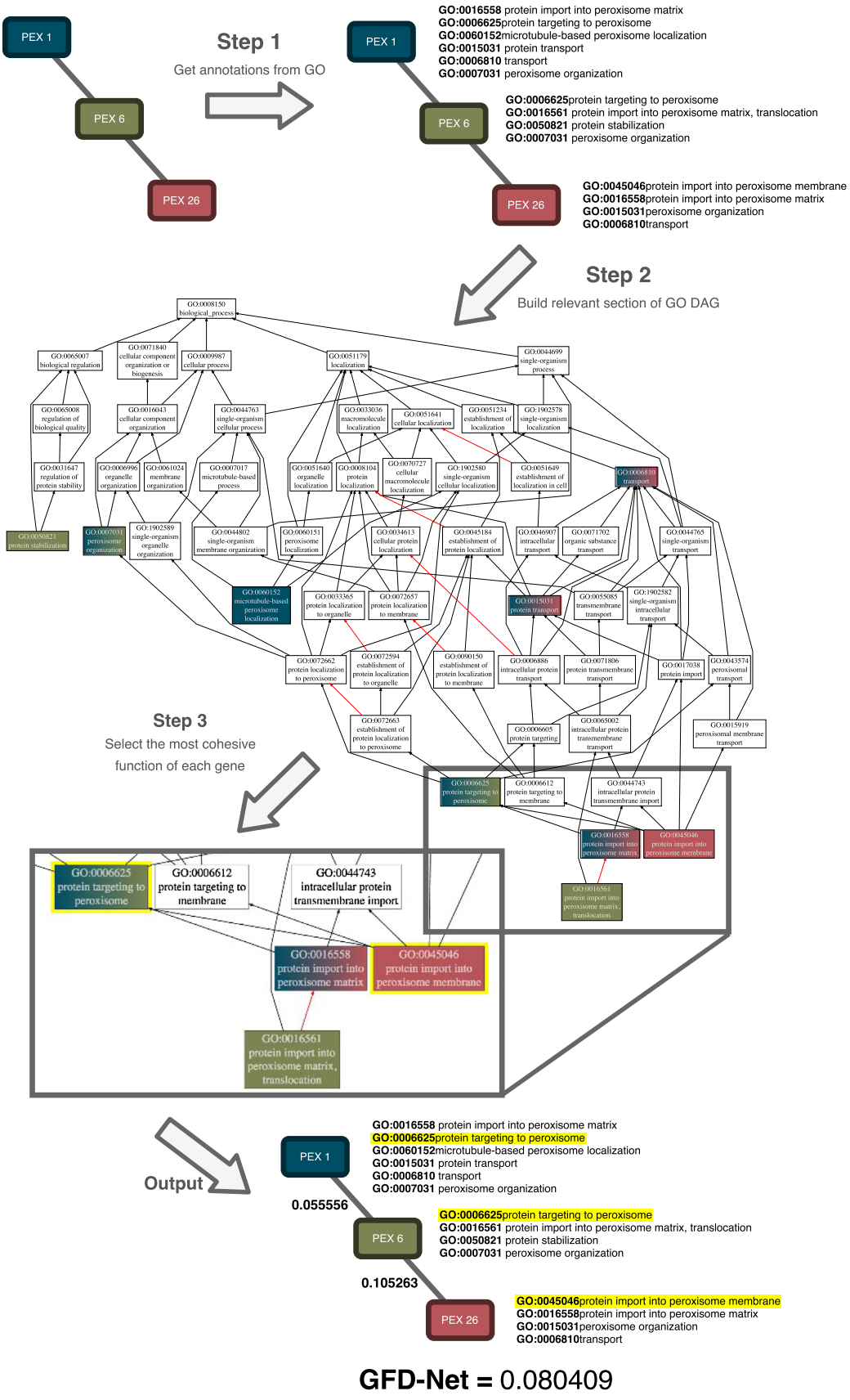


Fig. 1. Overall scheme of GFD-Net method using a sub-net of the Peroxisome pathway from KEGG as an example network being analyzed.

network as this type of undirected and unweighted network represents the highest level of abstraction when modeling biological networks [39,40].

GFD-Net outputs are the network dissimilarity and a network enriched with the information obtained. The network dissimilarity is represented by a numeric value ranging from 0 to 1, where values close to 0 mean "similar", and values near 1 mean "dissimilar". The output network is similar to the input network excluding the genes that are unknown or not annotated for the given ontology and weighting each edge by the functional dissimilarity between the genes at its ends. Also, each gene is enriched with the list of its associated GO terms in the given ontology as well as the selected GO term describing its predicted function in the context of the network.

2.2. Step 1: Get the genes annotations from GO

The first step of GFD-Net, as shown in Fig. 1, consists of retrieving all the relevant gene annotations from GO. GFD-Net assumes that the identifiers used for the genes are supported by GO. GFD-Net considers all the annotations associated to each gene and all its synonyms as well as all the annotations associated to the products encoded by them. All the duplicated, unknown or unannotated genes are discarded.

2.3. Step 2: Build the relevant section of the GO DAG

Once GFD-Net has all the necessary information, it generates the relevant subsection of the GO directed acyclic graph (DAG) based on the annotations of the genes in the network, as shown in the second step of Fig. 1. Unlike GFD, apart from the "is_ a" relationships, GFD-Net considers the "part_ of" and "occurs_ in" relationships. This is because our experiments highlighted the fact that important relationships were ignored, lowering GFD-Net accuracy. Including these extra relationships in the analysis adds more information and therefore leads to the selection of more cohesive functions for the genes. The fact that the ontologies may be "part_ of" and "occurs_ in" incomplete is irrelevant, because they are not used instead of but in addition to the "is_ a" relationships for which the ontologies are complete.

Fig. 2 illustrates an example of the accuracy loss when calculating the dissimilarity of GO:0005634 (nucleus) and GO:0000139 (Golgi membrane). If only the "is_ a" relationships were considered, it would be unknown that the Golgi apparatus is part of the endomembrane system and relationships as obvious as organelle membrane and membrane-bounded organelle, organelle part and organelle, cell part and cell, etc. would be missed. Including these missing relationships may increase the terms depth and create shorter paths between them, hence improving GFD-Net results. For example, considering only the "is_ a" relationships, the lowest common ancestor (LCA) of GO:0005634 (nucleus) and GO:0000139 (Golgi membrane) would be GO:0044424 (intracellular part) which is not a very precise term (being only 2 levels away from the root). The dissimilarity given by GFD-Net to these two GO terms would be 0.58. However, once the other relationships are added, the LCA becomes GO:0043231 (intracellular membrane-bounded organelle) which is more precise than GO:0044464 (cell part). Using this improved DAG the dissimilarity given by GFD-Net is 0.27. Therefore, it can be concluded that considering these extra relationship types when building the DAG results in a more informative DAG which in turn results in more accurate information extracted by GFD-Net and a more accurate dissimilarity measure.

2.4. Step 3: Select the most cohesive function of each gene

Once the DAG has been built, GFD-Net performs the actual analysis. The original idea of GFD was to try all the possible combinations of annotations for the given genes. However, the time consumed by this approach grows exponentially with the size of the gene set making it unusable for large sets. Therefore, a heuristic technique based on the Voronoi Diagram concept [5] was introduced in order to reduce the complexity of the search from exponential to polynomial order (see the Computational analysis section of Diaz-Diaz et al. [23]). GFD-Net uses the same approach which utilizes GO DAG as search space. For each node in the search space, GFD-Net searches for the closest annotations of each gene in the input network, obtaining a GO term (function) for each gene and creating a candidate solution.

The distance between two annotations (GO terms) $g_\alpha t_\gamma$ and $g_\beta t_\delta$ associated to the genes g_α and g_β respectively is given by:

$$\mathcal{R}(t_\alpha, t_\beta) = \frac{distance(t_\alpha, t_\beta)}{2 * depth(LCA(t_\alpha, t_\beta)) + distance(t_\alpha, t_\beta)} \quad (1)$$

where distance denotes the minimum number of nodes separating two nodes in the DAG, depth indicates the maximum number of nodes between a node and the root of the DAG and LCA represents the lowest common ancestor (LCA) between two nodes. Both length and depth include the node being considered i.e. the depth of the root node and the distance between a node and itself are both 1. This measure penalizes GO term pairs that are widely separated and rewards specialization assuming that the further a term is from the root the more specialized it is. $\mathcal{R}$ values range between 0 and 1, where values close to 0 mean "similar", and values near 1 mean "dissimilar" considering the prior knowledge contained in GO. It should be noted that this formula is essentially the same formula presented by GFD [23]. However, it enforces that only paths to the root that contain the LCA of the terms are considered to calculate the depth of the terms. This restriction ensures that the depth of each term is calculated in the context of the comparison and avoids biasing the result by considering a higher depth in some unrelated context. An example of this issue can be seen in Fig. 1 when calculating the distance between GO:0007031 (peroxisome organization) and GO:0060152 (microtubule-based peroxisome localization). Their LCA is GO:0009987 (cellular process) and thus the depth of GO:0060152 is 5. However, using GFD's original formula the depth of GO:0060152 would have been 6, choosing the path that contains GO:0051641 (cellular localization). This means that GO:0060152 is more informative in the context of localization than in the context of cellular processes. Since the terms being compared are related as both being cellular processes and not for being related to location, it should be regarded as less informative.

The main innovation of GFD-Net lies in the way this pairwise measure is used to analyze gene networks taking into account the network topology. While previous measures only consider all the possible term pairs (e.g. GFD averages the dissimilarity between each term pair given by $\mathcal{R}$), GFD-Net only takes into account term pairs associated to genes connected in the input network. This penalizes disconnected nodes that are functionally similar and connected nodes that are not functionally similar. Looking to Fig. 1, it is clear that the most specific (deepest) and most common (closest) nodes for the three genes in the networks are GO:0016558 (protein import into peroxisome matrix) for PEX1 and PEX26 and GO:0016561 (protein import into peroxisome matrix, translocation) for PEX6. However, PEX1 and PEX26 are not connected while PEX1 and PEX6 are connected. GFD-Net leverages this knowledge and chooses GO:0006625 (protein targeting to peroxisome) for PEX1 and PEX6 and GO:0045046 (protein import

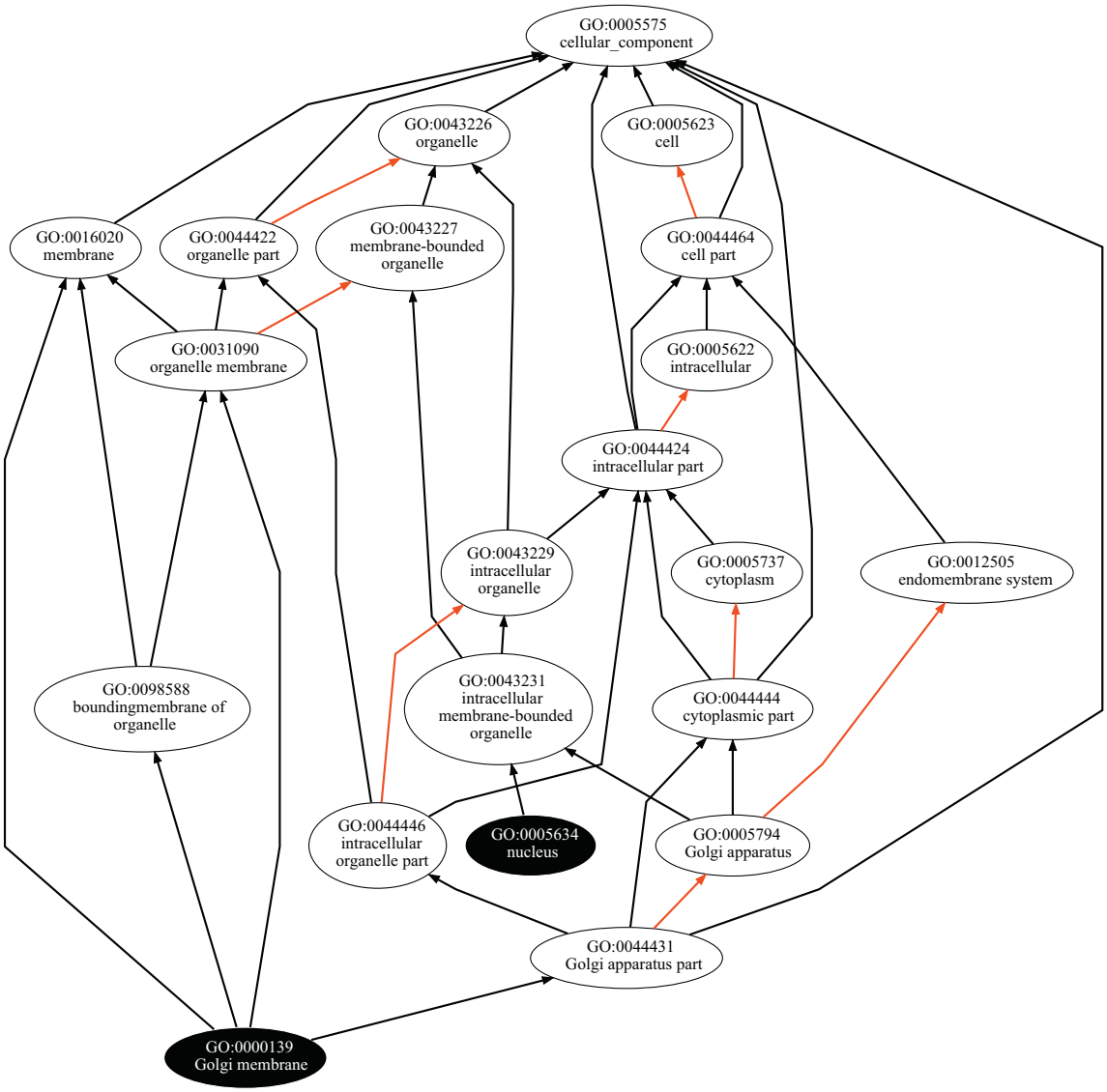


Fig. 2. Relevant section of GO DAG when comparing GO:0005634 (nucleus) and GO:0000139 (Golgi membrane). Black edges represent "is_ a" relationships while red edges represent "part_ of" and "occurs_ in" relationships.

into peroxisome membrane) for PEX26. This function selection is consistent with existing research [36,68]. By weighting the edges, GFD-Net is also capable of discovering that PEX1 and PEX6 are more closely related to each other than PEX6 and PEX26.

Considering E, the set of edges in the network, where $E[g_\alpha, g_\beta]$ is the weight of the edge connecting the genes g_α and g_β or 0 if they are disconnected, the dissimilarity of two gene annotations can be rewritten as:

$$\mathcal{R}'(g_\alpha t_\gamma, g_\beta t_\delta) = \begin{cases} 0 & \text{if } E[g_\alpha, g_\beta] = 0; \\ \mathcal{R}(g_\alpha t_\gamma, g_\beta t_\delta), & \text{if } E[g_\alpha, g_\beta] \neq 0; \end{cases} \quad (2)$$

Let V be a set of genes $g_1, g_2, \ldots, g_n$. The set of GO terms associated with a gene g_i is given by $T(g_i) = g_i t_1, g_i t_2, \ldots, g_i t_n$ so the Cartesian product $P(V) = T(g_1) \times T(g_2) \times \ldots \times T(g_n)$ defines the set of all possible sets of GO terms. Each $p \in P$ represents a possible solution where $p(\delta)$ represents the selected GO term associated to the gene g_δ. The dissimilarity $\mathcal{S}$ of an annotation set p is given by the average of the dissimilarity between the term pairs associated to genes connected in the input network:

$$\mathcal{S}(p) = \frac{\Sigma_{\forall \delta,\gamma | 0<\delta<\gamma<|p|} \mathcal{R}'(p(\delta), p(\gamma))}{\Sigma_{\forall \delta,\gamma | 0<\delta<\gamma<|V|} E[g_\delta, g_\gamma]} \quad (3)$$

Finally, GFD-Net is the lowest dissimilarity for all possible GO term sets for a given network.

$$GFDnet(V, E) = \min_{\forall p \in P} \mathcal{S}(p) \quad (4)$$

3. Results and discussion

3.1. GFD-Net as a gene network validator

GFD-Net was directly compared to its predecessor, GFD [23] to prove that considering the network topology yields more accurate results than only considering the nodes as a set. Also, its accuracy as a gene network validator was assessed by performing a ROC analysis comparing the results of analyzing functionally coherent networks with the results of analyzing randomly generated networks containing the same genes.

ROC analysis assesses the performance of classifiers and rankers as a trade-off between sensitivity (true positive rate) and specificity (false positive rate). It can be plotted as a curve and the area under the ROC curve (AUC) is often taken as a measure of the prediction performance. An area of 0.5 represents random forecasts, while an area of 1 reflects perfect forecasts.

The reason to keep the same genes between the known networks and the random ones is to focus on the ability of GFD-Net to consider the network topology in the analysis. It should be noted that since all the existing approaches ignore the topology they are unable to differentiate between the known networks and the random ones resulting in incorrect validation results and an AUC of 0.5 in the ROC analysis.

The dataset used consisted of all the gene networks mapped from the pathways contained in Biocarta [79], KEGG [58], NCI/Nature Pathway Interaction Database [92], PANTHER [74] and Reactome [55] databases. The mapping was performed using the Graphite package [91] from Bioconductor [37]. As said above, GFD-Net assumes that each gene in the network is involved in the same or a similar biological function as the genes that it is connected to so all the connected genes are somehow related in a biological sense. Even though a metabolic pathway may encompass more than one biological function, each pair of connected genes participate in the same biological process and thus the premise of GFD-Net is fulfilled.

For each gene network, 100 random networks were generated by shuffling the edges. This randomization process only changes the relationships between the genes (edges) while keeping the original gene set (nodes) and the number of edges. Keeping the number of edges avoids resulting in empty networks, which do not offer enough information, and complete networks, which are almost like gene sets and do not offer much information either.

In that sense, all the networks with less than the 5% or more than the 95% of the edges that it would have if they were complete were removed because randomizing networks that are too empty or too full will yield too many duplicated networks biasing the final results. This percentage filtering does not work well for networks with few nodes which seem to be sufficiently complete but cannot generate enough random networks so these were also removed.

To ensure that the randomization process could be computed in reasonable time all the networks with more than 1000 nodes were removed. This was done to reduce experimentation time to a reasonable scale and should not reduce the quality of the analysis because a network with so many nodes will most probably not fulfill the GFD-Net premise as mentioned above.

After this pre-processing of the input dataset, 1678 valid networks were considered and 832 invalid ones were removed; 608 were too small, 6 were too large, 63 were too empty, 101 were too complete and 54 did not contain any annotations in GO in any of the three ontologies (see supplementary material for more details).

First, GFD-Net was compared to GFD [23] since it already proved to be a better classifier than other semantic similarity measures. GFD-Net uses the concept of "most cohesive (common and specific) function of each gene" presented in GFD and their results can be directly compared. Because GFD is designed to validate gene sets, in order to apply it to gene networks, their topology was disregarded and only the sets of genes were considered. It should be noted that GFD was selected as an example and the results should be similar if comparing GFD-Net with other gene set semantic similarity measures.

Fig. 3 shows a sample of the comparison between GFD-Net and GFD on some networks extracted from KEGG pathways (see supplementary material for the full results). Because GFD ignores the network topology, the average value of GFD for the randomized networks is the same as for the real network. GFD-Net consistently outperformed GFD and the average dissimilarity of the randomized networks given by GFD-Net is very similar to GFD. It makes sense that using random edges gives similar results than not considering the edges at all.

Once it has been noted that GFD-Net appears to perform better than previous measures and is able to correctly take into consideration the network topology during the analysis, a ROC analysis was performed to statistically prove that GFD-Net is better than previously existing measures. It was decided to take the analysis to the most extreme case and assess the performance of GFD-Net to validate a gene network based solely on its topology, i.e. given two networks that contain functionally coherent genes, be able to assess the correctness of the relationship among them (edges). It should be noted that this is an edge case where the nodes are always biologically meaningful but the edges might not be.

Since the relationships among the elements in a biological pathway are defined by their role in the biological process, only the Biological Process ontology was considered for the analysis.

The results of applying GFD-Net on the real networks were compared with the average results obtained for the random ones and a ROC analysis was performed, so that a total of 164024 networks were analyzed using GFD-Net. The ROC curve is plotted over the interval $[0, 1]$ with increments of 0.01 and the area under the ROC curve (AUC) is enclosed in brackets as illustrated in Fig. 4.

Figs. 4 shows the ROC curves obtained by applying GFD-Net to the different repositories. The results were satisfactory, especially considering that networks containing exactly the same genes were compared. As previously stated in the introduction section, there are many approaches to discriminate gene sets based on their biological meaning. However, none of them solve the problem of assessing whether a network, i.e. a gene set and the relationships among them, is biologically meaningful or not. All the existing approaches ignore the network topology and are not able to distinguish the meaningful networks from the random ones yielding an AUC of 0.5. GFD-Net, however, is able to evaluate the same gene set based on the relationship among the genes given by the network topology.

GFD-Net is an effective gene network validator that correctly takes the network topology into consideration. Furthermore, it was proved that, given a network containing functionally coherent genes, GFD-Net is able to assess the correctness of the relationships among them given by the network structure.

3.2. GFD-Net as an analysis tool

GFD-Net is not only a validation tool. It can also be used to analyze existing networks and extract knowledge from them. In order to demonstrate this capability, the gene network published by Hazbun et al. [45] was analyzed. This network is provided by YeastNet V3 [43] and represents the interactome for 100 essential genes mapped by a combination of four complementary methods (affinity purification and mass spectrometry analysis to identify co-purifying proteins, two-hybrid analysis to identify interacting proteins, fluorescence microscopy to localize the proteins, and structure prediction methodology to predict structural domains or identify remote homologies). Hazbun et al. used GO term finder in order to annotate the network on each GO ontology. The biological process term was associated based on the common GO terms among the protein purification data set and the two-hybrid data set, the cellular component term based on fluorescent microscopy and the molecular function based on the remote homologies using PSI-BLAST.

GFD-Net results were positive for all three ontologies and GFD-Net was able to correctly identify the function of each protein yielding very similar results to Hazbun et al. The rest of this section is focused on the analysis performed on the Biological Process

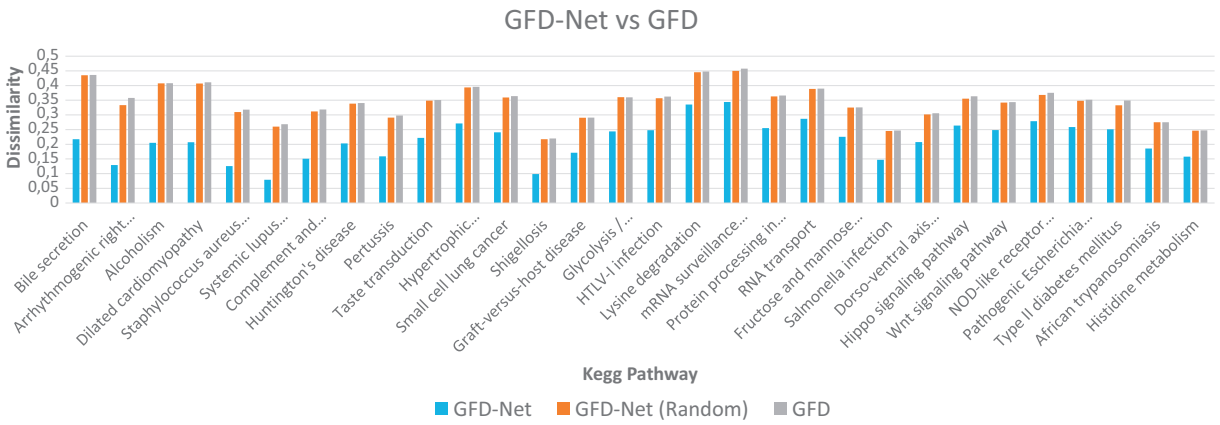


Fig. 3. Comparison of the dissimilarity values given by GFD and GFD-Net when applied to some KEGG pathways and the dissimilarity value given by GFD-Net when applied to the same KEGG pathways but randomizing the interactions among the genes.

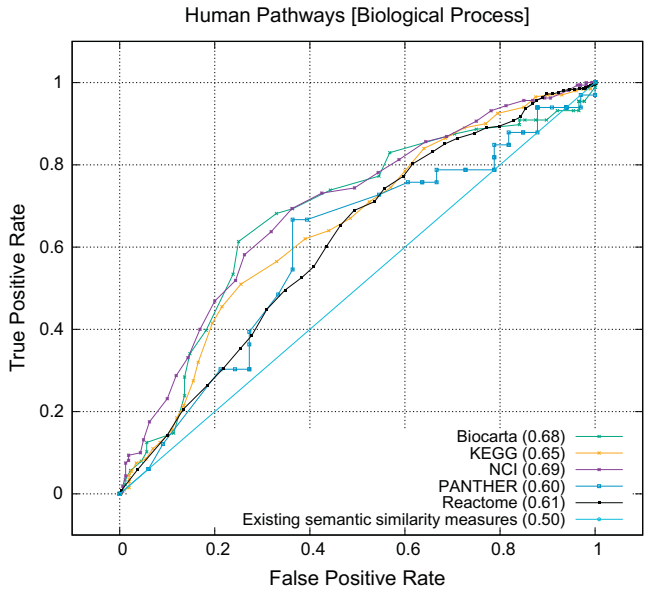


Fig. 4. ROC analysis for GFD-Net in the different repositories, as applied to the biological process ontology in GO. The area under the ROC curve is indicated in brackets.

ontology, but the same approach can be used on the other ontologies obtaining similar results (see supplementary material for the data related to the other ontologies).

Table 1 shows the comparison between the GO terms in the Biological Process ontology assigned by the author and GFD-Net to each of the genes contained in the original paper by Hazbun et al. that have not changed through the YeastNet revisions. GFD-Net annotation selection was consistent with the annotations originally selected by Hazbun et al.

Fig. 5 shows the results of the analysis as displayed by GFD-Net app. Each gene has been colored according to its selected annotation as shown in Table 1. These results match the different sets identified by Hazbun et al. proving the capability of GFD-Net to leverage the network topology during the analysis. For example: (a) The complex containing YIR010W (DSN1) and YPL233W (NSL1) was associated by the author to the "chromosome segregation" process. GFD-Net associates almost all the genes to the "mitotic nuclear division" process which involves the "chromosome segregation" process. The result is correct, but could have been more specific. (b) The complex containing YKL088W was associated by the author to the "Coenzyme A biosynthesis" process and GFD-Net associates all the genes to the same process. (c) The complexes containing YJL072C (PSF2), YDR288W and YML023C were

Table 1
Hazbun et al. vs GFD-Net.

Gene	GFDnet BP	BP
YDL209C	mRNA processing	mRNA splicing
YKR022C	mRNA processing	mRNA splicing
YLR424W	mRNA processing	mRNA splicing
YLR145W,	rRNA processing	TRNA processing
YIR010W	Mitotic nuclear division	Chromosome segregation
YPL233W	Mitotic nuclear division	Chromosome segregation
YLR132C	mRNA processing	mRNA splicing
YNL313C	Cell wall organization	Nuclear membrane fusion
YLL034C	Ribosomal large subunit export from nucleus	Organelle organization and biogenesis
YJL091C	GPI anchor biosynthetic process	Secretory pathway
YPR169W	Ribosomal large subunit biogenesis	Protein mono-ubiquitination
YDR288W	DNA recombination	DNA repair
YML023C	DNA recombination	DNA repair
YDR013W	DNA replication	DNA repair
YDR489W	DNA replication	DNA repair
YJL072C	DNA replication	DNA repair
YOL146W	DNA replication	DNA repair
YJR012C	biological_process	Transport
YGR002C	Transcription, DNA-templated	Histone acetylation
YJL010C	rRNA processing	rRNA processing
YDR365C	rRNA processing	rRNA processing
YGR145W	rRNA processing	rRNA processing
YKL195W	Protein folding	Protein targeting
YFR003C	Glycogen metabolic process	Cell cycle
YJR136C	Chromatin modification	Protein biosynthesis
YGR046W	Phospholipid biosynthetic process	Mitochondrial translocation
YGR198W	Establishment of protein localization to plasma membrane	MAPKKK cascade
YHR085W	rRNA processing	Unknown
YHR197W	rRNA processing	Unknown
YNL182C	rRNA processing	Unknown
YNL245C	mRNA processing	mRNA splicing
YKR038C	tRNA processing	Response to dessication
YKR079C	tRNA processing	DNA catabolism, RNA catabolism
YNL260C	Translational initiation	Unknown
YKL088W	Coenzyme A biosynthetic process	Coenzyme A biosynthesis
YNL124W	Pseudouridine synthesis	rRNA processing
YPL063W	Protein import into mitochondrial matrix	Mitochondrial translocation

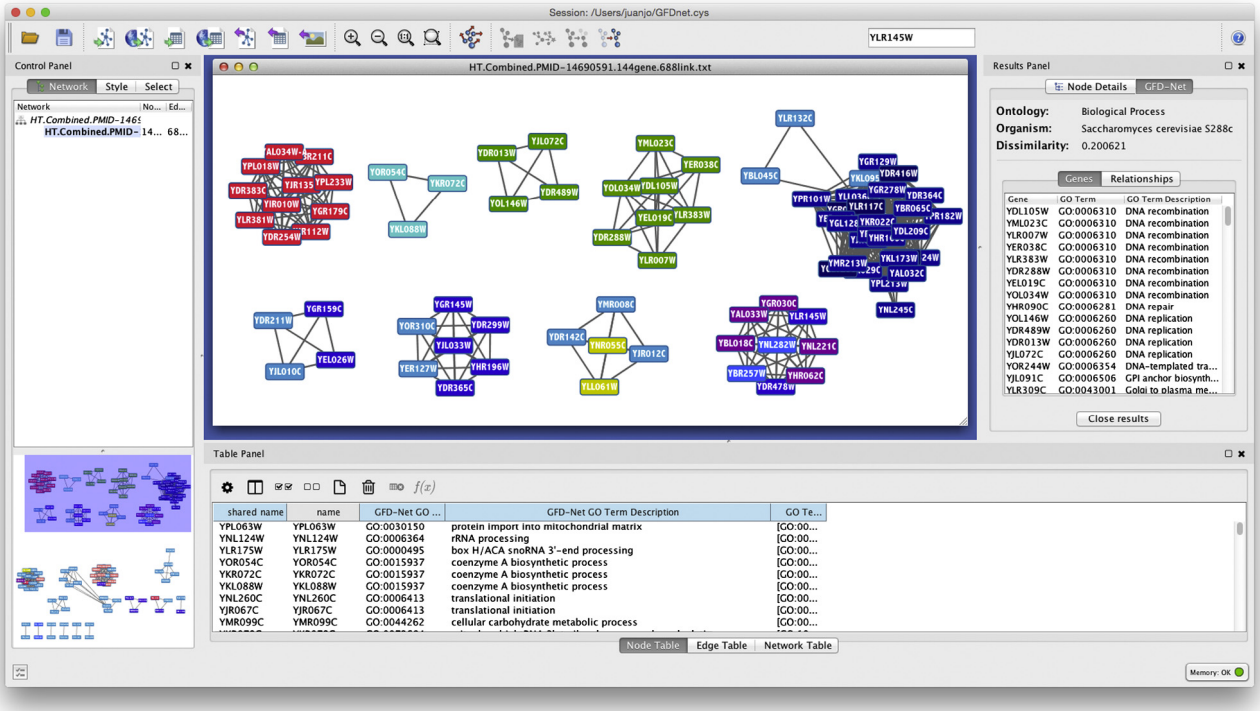


Fig. 5. Screenshot of an analysis of Hazbun et al. gene network using GFD-Net app. Genes are colored according to the GO terms selected by GFD-Net.

associated by the author to the "DNA repair" process. GFD-Net associates the first to the "DNA replication" process and the other two, which were combined into one by YeastNet, to the "DNA recombination" process. The three terms belong to the "DNA metabolic process" and are very closely related. (d) The complexes containing YDL209C (CWC2), YLR424W and YKR022C, YLR132C and YNL245C (CWC25) were associated by the author to the "mRNA splicing" process which is obsolete because it represents several different mRNA-related processes. All these complexes were combined into one by YeastNet and GFD-Net associated most of their genes to the "mRNA processing" process, a more generic term of covering the same processes. (e) The complexes containing YJL010C and YDR365C were associated by the author to the "rRNA processing" process. GFD-Net associates half of the genes to the same process and one gene to the very specific "endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)" process. Here GFD-Net was able to provide more concrete information than the original paper. (f) The complex containing YJR012C was associated by the author to the "transport" process. Once again, GFD-Net was able to provide more concrete information associating the genes in the complex to the "transmembrane transport", "protein import into peroxisome matrix, docking" and "lipid metabolic process" processes. (g) The complex containing YLR145W was associated by the author to the "tRNA processing" process. GFD-Net associates some of the genes to the same process but, once again, was able to provide more specific information associating half of the genes to the "RNA phosphodiester bond hydrolysis, endonucleolytic" process.

It is clear in Fig. 5 that the analyzed network was sparse and composed by several small subnetworks. GFD-Net was applied to each of these subnetworks separately, yielding very similar results to the ones obtained when executed on the whole network. This proves that GFD-Net is able to differentiate relevant components within the analyzed networks which is due to the fact that it performs the analysis considering the network topology as well as the nodes. As it has already been mentioned, none of the existing measures are capable of that.

It is proved that GFD-Net was able to take into consideration the network topology and identify the existing subnetworks. Furthermore, the genes in the network could be annotated in a consistent manner with the existing research, providing even more information than previous approaches.

3.3. GFD-Net applied to the study of human diseases

The results of GFD-Net assign a functionality to each gene in the input network. Although this is useful by itself, it is normally more desirable to infer information about the network as a whole.

Here, an example of how GFD-Net can be used to associate relevant diseases to a specific gene network is presented. The aim of this experiment is to show how GFD-Net is able to identify human diseases that are related to a specific network.

To that end, some of the best-known pathways from KEGG related to diseases are analyzed. As in the subSection 3.1, the networks were extracted from the pathways using Graphite. Concretely, three pathways were analyzed: "Type II diabetes mellitus" (hsa04930), "Insulin resistance" (hsa04910) and "Alzheimer's disease" (hsa05010). For brevity, the analysis process is detailed only for the "Type II diabetes mellitus" pathway, but the results for the three pathways are discussed.

First, the most cohesive (common and specific) function of each gene in the network is identified using GFD-Net (see Table 2 column 1 and 2).

Next, curated GODiseaseGene inference data from the Comparative Toxicogenomics Database (CTD) [20] is used to associate GFD-Net results with human diseases. CTD associates each disease which one or more GO term and each GO term in the context of the disease with one or more genes. Each gene–GO term pair obtained by GFD-Net is assigned all the diseases in CTD associated to them,

Table 2
Diabetes Type II genes.

Gene	GO-Term	GO-Term description	Related diseases
SOCS4	GO:0035556	Intracellular signal transduction	
PIK3R5	GO:0007186	G-protein coupled receptor signaling pathway	
MTOR	GO:0007165	Signal transduction	Carcinoma, Hepatocellular—Hypertension—Adenocarcinoma—Ovarian Neoplasms—Schizophrenia—...
GCK	GO:0035556	Intracellular signal transduction	
HK1	GO:0001678	Cellular glucose homeostasis	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Myocardial Ischemia—HEMOLYTIC ANEMIA, NONSPHEROCYTIC, DUE TO HEXOKINASE DEFICIENCY—...
HK2	GO:0008637	Apoptotic mitochondrial changes	Myocardial Ischemia—Obesity
HK3	GO:0001678	Cellular glucose homeostasis	Liver Diseases
IKBKB	GO:0007249	I-kappaB kinase/NF-kappaB signaling	
INS	GO:0007186	G-protein coupled receptor signaling pathway	Diabetes Mellitus, Type 2—Insulin Resistance—Liver Cirrhosis, Experimental—Prostatic Neoplasms—Hypertension—...
INSR	GO:0023014	Signal transduction by protein phosphorylation	Diabetes Mellitus, Experimental—Insulin Resistance—Hyperalgesia—Alzheimer Disease—Hyperglycemia—...
PDX1	GO:0007224	Smoothened signaling pathway	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Pancreatic Neoplasms—Maturity-Onset Diabetes of the Young, Type 4—Pancreatic Agenesis, Congenital—...
IRS1	GO:0007165	Signal transduction	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Insulin Resistance—Carcinoma, Hepatocellular—Prostatic Neoplasms—...
MAFA	GO:0007166	Cell surface receptor signaling pathway	
PIK3CA	GO:0043491	Protein kinase B signaling	Carcinoma, Hepatocellular—Prostatic Neoplasms—Stomach Neoplasms—Colorectal Neoplasms—Lymphoma, Large B-Cell, Diffuse—...
PIK3CB	GO:0007165	Signal transduction	Lymphoma, Large B-Cell, Diffuse—Schizophrenia
PIK3CD	GO:0007165	Signal transduction	Prostatic Neoplasms—Lymphoma, Large B-Cell, Diffuse—Lymphoma, Mantle-Cell
PIK3CG	GO:0007186	G-protein coupled receptor signaling pathway	Heart Failure—Lymphoma, Large B-Cell, Diffuse—Cardiomegaly—Autistic Disorder—Fibrosis—...
PIK3R1	GO:0007165	Signal transduction	Insulin Resistance—Carcinoma—Mammary Neoplasms, Animal—Mammary Neoplasms, Experimental—Burkitt Lymphoma—...
PIK3R2	GO:0007165	Signal transduction	Megalanecephaly Polymicrogyria-Polydactyly Hydrocephalus Syndrome
PRKCD	GO:0007165	Signal transduction	Diabetes Mellitus, Experimental—Liver Cirrhosis, Experimental—Hypertension—Seizures—Neurotoxicity Syndromes—...
PRKCE	GO:0007165	Signal transduction	Diabetes Mellitus, Experimental—Hyperalgesia—Myocardial Ischemia—Colorectal Neoplasms—Cardiomyopathies—...
PRKCZ	GO:0007165	Signal transduction	Prostatic Neoplasms—Leukemia
MAPK1	GO:0007165	Signal transduction	Hyperalgesia—Stomach Neoplasms—Disease Models, Animal—Trigeminal Neuralgia—Neoplasm Metastasis—...
MAPK3	GO:0035556	Intracellular signal transduction	Hyperalgesia—Prostatic Neoplasms—Stomach Neoplasms—Disease Models, Animal—Trigeminal Neuralgia—...
MAPK8	GO:0000165	MAPK cascade	Hyperalgesia—Stomach Neoplasms—Disease Models, Animal—Trigeminal Neuralgia—Reperfusion Injury—...
MAPK9	GO:0000165	MAPK cascade	Hyperalgesia—Disease Models, Animal—Trigeminal Neuralgia—Neoplasm Metastasis—Reperfusion Injury—...
MAPK10	GO:0007165	Signal transduction	Epileptic encephalopathy, Lennox-Gastaut type
ABCC8	GO:0007165	Signal transduction	Diabetes Mellitus, Type 2—Colorectal Neoplasms—Diabetes Mellitus, Type 1—Diabetes Mellitus, Permanent Neonatal—Congenital Hyperinsulinism—...
TNF	GO:0097527	Necroptotic signaling pathway	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Insulin Resistance—Hyperalgesia—Carcinoma, Hepatocellular—...
CACNA1A	GO:0007214	Gamma-aminobutyric acid signaling pathway	Ataxia—Epilepsy, Absence—Episodic Ataxia, Type 2—Exfoliation Syndrome—Hemiplegic migraine, familial type 1—...
CACNA1B	GO:0034765	Regulation of ion transmembrane transport	Peripheral Nervous System Diseases
CACNA1C	GO:0035585	Calcium-mediated signaling using extracellular calcium source	Hypertension—Hypoglycemia—Autistic Disorder—Bipolar Disorder—Arrhythmias, Cardiac—...
CACNA1D	GO:0007188	Adenylate cyclase-modulating G-protein coupled receptor signaling pathway	Adenoma—Deafness—Hyperaldosteronism
CACNA1E	GO:0034765	Regulation of ion transmembrane transport	
HKDC1	GO:0001678	Cellular glucose homeostasis	
IRS4	GO:0007165	Signal transduction	
PIK3R3	GO:0008286	Insulin receptor signaling pathway	
SOCS1	GO:0035556	Intracellular signal transduction	Liver Cirrhosis, Experimental—Reperfusion Injury—Liver Diseases—Dermatitis, Allergic Contact—Liver Neoplasms, Experimental—...
IRS2	GO:0007165	Signal transduction	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Insulin Resistance—Carcinoma, Hepatocellular—Stomach Neoplasms—...
SOCS2	GO:0035556	Intracellular signal transduction	Liver Cirrhosis, Experimental—Dermatitis, Allergic Contact—Narcolepsy
CACNA1G	GO:0034765	Regulation of ion transmembrane transport	Intellectual Disability
SOCS3	GO:0007165	Signal transduction	Carcinoma, Hepatocellular—Myocardial Ischemia—Liver Diseases—Atherosclerosis—Dermatitis, Atopic, 4—...
ADIPOQ	GO:0033211	Adiponectin-activated signaling pathway	Diabetes Mellitus, Type 2—Insulin Resistance—Liver Cirrhosis, Experimental—Hypertension—Myocardial Ischemia—...
SLC2A4	GO:0071456	Cellular response to hypoxia	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Insulin Resistance—Alzheimer Disease—Cardiomegaly—...
PKLR	GO:0032869	Cellular response to insulin stimulus	Diabetes Mellitus, Experimental—Liver Cirrhosis, Experimental—Adenosine Triphosphate,

(*continued on next page*)

Table 2 (*continued*)

Gene	GO-Term	GO-Term description	Related diseases
PKM	GO:0012501	Programmed cell death	Elevated, Of Erythrocytes—Pyruvate Kinase Deficiency of Red Cells—... Carcinoma, Hepatocellular—Neoplasm Invasiveness—Carcinoma—Mammary Neoplasms, Animal—Mammary Neoplasms, Experimental—...
KCNJ11	GO:0034765	Regulation of ion transmembrane transport	Diabetes Mellitus, Experimental—Diabetes Mellitus, Type 2—Insulin Resistance—Diabetes Mellitus, Type 1—Diabetes Mellitus, Permanent Neonatal—..."

if any (see table Table 2 column 4). For example, INSR (the insulin receptor) in the "Type II diabetes mellitus" pathway is assigned "signal transduction by protein phosphorylation" by GFD-Net and this gene-GO term pair is associated with several disease in CTD (Diabetes Mellitus, Experimental, Insulin Resistance, Hyperalgesia, Alzheimer Disease, Hyperglycemia, ...).

Finally, the network is associated to all the diseases associated to each of its genes (according to their selected GO terms) and these diseases are ranked based on their relevance which is given by the number of associated genes within the network. Table 3 shows the five top diseases associated with each pathway (see column 2). It should be noted, that not only the known related disease for each networks was correctly identified, but also, some diseases indirectly related according to existing literature were correctly identified. All those indirectly related disease can be verified in existing literature as shown in column 3. For example, Alzheimer's disease and Prostatic Neoplasms might seem unrelated. However, Nead et al. [78] states: "Our results support an association between the use of ADT in the treatment of prostate cancer and an increased risk of Alzheimers disease in a general population cohort".

This demonstrates how GFD-Net can be used for much more than identifying gene functionalities in the context of a network. As seen above, analyzing a gene network obtained experimentally, GFD-Net can be utilize in identifying the diseases related to it. Also demonstrated above, when a network is related to a known disease, GFD-Net can be used to identify further related diseases. Furthermore, GFD-Net could be used to infer new annotations and increase the knowledge in existing databases/repositories. For example, SOCS2 in the "Type II diabetes mellitus" pathway is associated by GFD-Net to GO:0035556 (intracellular signal transduction) and the pair SOCS2-GO:0035556 is associated in CTD to "Liver Cirrhosis, Experimental", "Dermatitis, Allergic Contact" and "Narcolepsy". However, according to Isshiki et al. [53], a decrease in the expression of SOCS2 involved in the signal transduction process is related to diabetes proving GFD-Net inference capabilities.

Table 3
Diseases associated to the pathways studied.

Pathway	Diseases related	Literature references
Type II diabetes mellitus	Diabetes Mellitus, Experimental	The known related disease
	Diabetes Mellitus, Type 2	The known related disease
	Insulin Resistance	[21,67,107]
	Hyperalgesia	[19,87,11,32]
	Carcinoma, Hepatocellular	[61,105,103,30,29,106]
Insulin resistance	Insulin Resistance	The known related disease
	Liver Cirrhosis, Experimental	[76,15,93,14]
	Diabetes Mellitus, Experimental	[21,67,107]
	Diabetes Mellitus, Type 2	[21,67,107]
	Carcinoma, Hepatocellular	[52,3,88,30,28,98,7]
Alzheimer's disease	Alzheimer Disease	The known related disease
	Prostatic Neoplasms	[78,16]
	Nerve Degeneration	[47,102,48,12,90,64]
	Brain Ischemia	[57,56]
	Cognition Disorders	[18,22,73,6,94]

4. Conclusions

GFD-Net is a novel methodology that applies the concept of semantic similarity to gene network analysis. GFD-Net allows the analysis and validation of gene networks based on their topology and the prior knowledge contained in Gene Ontology.

To demonstrate the robustness of GFD-Net, a ROC analysis was performed using several network repositories in order to analyze the discriminatory power of GFD-Net in assessing the biological meaning of networks where the genes are biologically meaningful but the relationships between them might not be.

Furthermore, a known network was analyzed in order to demonstrate the capabilities of GFD-Net as an analysis tool. The results obtained by GFD-Net were coherent with the original results obtained using affinity purification and mass spectrometry, two-hybrid analysis, fluorescence microscopy and structure prediction methodology showing the potential of GFD-Net to infer knowledge *in silico*.

Moreover,GFD-Net's application to the study of human diseases has been demonstrated through an example experiment.

Finally, it could be argued that GFD-Net is a novel and robust solution for the problem of gene network validation and analysis based on GO.

Conflict of interest

None declared.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.02.013.

References

[1] A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtiö, Y. Pawitan, Network enrichment analysis: extension of gene-set enrichment analysis to gene networks, BMC Bioinform. 13 (1) (2012) 226.
[2] M.A. Alvarez, C. Yan, A graph-based semantic similarity measure for the gene ontology, J. Bioinform. Comput. Biol. 9 (06) (2011) 681–695.
[3] B. Arcidiacono, S. Iiritano, A. Nocera, K. Possidente, M.T. Nevolo, V. Ventura, D. Foti, E. Chiefari, A. Brunetti, Insulin resistance and cancer risk: an overview of the pathogenetic mechanisms, Exp. Diabetes Res. (2012) 2012.
[4] M. Ashburner, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[5] F. Aurenhammer, R. Klein, Voronoi Diagrams, Handbook of Computational Geometry, vol. 5, 2000, pp. 201–290.
[6] L. Bäckman, S. Jones, A.-K. Berger, E.J. Laukka, B.J. Small, Cognitive impairment in preclinical alzheimer's disease: a meta-analysis, 2005.
[7] G. Baffy, E.M. Brunt, S.H. Caldwell, Hepatocellular carcinoma in non-alcoholic fatty liver disease: an emerging menace, J. Hepatol. 56 (6) (2012) 1384–1391.
[8] J.B. Bard, S.Y. Rhee, Ontologies in biology: design, applications and future challenges, Nat. Rev. Genet. 5 (3) (2004) 213–222.
[9] C. Bettembourg, C. Diot, O. Dameron, Semantic particularity measure for functional characterization of gene sets using gene ontology, PloS One 9 (1) (2014) e86525.

[10] C. Bettembourg, C. Diot, O. Dameron, Optimal threshold determination for interpreting semantic similarity and particularity: application to the comparison of gene sets and metabolic pathways using go and chebi, PloS One 10 (7) (2015) e0133579.
[11] A. Bierhaus, T. Fleming, S. Stoyanov, A. Leffler, A. Babes, C. Neacsu, S.K. Sauer, M. Eberhardt, M. Schnölzer, F. Lasitschka, Methylglyoxal modification of nav1. 8 facilitates nociceptive neuron firing and causes hyperalgesia in diabetic neuropathy, Nat. Med. 18 (6) (2012) 926–933.
[12] J.C. Blanks, D.R. Hinton, A.A. Sadun, C.A. Miller, Retinal ganglion cell degeneration in alzheimer's disease, Brain Res. 501 (2) (1989) 364–372.
[13] F. Borelli, R. de Camargo, D. Martins, L. Rozante, Gene regulatory networks inference using a multi-gpu exhaustive search algorithm, BMC Bioinform. 14 (Suppl. 18) (2013) S5.
[14] E. Bugianesi, A. Gastaldelli, E. Vanni, R. Gambino, M. Cassader, S. Baldi, V. Ponti, G. Pagano, E. Ferrannini, M. Rizzetto, Insulin resistance in non-diabetic patients with non-alcoholic fatty liver disease: sites and mechanisms, Diabetologia 48 (4) (2005) 634–642.
[15] P. Cavallo-Perin, M. Cassader, C. Bozzo, A. Bruno, P. Nuccio, A. Dall'Omo, M. Marucci, G. Pagano, Mechanism of insulin resistance in human liver cirrhosis: evidence of a combined receptor and postreceptor defect, J. Clin. Invest. 75 (5) (1985) 1659.
[16] D. Chen, Q.C. Cui, H. Yang, R.A. Barrea, F.H. Sarkar, S. Sheng, B. Yan, G.P.V. Reddy, Q.P. Dou, Clioquinol, a therapeutic agent for alzheimer's disease, has proteasome-inhibitory, androgen receptor–suppressing, apoptosis-inducing, and antitumor activities in human prostate cancer cells and xenografts, Cancer Res. 67 (4) (2007) 1636–1644.
[17] E.S.H. Cheow, W.C. Cheng, C.N. Lee, D. de Kleijn, V. Sorokin, S.K. Sze, Plasma-derived extracellular vesicles contain predictive biomarkers and potential therapeutic targets for myocardial ischemic injury, Molec. Cell. Proteom. (2016). pp. mcp–M115.
[18] J.T. Coyle, D.L. Price, M.R. Delong, Alzheimer's disease: a disorder of cortical cholinergic innervation, Science 219 (4589) (1983) 1184–1190.
[19] L. Daulhac, C. Mallet, C. Courteix, M. Etienne, E. Duroux, A.-M. Privat, A. Eschalier, J. Fialip, Diabetes-induced mechanical hyperalgesia involves spinal mitogen-activated protein kinase activation in neurons and microglia via n-methyl-d-aspartate-dependent mechanisms, Molec. Pharmacol. 70 (4) (2006) 1246–1254.
[20] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, B.L. King, R. McMorran, J. Wiegers, T.C. Wiegers, C.J. Mattingly, The comparative toxicogenomics database: update 2017, Nucl. Acids Res. 45 (D1) (2017) D972–D978.
[21] R. DeFronzo, D. Simonson, E. Ferrannini, Hepatic and peripheral insulin resistance: a common feature of type 2 (non-insulin-dependent) and type 1 (insulin-dependent) diabetes mellitus, Diabetologia 23 (4) (1982) 313–319.
[22] S.T. DeKosky, S.W. Scheff, Synapse loss in frontal cortex biopsies in alzheimer's disease: correlation with cognitive severity, Ann. Neurol. 27 (5) (1990) 457–464.
[23] N. Diaz-Diaz, J.S. Aguilar-Ruiz, GO-based functional dissimilarity of gene sets, BMC Bioinform. 12 (2011) 360.
[24] J.J. Diaz-Montana, N. Diaz-Diaz, Development and use of the cytoscape app gfd-net for measuring semantic dissimilarity of gene networks, F1000Research, 2014.
[25] X. Dong, A. Yambartsev, S.A. Ramsey, L.D. Thomas, N. Shulzhenko, A. Morgun, Reverse engeneering of regulatory networks from big data: a roadmap for biologists, Bioinform. Biol. Insights 9 (2015) 61.
[26] E.R. Dougherty, Validation of gene regulatory networks: scientific and inferential, Brief. Bioinform. 12 (3) (2011) 245–252.
[27] D. Eisenberg, E.M. Marcotte, I. Xenarios, T.O. Yeates, Protein function in the post-genomic era, Nature 405 (6788) (2000) 823–826.
[28] H.B. El-Serag, Hepatocellular carcinoma: recent trends in the united states, Gastroenterology 127 (5) (2004) S27–S34.
[29] H.B. El-Serag, H. Hampel, F. Javadi, The association between diabetes and hepatocellular carcinoma: a systematic review of epidemiologic evidence, Clin. Gastroenterol. Hepatol. 4 (3) (2006) 369–380.
[30] H.B. El-serag, T. Tran, J.E. Everhart, Diabetes increases the risk of chronic liver disease and hepatocellular carcinoma, Gastroenterology 126 (2) (2004) 460–468.
[31] F. Emmert-Streib, M. Dehmer, B. Haibe-Kains, Untangling statistical and biological models to understand network inference: the need for a genomics network ontology, Front. Genet. 5 (2014) 299.
[32] P. Facer, M.A. Casula, G.D. Smith, C.D. Benham, I.P. Chessell, C. Bountra, M. Sinisi, R. Birch, P. Anand, Differential expression of the capsaicin receptor trpv1 and related novel receptors trpv3, trpv4 and trpm8 in normal human tissues and changes in traumatic and diabetic neuropathy, BMC Neurol. 7 (1) (2007) 11.
[33] L. Franke, H. Van Bakel, L. Fokkens, E.D. De Jong, M. Egmont-Petersen, C. Wijmenga, Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, Am. J. Human Genet. 78 (6) (2006). 1011omranian1025.
[34] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J.S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J.A. Castro-Mondragón, Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, Nucl. Acids Res. (2015) gkv1156.
[35] M. Gan, X. Dou, R. Jiang, From ontology to semantic similarity: calculation of ontology-based semantic similarity, Sci. World J. (2013).
[36] B.V. Geisbrecht, C.S. Collins, B.E. Reuber, S.J. Gould, Disruption of a pex1–pex6 interaction is the most common cause of the neurologic disorders zellweger syndrome, neonatal adrenoleukodystrophy, and infantile refsum disease, Proc. Natl. Acad. Sci. 95 (15) (1998) 8630–8635.
[37] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (10) (2004).
[38] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia, Enrichnet: network-based gene set enrichment analysis, Bioinformatics 28 (18) (2012) i451–i457.
[39] F. Gómez-Vela, N. Díaz-Díaz, Gene network biological validity based on gene-gene interaction relevance, Sci. World J. (2014).
[40] F. Gómez-Vela, J.A. Lagares, N. Díaz-Díaz, Gene network coherence based on prior knowledge using direct and indirect relationships, Comput. Biol. Chem. 56 (2015) 142–151.
[41] P.H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Brief. Bioinform. 13 (5) (2012) 569–585.
[42] B. Haibe-Kains, F. Emmert-Streib, Quantitative Assessment and Validation of Network Inference Methods in Bioinformatics, Frontiers Media, SA, 2015.
[43] K. Hanhae, J. Shin, K. Eiru, K. Hyojin, H. Sohyun, S. Jung Eun, L. Insuk, Yeastnet v3: a public database of data-specific and integrated functional gene networks for saccharomyces cerevisiae, Nucl. Acids Res. (2013) gkt981.
[44] M. Harrell, J. Xia, Z. Zhao, Network analysis of gene fusions in human cancer, BMC Bioinform. 14 (Suppl. 17) (2013) A13.
[45] T.R. Hazbun, L. Malmström, S. Anderson, B.J. Graczyk, B. Fox, M. Riffle, B.A. Sundin, J.D. Aranda, W.H. McDonald, C.-H. Chiu, B.E. Snydsman, P. Bradley, E. G. Muller, S. Fields, D. Baker, J.R. Yates III, T.N. Davis, Assigning function to yeast proteins by integration of technologies, Molec. Cell 12 (6) (2003) 1353–1365.
[46] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, R. Guthke, Gene regulatory network inference: data integration in dynamic models – a review, Biosystems 96 (1) (2009) 86–103.
[47] F. Hefti, W.J. Weiner, Nerve growth factor and alzheimer's disease, Ann. Neurol. 20 (3) (1986) 275–281.
[48] D.R. Hinton, A.A. Sadun, J.C. Blanks, C.A. Miller, Optic-nerve degeneration in alzheimer's disease, New England J. Med. 315 (8) (1986) 485–487.
[49] Q. Hu, Z. Wang, Z. Zhang, Fsim: a novel functional similarity search algorithm and tool for discovering functionally related gene products, BioMed Res. Int. (2014).
[50] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucl. Acids Res. 37 (1) (2009) 1–13.
[51] Y. Ichihashi, J.A. Aguilar-Martínez, M. Farhi, D.H. Chitwood, R. Kumar, L.V. Millon, J. Peng, J.N. Maloof, N.R. Sinha, Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape, Proc. Natl. Acad. Sci. 111 (25) (2014) E2616–E2621.
[52] K. Imai, K. Takai, Y. Nishigaki, S. Shimizu, T. Naiki, H. Hayashi, T. Uematsu, J. Sugihara, E. Tomita, M. Shimizu, Insulin resistance raises the risk for recurrence of stage i hepatocellular carcinoma after curative radiofrequency ablation in hepatitis c virus-positive patients: a prospective, case series study, Hepatol. Res. 40 (4) (2010) 376–382.
[53] K. Isshiki, Z. He, Y. Maeno, R.C. Ma, Y. Yasuda, T. Kuroki, G.S. White, M.E. Patti, G.C. Weir, G.L. King, Insulin regulates socs2 expression and the mitogenic effect of igf-1 in mesangial cells, Kidney Int. 74 (11) (2008) 1434–1443.
[54] J.C. Jeong, X. Chen, A new semantic functional similarity over gene ontology, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 12 (2) (2015) 322–334.
[55] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, Reactome: a knowledgebase of biological pathways, Nucl. Acids Res. 33 (Suppl. 1) (2005) D428–D432.
[56] R. Kalaria, D. Cohen, D. Premkumar, S. Nag, J. LaManna, W. Lust, Vascular endothelial growth factor in alzheimer's disease and experimental cerebral ischemia, Molec. Brain Res. 62 (1) (1998) 101–105.
[57] R.N. Kalaria, The role of cerebral ischemia in alzheimers disease, Neurobiol. Aging 21 (2) (2000) 321–330.
[58] M. Kanehisa, S. Goto, Kegg: kyoto encyclopedia of genes and genomes, Nucl. Acids Res. 28 (1) (2000) 27–30.
[59] I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, P.D. Karp, Ecocyc: a comprehensive database resource for escherichia coli, Nucl. Acids Res. 33 (Suppl. 1) (2005) D334–D337.
[60] H. Kim, J. Shin, E. Kim, H. Kim, S. Hwang, J.E. Shim, I. Lee, Yeastnet v3: a public database of data-specific and integrated functional gene networks for saccharomyces cerevisiae, Nucl. Acids Res. (2013) gkt981.
[61] M.-S. Lai, M.-S. Hsieh, Y.-H. Chiu, T.H.-H. Chen, Type 2 diabetes and hepatocellular carcinoma: a cohort study in high prevalence area of hepatitis virus infection, Hepatology 43 (6) (2006) 1295–1302.
[62] K. Laukens, S. Naulaerts, W.V. Berghe, Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis, Proteomics 15 (5–6) (2015) 981–996.
[63] B. Liu, M. Jin, P. Zeng, Prioritization of candidate disease genes by combining topological similarity and semantic similarity, J. Biomed. Inform. 57 (2015) 1–5.

[64] Y. Lu, Z. Li, X. Zhang, B. Ming, J. Jia, R. Wang, D. Ma, Retinal nerve fiber layer structure abnormalities in early alzheimer's disease: evidence in optical coherence tomography, Neuroscience Lett. 480 (1) (2010) 69–72.
[65] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, G. Stolovitzky, Wisdom of crowds for robust gene network inference, Nat. Methods 9 (8) (2012) 796–804.
[66] D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, G. Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference, Proc. Natl. Acad. Sci. 107 (14) (2010) 6286–6291.
[67] B.C. Martin, J.H. Warram, A. Krolewski, J. Soeldner, C. Kahn, R. Bergman, Role of glucose and insulin resistance in development of type 2 diabetes mellitus: results of a 25-year follow-up study, Lancet 340 (8825) (1992) 925–929.
[68] N. Matsumoto, S. Tamura, Y. Fujiki, The pathogenic peroxin pex26p recruits the pex1p–pex6p aaa atpase complexes to peroxisomes, Nat. Cell Biol. 5 (5) (2003) 454–460.
[69] G.K. Mazandu, E.R. Chimusa, N.J. Mulder, Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery, Brief. Bioinform. (2016) bbw067.
[70] G.K. Mazandu, N.J. Mulder, A topology-based metric for measuring term similarity in the gene ontology, Adv. Bioinform. (2012).
[71] G.K. Mazandu, N.J. Mulder, Information content-based gene ontology functional similarity measures: which one to use for a given biological data type?, PloS One 9 (12) (2014) e113859
[72] T. McCormack, O. Frings, A. Alexeyenko, E.L. Sonnhammer, Statistical assessment of crosstalk enrichment between gene groups in biological networks, PloS One 8 (1) (2013) e54945.
[73] G.M. McKhann, D.S. Knopman, H. Chertkow, B.T. Hyman, C.R. Jack, C.H. Kawas, W.E. Klunk, W.J. Koroshetz, J.J. Manly, R. Mayeux, The diagnosis of dementia due to alzheimers disease: recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease, Alzheimer's Dementia 7 (3) (2011) 263–269.
[74] H. Mi, S. Poudel, A. Muruganujan, J.T. Casagrande, P.D. Thomas, Panther version 10: expanded protein families and functions, and analysis tools, Nucl. Acids Res. 44 (D1) (2016) D336–D342.
[75] M. Mistry, P. Pavlidis, Gene ontology term overlap as a measure of gene functional similarity, BMC Bioinform. 9 (1) (2008) 1.
[76] M.J. Müller, O. Willmann, A. Rieger, A. Fenk, O. Selberg, H.U. Lautz, M. Bürger, H.J. Balks, A. Von Zur Mühlen, F.W. Schmidt, Mechanism of insulin resistance associated with liver cirrhosis, Gastroenterology 102 (6) (1992) 2033–2041.
[77] A. Nagar, H. Al-Mubaid, A hybrid semantic similarity measure for gene ontology based on offspring and path length, in: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2015, pp. 1–7.
[78] K.T. Nead, G. Gaskin, C. Chester, S. Swisher-McClure, J.T. Dudley, N.J. Leeper, N. H. Shah, Androgen deprivation therapy and future alzheimers disease risk, J. Clin. Oncol. 34 (6) (2015) 566–571.
[79] D. Nishimura, Biocarta, Biotech Softw. Internet Rep.: Comput. Softw. J. Sci. 2 (3) (2001) 117–120.
[80] C. Olsen, K. Fleming, N. Prendergast, R. Rubio, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains, J. Quackenbush, Inference and validation of predictive gene networks from biomedical literature and gene expression data, Genomics 103 (5) (2014) 329–336.
[81] N. Omranian, J.M. Eloundou-Mbebi, B. Mueller-Roeber, Z. Nikoloski, Gene regulatory network inference using fused lasso on multiple data sets, Sci. Rep. 6 (2016).
[82] M. Pathan, S. Keerthikumar, C.-S. Ang, L. Gangoda, C.Y. Quek, N.A. Williamson, D. Mouradov, O.M. Sieber, R.J. Simpson, A. Salim, Funrich: an open access standalone functional enrichment and interaction network analysis tool, Proteomics 15 (15) (2015) 2597–2601.
[83] J. Peng, H. Li, Q. Jiang, Y. Wang, J. Chen, An integrative approach for measuring semantic similarities using gene ontology, BMC Syst. Biol. 8 (Suppl. 5) (2014) S8.
[84] J. Peng, S. Uygun, T. Kim, Y. Wang, S.Y. Rhee, J. Chen, Measuring semantic similarities by combining gene ontology annotations and gene co-function networks, BMC Bioinform. (2015).
[85] A. Pesaranghader, S. Matwin, M. Sokolova, R.G. Beiko, simdef: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes, Bioinformatics (2015) btv755.
[86] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, F.M. Couto, Semantic similarity in biomedical ontologies, PLoS Comput. Biol. 5 (7) (2009) e1000443+.
[87] K.M. Ramos, Y. Jiang, C.I. Svensson, N.A. Calcutt, Pathogenesis of spinally mediated hyperalgesia in diabetes, Diabetes 56 (6) (2007) 1569–1576.
[88] V. Ratziu, L. Bonyhay, V. Di Martino, F. Charlotte, L. Cavallaro, M.-H. Sayegh-Tainturier, P. Giral, A. Grimaldi, P. Opolon, T. Poynard, Survival, liver failure, and hepatocellular carcinoma in obesity-related cryptogenic cirrhosis, Hepatology 35 (6) (2002) 1485–1493.
[89] B. Ristevski, A survey of models for inference of gene regulatory networks, Nonlinear Anal.: Modell. Control 18 (4) (2013) 444–465.
[90] A.A. Sadun, C.J. Bassi, Optic nerve damage in alzheimer's disease, Ophthalmology 97 (1) (1990) 9–17.
[91] G. Sales, E. Calura, D. Cavalieri, C. Romualdi, graphite – a Bioconductor package to convert pathway topology to gene network, BMC Bioinform. 13 (1) (2012) 20+.
[92] C.F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, K.H. Buetow, Pid: the pathway interaction database, Nucl. Acids Res. 37 (Suppl. 1) (2009) D674–D679.
[93] O. Selberg, W. Burchert, J. Vd Hoff, G.J. Meyer, H. Hundeshagen, E. Radoch, H. Balks, M. Müller, Insulin resistance in liver cirrhosis: positron-emission tomography scan analysis of skeletal muscle glucose metabolism, J. Clin. Invest. 91 (5) (1993) 1897.
[94] D.J. Selkoe, Alzheimer's disease is a synaptic failure, Science 298 (5594) (2002) 789–791.
[95] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.
[96] J. Sławek, T. Arodź, Ennet: inferring large gene regulatory networks from expression data using gradient boosting, BMC Syst. Biol. 7 (1) (2013) 1.
[97] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (11) (2007) 1251–1255.
[98] B.Q. Starley, C.J. Calcagno, S.A. Harrison, Nonalcoholic fatty liver disease and hepatocellular carcinoma: a weighty connection, Hepatology 51 (5) (2010) 1820–1832.
[99] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. 102 (43) (2005) 15545–15550.
[100] M.C. Teixeira, P.T. Monteiro, J.F. Guerreiro, J.P. Gonçalves, N.P. Mira, S.C. dos Santos, T.R. Cabrito, M. Palma, C. Costa, A.P. Francisco, The yeastract database: an upgraded information system for the analysis of gene and genomic transcription regulation in saccharomyces cerevisiae, Nucl. Acids Res. (2013) gkt1015.
[101] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, P. Xuan, Measuring gene functional similarity based on group-wise comparison of go terms, Bioinformatics (2013) btt160.
[102] C.S. Tsai, R. Ritch, B. Schwartz, S.S. Lee, N.R. Miller, T. Chi, F. Hsieh, Optic nerve head and nerve fiber layer in alzheimer's disease, Arch. Ophthalmol. 109 (2) (1991) 199–204.
[103] B.J. Veldt, W. Chen, E.J. Heathcote, H. Wedemeyer, J. Reichen, W.P. Hofmann, R.J. de Knegt, S. Zeuzem, M.P. Manns, B.E. Hansen, Increased risk of hepatocellular carcinoma among patients with hepatitis c cirrhosis and diabetes mellitus, Hepatology 47 (6) (2008) 1856–1862.
[104] N.X. Vinh, M. Chetty, R. Coppel, P.P. Wangikar, Issues impacting genetic network reverse engineering algorithm validation using small networks, Biochim. Biophys. Acta (BBA)-Proteins Proteom. 1824 (12) (2012) 1434–1441.
[105] C. Wang, X. Wang, G. Gong, Q. Ben, W. Qiu, Y. Chen, G. Li, L. Wang, Increased risk of hepatocellular carcinoma in patients with diabetes mellitus: a systematic review and meta-analysis of cohort studies, Int. J. Cancer 130 (7) (2012) 1639–1648.
[106] P. Wang, D. Kang, W. Cao, Y. Wang, Z. Liu, Diabetes mellitus and risk of hepatocellular carcinoma: a systematic review and meta-analysis, Diabetes/Metabol. Res. Rev. 28 (2) (2012) 109–122.
[107] C. Weyer, C. Bogardus, D.M. Mott, R.E. Pratley, The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus, J. Clin. Invest. 104 (6) (1999) 787–794.
[108] X. Wu, R. Jiang, M.Q. Zhang, S. Li, Network-based global inference of human disease genes, Molec. Syst. Biol. 4 (1) (2008) 189.
[109] X. Wu, E. Pang, K. Lin, Z.-M. Pei, Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge- and ic-based hybrid method, PLOS ONE (May) (2013).
[110] H. Xu, Y.-S. Ang, A. Sevilla, I.R. Lemischka, A. Ma'ayan, Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells, PLoS Comput. Biol. 10 (8) (2014) e1003777.
[111] Y. Xu, M. Guo, W. Shi, X. Liu, C. Wang, A novel insight into gene ontology semantic similarity, Genomics 101 (6) (2013) 368–375.
[112] S.-B. Zhang, J.-H. Lai, A hybrid measure for the semantic similarity of gene ontology terms, in: 2014 2nd International Conference on Systems and Informatics (ICSAI), IEEE, 2014, pp. 911–916.
[113] S.-B. Zhang, J.-H. Lai, An integrated information-based similarity measurement of gene ontology terms, Comput. Sci. Inform. Syst. 12 (4) (2015).
[114] S.-B. Zhang, Q.-R. Tang, Protein–protein interaction inference based on semantic similarity of gene ontology terms, J. Theor. Biol. 401 (2016) 30–37.

Capítulo 6

GNC–app: A new Cytoscape app to rate gene networks biological coherence using gene–gene indirect relationships

BioSystems 166 (2018) 61–65

Contents lists available at ScienceDirect

BioSystems

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

Short Communication

GNC–app: A new Cytoscape app to rate gene networks biological coherence using gene–gene indirect relationships☆

Juan J. Díaz-Montaña*, Francisco Gómez-Vela*, Norberto Díaz-Díaz

Intelligent Data Analysis (DATAi), Division of Computer Science, Pablo de Olavide University, ES-41013 Seville, Spain

ARTICLE INFO

Article history:
Received 13 August 2017
Received in revised form
22 December 2017
Accepted 27 January 2018
Available online 14 February 2018

Keywords:
Cytoscape
Gene networks
Gene networks validation
Gene networks analysis

ABSTRACT

Motivation: Gene networks are currently considered a powerful tool to model biological processes in the Bioinformatics field. A number of approaches to infer gene networks and various software tools to handle them in a visual simplified way have been developed recently. However, there is still a need to assess the inferred networks in order to prove their relevance.

Results: In this paper, we present the new GNC-app for Cytoscape. GNC-app implements the GNC methodology for assessing the biological coherence of gene association networks and integrates it into Cytoscape. Implemented de novo, GNC-app significantly improves the performance of the original algorithm in order to be able to analyse large gene networks more efficiently. It has also been integrated in Cytoscape to increase the tool accessibility for non-technical users and facilitate the visual analysis of the results. This integration allows the user to analyse not only the global biological coherence of the network, but also the biological coherence at the gene–gene relationship level. It also allows the user to leverage Cytoscape capabilities as well as its rich ecosystem of apps to perform further analyses and visualizations of the network using such data.

Availability: The GNC-app is freely available at the official Cytoscape app store: http://apps.cytoscape.org/apps/gnc.

1. Introduction

Gene networks are one of the most used tools to model biological processes in Bioinformatics. In the literature it is possible to find several works in this sense like the works presented by Gallo et al. (2011) or the recent work by Gómez-Vela et al. (2016). Although the assessment of the inferred gene networks is a crucial step in any study, usually it is not easily performed (Dougherty, 2010) and is considered an unsolved problem.

Network analysis platforms such as Cytoscape (Shannon et al., 2003), allow researchers to develop apps that provide visual and easy-to-use means to analyse gene networks and their biological coherence, thus, simplifying the network validation and analysis. Although there are existing Cytoscape apps that perform various gene networks analyses (Spinelli et al., 2013; Tang et al., 2015; Díaz-Montaña et al., 2017), there are none that give a direct comparison of the input network with a reference network in order to offer a global evaluation of the biological coherence of the input network. The biological coherence in this context is defined as the correctness of the information encoded by the input network based on the information encoded by the reference network (Gómez-Vela et al., 2015). Moreover, there are no existing apps that use the indirect relationships that are present in the networks to perform the analysis. It is worth mentioning that the use of indirect relationships offers a fairer and more reliable validation of gene networks.

In this paper, we present GNC-app for Cytoscape, which is an improved implementation of the GNC methodology presented in the work by Gómez-Vela et al. (2015). GNC is able to use any biological database (presented as a network) to assess the biological coherence of a given gene network. GNC–app significantly improves the original algorithm performance so that large gene networks are analysed more efficiently. It has also been integrated in Cytoscape to increase the tool accessibility for non-technical users and facilitate the visual analysis of the results. This integration allows the user to analyse not only the global biological coherence of the network but also the biological coherence at the gene–gene relationship level. It also allows the user to leverage the capabili-

☆ The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

* Corresponding authors.
E-mail addresses: jjdiamon@alumno.upo.es (J.J. Díaz-Montaña), fgomez@upo.es (F. Gómez-Vela), ndiaz@upo.es (N. Díaz-Díaz).

https://doi.org/10.1016/j.biosystems.2018.01.007
0303-2647/

ties of Cytoscape as well as its rich ecosystem of apps to perform further analyses and visualizations of the network using such data.

2. Software description

2.1. App overview

GNC-app is an implementation of GNC methodology (Gómez-Vela et al., 2015) that extends the functionality of Cytoscape to perform an evaluation of gene networks using the indirect relationships.

The app allows the user to select any gene network as reference network. To do so, a configuration dialogue offers the user the possibility of choosing either a pre-loaded network such as BioGRID (Stark et al., 2006) or YeastNet (Kim et al., 2013), or any other biological database in the form of a network to perform the evaluation. GNC considers both the input and reference networks as association networks (undirected and unweighted). Thus, the direction, type and weight of the edges in the networks are ignored if they exist.

After GNC execution has been completed, the information is shown in Cytoscape as can be seen in Fig. 1.

The "Results Panel" offers a global analysis of the input network. It shows, not only the biological coherence value given by GNC for the input network, but also other well-known measures for the evaluation of networks, specifically PPV (Dougherty, 2010) and *F*-measure (Powers, 2011). Additionally, it offers information about the input and output networks (nodes, edges, density and common nodes). The results described in this panel allow the user to assess the overall biological coherence of the network under study.

Furthermore, this new implementation offers, not only an overall assessment of the input network but also detailed information about the genes and their relationships (edges). The "Table Panel" provides specific information on each relationship of the input network. In this panel, the information of each relationship is enriched with the relationship's biological coherence value obtained from GNC (see Fig. 1), each node is marked by whether they are known or not in the biological database used for the analysis and the network information is enriched by the global information as displayed in the "Results Panel". This data integration allows the users to leverage Cytoscape analysis functionality like filtering, styling and other useful functions in order to, for example, sort relations according to their biological relevance according to GNC, filter relationships with low biological coherence or filter nodes that are not present in the biological database used. This information can also be used by any of the Cytoscape's apps for more complex analyses.

The same information is also displayed in the "network view" to facilitate visual exploration and analysis.

GNC-app styles the "network view" so nodes that are not present in the database used to carry out the network validation are indicated in pink while nodes that are common in both, input network and database, maintain the original blue (see Fig. 1). This facilitates the quick identification of nodes for which there is no relative information in the database. These nodes can be considered candidates for new relationships to be studied whose information is not yet known in the database, or nodes that are not relevant to the analysis of the network. It is worth to mention that these genes are also used to compute the distance matrices and not discarded by GNC (see original paper of GNC for more details). Moreover, edges are coloured based on the biological coherence value obtained by GNC between the genes at its edges. Edges related to genes that are not in the biological database are indicated in pink, relationships between genes that are directly related in both the input network and the biological database are highlighted in green and the rest of the relationships are coloured in grey scale where a darker colour represents higher biological coherence (see Fig. 1).

Finally, it is worth mentioning that the app also allows the user to download the coherence matrix of the input network calculated

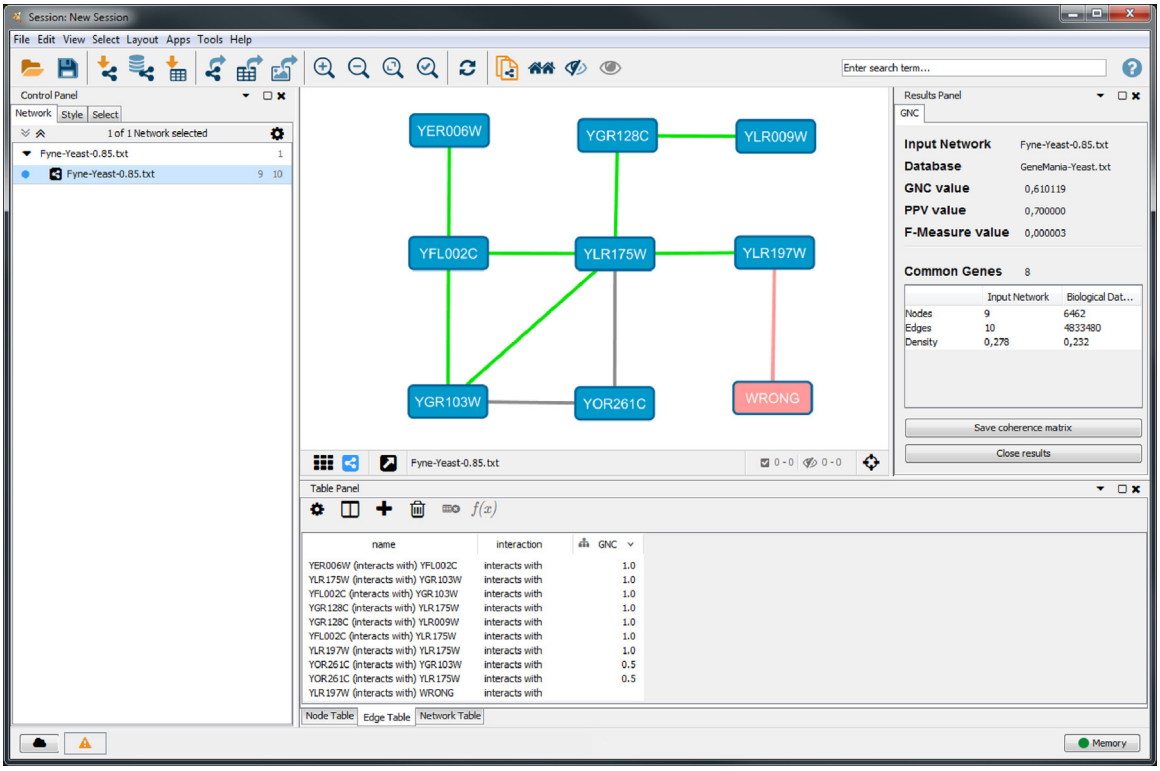
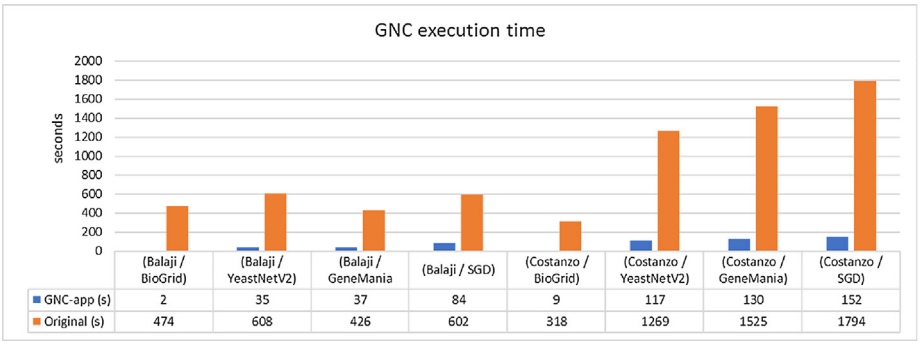


Fig. 1. Screenshot of Cytoscape showing the results obtained after executing GNC-app on the FyNet network using GeneMANIA as biological database. A non-existing node labelled as "WRONG" was added to the original FyNet network to show the example of use.



Network	Nodes	Edges
Balaji	441	12873
Contanzo	4273	74840
BioGrid	2554	22651
YeastNetV2	5436	102456
GeneMania	5514	125849
SGD	5726	213673

Fig. 2. Comparison of the execution times of GNC-app and the original implementation on several known networks (Balaji and Constanzo) and databases (Biogrid, YeastNet, GeneMANIA and SGD).

by GNC in.csv format. The coherence matrix is a matrix containing one column and one row per common gene between the input and reference networks. The value of each cell corresponds to the coherence between the nodes corresponding to that particular column and row as calculated by GNC. This file allows the user to perform further analysis of GNC results using external tool.

Therefore, these presented characteristics show the relevance of this new app for the study, validation and analysis of genetic association networks.

2.2. Implementation

The original implementation of GNC considered the networks as gene association networks, calculated their adjacency matrices, and then used Floyd–Warshall ($\Theta(V^3)$ where V is the number of vertices in the network) to calculate the distance matrices. Because gene association networks are unweighed and undirected, the Floyd–Warshall algorithm could be replaced by Breadth-First (Moore, 1959) search on each node ($\Theta(E+V)$ where V and E are the number of vertices and edges in the network respectively). Breadth-First search removes the need of calculating the adjacency matrix and can be executed directly on Cytoscape native networks. Furthermore, apart from its lower complexity, there are some characteristics of Breadth-First search that make it better suited for gene networks than Floyd–Warshall. Gene networks usually present small-world characteristics; i.e. most nodes are not neighbours of one another, but the neighbours of any given node are likely to be neighbours of each other and most nodes can be reached from every other node by a small number of hops or steps. Since Breadth-First search is based on hoping from one node to its neighbours the analysis runtime is reduced. Also, in many cases the biological database might be significantly larger than the input network. Breadth-First splits the all-shortest-paths problem into an independent problem for each node to calculate the shortest paths between itself and all the other nodes. GNC only requires the distances between nodes that are present in both the input network and the biological database and the new implementation takes advantage of this to skip every unnecessary distance calculation. GNC-app also utilizes heavy caching in order to calculate the common nodes faster than the original version (from $\Theta(V1\,V2)$ to $\Theta(V1\,log(V2))$ in the average case and $\Theta(V1)$ in the best case) and speed up the coherence calculations in most cases (from $\Theta(V1\,V2)$ to $\Theta(\cap(V1,\ V2)^2)$). With all this improvements, GNC complexity is reduced from $\Theta(V1^3)+\Theta(V2^3)+\Theta(V1\,V2)$ to $\Theta(E1+V1)+\Theta(E2+V2)+\Theta(\cap(V1,\ V2)^2)$ where $V1$ and $E1$ are the number of vertices and edges in the input network and $V2$ and $E2$ are the number of vertices and edges in the reference network. This complexity includes calculating the distance matrix for both the input and reference network and evaluating every edge existing between common nodes in both networks.

Furthermore, modern processors normally include several cores. This allows programmers to parallelize algorithms using multiple threads. Using multiple cores can reduce the runtime of the analysis by roughly the number of cores available; i.e. having four cores should speed up the analysis to take one quarter of the time taken in a single core. Breadth-First search is well designed for paralellization because, as mentioned above, it splits the all-shortest-paths problem into a set of independent problems, one for each node. Also, calculating the biological coherence between each pair of genes can be parallelized. GNC–app is designed to distribute its workload among all the available cores in the computer and leverage hyper-threading if available. Furthermore, this approach could be extended to use distributed approaches like map-reduce (Dean and Ghemawat, 2008) or Spark (Zaharia et al., 2010) if the need arises, for example to do genome-wide analysis.

GNC-app also includes two other measures as was described before: PPV and F-measure. The original implementation calculated each element in the confusion matrix separately by looping over all the edges in the input network and the reference network ($\Theta(E1\,E2)$). The new implementation calculates all the elements in the confusion matrix in one pass over the networks and ensures that the same edge is never visited twice once processed ($\Theta(E1\,log(E2))$ in the average case and $\Theta(E1)$ in the best case). It also uses parallelism as explained above.

Finally, the complexity of GNC-app combines the complexity of GNC and the other measures resulting on and average case complexity of $\Theta(E1+V1)+\Theta(E2+V2)+\Theta(\cap(V1,\ V2)^2)+\Theta(E1)$. This is noticeably better than the original implementation ($\Theta(V1^3)+\Theta(V2^3)+\Theta(V1^3)+\Theta(V2^3)+\Theta(V1\,V2)+\Theta(2\,E1\,E2)$). It should also be noted that not only the complexity is lower, but also the analysis runtime is reduced by the optimizations mentioned above and the use of multiple cores if they are available.

Fig. 2 shows a comparison of the average execution time between the old implementation and the new implementation performing the analysis presented in Gómez-Vela et al. (2015). Specifically, the networks presented by Balaji et al. (2006) and Costanzo et al. (2010) were analysed using Biogrid, YeastNet, GeneMANIA and the *Saccharomyces cerevisiae* Genome Database (SGD) as biological databases (see Gómez-Vela et al. (2015) for more details). The executions were performed using commodity hardware, a MacBook Pro from early 2011 with a 1.2 GHz processor with 4 cores, hyper-threading enabled and 8 Gbs of memory.

As it is depicted in Fig. 2, GNC-app was consistently an order of magnitude faster than the original GNC implementation. It is worth to mention the improvement obtained for the experiment with Constanzo's network where the GNC-app is able to perform the analysis significantly faster than the original implementation. This allows researchers to obtain results faster and improves their iteration time when doing exploratory analysis. Also, more impor-

tantly, it allows the analysis of large networks in a reasonable time using commodity hardware such as those used in our analysis. This is especially relevant considering that gene networks and biological databases tend to be large, making the previous implementation unusable after a certain network-size because of the excessive time used by the old algorithm to complete the analysis.

3. Application example

In order to demonstrate its usefulness, GNC-app was used to evaluate the coherence of a fuzzy gene association network considering the information stored in GeneMANIA (Warde-Farley et al., 2010) (see Supplementary material). GeneMANIA provides a composite gene–gene functional interaction network where genes (nodes) are related (edges) if at least one piece of evidence of this relation exists in the literature.

The selected input network was extracted using the FyNet (Gómez-Vela et al., 2016) algorithm, a fuzzy approach for modelling gene association networks combining gene co-expression and biological knowledge from Gene Ontology. More specifically, FyNet was applied to the well-known yeast cell-cycle microarray (Spellman et al., 1998) and the largest subnetwork with 0.85 as α-cut was selected. We have selected a yeast network as example, because it is the most representative organism in Bioinformatics studies. A non-existing node labelled as "WRONG" was added to the network in order to showcase how nodes that are not present in the reference network are treated.

In order to perform the analysis, FyNet network (provided as Supplemental material) needs to be imported into Cytoscape and selected as current network (Menu "Import" ⟶ "Network" ⟶ "File" of Cytoscape, see Cytoscape's documentation for more details).

Then, GNC-App needs to be installed in Cytoscape from the App Store. This will create an item in the "Apps" menu called GNC which opens the configuration panel for GNC analysis execution. The configuration dialogue allows the user to select a predefined reference networks or upload a custom one. GeneMANIA's reference network (also provided as Supplemental material) can be uploaded and selected by using the "Load custom database" button (see Fig. 3 for more details).

Finally, the analysis is launched by clicking on the "Run GNC" button. After a while the results of the analysis will be presented in the Results panel on the right.

Fig. 1 shows the results generated by GNC-app as they are presented in Cytoscape once the analysis completes. As mentioned in the previous section, the "Results Panel" shows the overall gene network coherence and other general information while the "Table Panel" and the "Network View" show each gene–gene relationship coherence and other gene/relationship-specific information.

The "Results Panel" shows that the input network coherence according to GNC is 0.61 while the values of PPV and *F*-measure are 0.78 and 0.000003, respectively. PPV yielded a high value because it does not consider the true and false negatives (TNs and FNs), i.e. it only considers the edges from the input network that are present in GeneMANIA. *F*-measure, on the other hand, considers true and false negatives and yields a very low value, i.e. there are a significant number of true and false negatives. GNC gives a more conservative coherence value than PPV but much higher than *F*-measure. This is because GNC considers indirect relationships and, although the relationships between some genes might not be directly present in GeneMANIA, they are present as indirect relationships through some intermediary gene(s). The missing direct relationship can be considered as a candidate to be studied whose information is not yet known in the database or as an incorrect relationship.

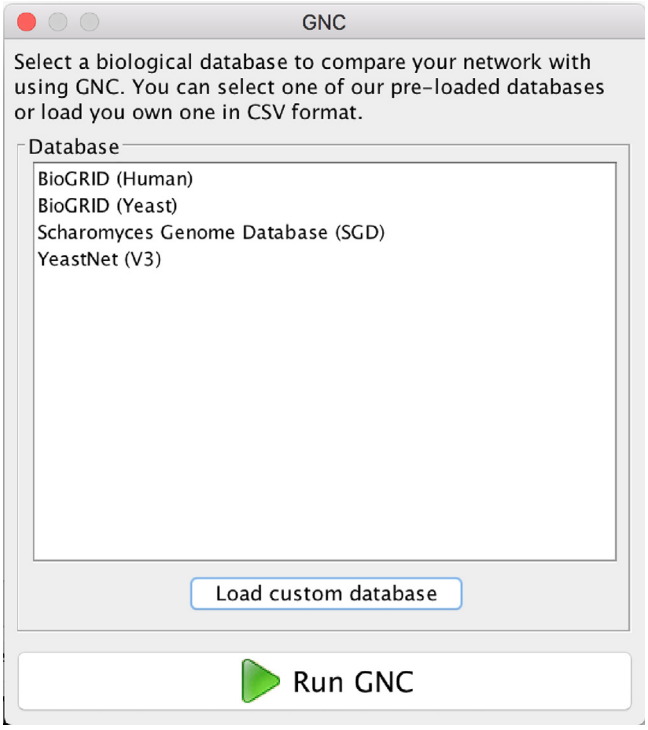


Fig. 3. GNC-App configuration panel. This panel allows the user to select a configure and run GNC-App.

Visual inspection of the "network view" reveals that all the genes in the network are present in GeneMANIA except for the one labelled as "WRONG" which, in this case, can be discarded. It also allows the user to quickly differentiate between relationships that are direct (green) or indirect (grey scale) in GeneMANIA. For example, the relationship between YLR175W and YGR103W is coloured in green which is correct since it is presented in Hong et al. (2011) as co-expression evidence. Indirect relationships' colour reveals at a glance whether the nodes are separated just by fewer genes (darker) or by many (lighter). An example of this type of case is the relationship between YOR261C and YGR103W which is interconnected by YAL005C in GeneMANIA. This is correct since evidence of YOR261C-YAL005C and YAL005C-YGR103W was presented by Myers et al. (2005).

Visual inspection is of course limited to small networks (or subnetworks). For larger networks, more automated analysis can be performed by using Cytoscape's core capabilities since the "Table Panel" contains all the information generated by GNC. This includes the general network values, each gene–gene relationship coherence and a flag indicating whether or not a gene in the input network was present in the biological database. The standard filters of Cytoscape can be used to select only the unknown genes and their neighbours or filter the edges with low coherence. Once the network is reduced to a more manageable size, visual analysis can be applied again. It should be noted that this information is also accessible by any Cytoscape app, giving the user a much wider range of analytic tools than what Cytoscape offers out-of-the-box.

4. Conclusion

In this paper, a new Cytoscape app to perform gene network analysis is presented. This is the first app in the Cytoscape ecosystem that validates a network against any reference network and the utilization of the indirect relationships presented in the networks. The app, which is an improved implementation of the GNC methodology, is able to more efficiently perform analyses of larger

gene networks due to the dramatic reduction of computational cost. Moreover, the relevance of the app is furthered with the integration of the information computed by GNC into Cytoscape. This new information combined with Cytoscape's core features and its rich ecosystem of apps can be used to carry out a more exhaustive analysis of the input network, such as filtering or redefining the input network to obtain networks with more biological meaning.

The capabilities of GNC-app are showed by analysing a fuzzy gene association network using GeneMANIA as a biological database (reference network).

Finally, the app and a complete tutorial to use it are freely available on the official Cytoscape app store: http://apps.cytoscape.org/apps/gnc.

References

Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M., Aravind, L., 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J. Mol. Biol. 360 (1), 213–227.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., et al., 2010. The genetic landscape of a cell. Science 327 (5964), 425–431.

Díaz-Montaña, J.J., Díaz-Díaz, N., Gómez-Vela, F., 2017. GFD-Net: a novel semantic similarity methodology for the analysis of gene networks. J. Biomed. Inform. 68, 71–82.

Dean, J., Ghemawat, S., 2008. Mapreduce: simplified data processing on large clusters. Commun. ACM 51 (1), 107–113.

Dougherty, E.R., 2010. Validation of gene regulatory networks: scientific and inferential. Brief. Bioinform. 12 (3), 245–252.

Gómez-Vela, F., Lagares, J.A., Díaz-Díaz, N., 2015. Gene network coherence based on prior knowledge using direct and indirect relationships. Comput. Biol. Chem. 56, 142–151.

Gómez-Vela, F., Barranco, C.D., Díaz-Díaz, N., 2016. Incorporating biological knowledge for construction of fuzzy networks of gene associations. Appl. Soft Comput. 42 (C), 144–155.

Gallo, C.A., Carballido, J.A., Ponzoni, I., 2011. Discovering time-lagged rules from microarray data using gene profile classifiers. BMC Bioinform. 12 (1), 123.

Hong, K.-K., Vongsangnak, W., Vemuri, G.N., Nielsen, J., 2011. Unravelling evolutionary strategies of yeast for improving galactose utilization through integrated systems level analysis. Proc. Natl. Acad. Sci. U. S. A. 108 (29), 12179–12184.

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J.E., Lee, I., 2013. Yeastnet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. Nucleic Acids Res. 42 (D1), D731–D736.

Moore, E.F., 1959. The shortest path through a maze. Proc. Int. Symp. Switching Theory, 285–292.

Myers, C.L., Robson, D., Wible, A., Hibbs, M.A., Chiriac, C., Theesfeld, C.L., Dolinski, K., Troyanskaya, O.G., 2005. Discovery of biological networks from diverse functional genomic data. Genome Biol. 6 (13), R114.

Powers, D.M.W., 2011. Evaluation: from precision, recall and *F*-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2 (1), 37–63.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle, regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9 (12), 3273–3297.

Spinelli, L., Gambette, P., Chapple, C.E., Robisson, B., Baudot, A., Garreta, H., Tichit, L., Guénoche, A., Brun, C., 2013. Clust&see: a Cytoscape plugin for the identification, visualization and manipulation of network clusters. BioSystems 113 (2), 91–95.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., 2006. Biogrid: a general repository for interaction datasets. Nucleic Acids Res. 34 (Suppl. 1), D535–D539.

Tang, Y., Li, M., Wang, J., Pan, Y., Wu, F.-X., 2015. Cytonca: a Cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems 127, 67–72.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al., 2010. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38 (Suppl. 2), W214–W220.

Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., 2010. Spark: cluster computing with working sets. HotCloud 10 (10-10), 95.

Capítulo 7

Development and use of a Cytoscape app for GRNCOP2

Computer Methods and Programs in Biomedicine 177 (2019) 211–218

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Development and use of a Cytoscape app for GRNCOP2

Juan J. Díaz–Montaña[a,*], Norberto Díaz–Díaz[a], Carlos D. Barranco[a], Ignacio Ponzoni[b]

[a] Intelligent Data Analysis (DATAi), Division of Computer Science, Pablo de Olavide University, Seville ES-41013, Spain
[b] Instituto de Ciencias e Ingeniería de la Computación (UNS, CONICET), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 17 September 2018
Revised 5 May 2019
Accepted 29 May 2019

Keywords:
Machine learning
Gene regulatory networks
Cytoscape
Pathway crosstalk
Alzheimer's disease

ABSTRACT

Background and Objective: Gene regulatory networks (GRNs) are essential for understanding most molecular processes. In this context, the so-called model-free approaches have an advantage modeling the complex topologies behind these dynamic molecular networks, since most GRNs are difficult to map correctly by any other mathematical model. Abstract model-free approaches, also known as rule-based extraction methods, offer valuable benefits when performing data-driven analysis; such as requiring the least amount of data and simplifying the inference of large models at a faster analysis speed. In particular, GRNCOP2 is a combinatorial optimization method with an adaptive criterion for the discretization of gene expression data and high performance, in contrast to other rule-based extraction methods for discovering GRNs. However, the analysis of the large relational structures of the networks inferred by GRNCOP2 requires the support of effective tools for interactive network visualization and topological analysis of the extracted associations. This need motivated the possibility of integrating GRNCOP2 in the Cytoscape ecosystem in order to benefit from Cytoscapes core functionality, as well as all the other apps in its ecosystem.

Methods: In this paper, we introduce the implementation of a GRNCOP2 Cytoscape app. This incorporation to Cytoscape platform includes new functionality for GRN visualizations, dynamic user-interaction and integration with other apps for topological analysis of the networks.

Results: In order to demonstrate the usefulness of integrating GRNCOP2 in Cytoscape, the new app was used to tackle a novel use case for GRNCOP2: the analysis of crosstalk between pathways. In this regard, datasets associated with Alzheimer's disease (AD) were analyzed using GRNCOP2 app and other apps of the Cytoscape ecosystem by performing a topological analysis of the AD progression and its synchronization with the Ubiquitin Mediated Proteolysis pathway. Finally, the biological relevance of the findings achieved by this new app were evaluated by searching for evidence in the literature.

Conclusions: The proposed crosstalk analysis with the new GRNCOP2 app focused on assessing the phase of the Alzheimer's disease progression where the coordination with the Ubiquitin Mediated Proteolysis pathway increase, and identifying the genes that explain the signalling between these cellular processes. Both questions were explored by topological contrastive analysis of the GRNs generated for the GRNCOP2 app, where several facilities of Cytoscape were exploited. The topological patterns inferred by this new App have been consistent with biological evidence reported in the scientific literature, illustrating the effectiveness of using this new GRNCOP2 App in pathway analysis.

Availability: The GRNCOP2 App is freely available at the official Cytoscape app store: http://apps.cytoscape.org/apps/grncop2

1. Introduction

The inference of gene regulatory networks (GRNs) is a computational procedure for assessing the interactions within a cellular system from experimental data [1,2]. These molecular interactions can be extracted from different kinds of data, such as DNA microarray, RNA-seq and proteomics. Therefore, the biological interpretation of the relationships in these networks is related to the semantic behind the data used to discover them [3]. In particular, several computational methods have been proposed for inferring GRNs from gene expression data, where the interactions may preferably indicate transcription regulation, but can also correspond to some protein-protein interactions [4]. Therefore, the semantic of these gene associations, which ensures a meaningful bi-

* Corresponding author.
E-mail addresses: jjdiamon@alumno.upo.es (J.J. Díaz–Montaña), ndiaz@upo.es (N. Díaz–Díaz), cbarranco@upo.es (C.D. Barranco), ip@cs.uns.edu.ar (I. Ponzoni).

https://doi.org/10.1016/j.cmpb.2019.05.030
0169-2607/

ological interpretation, can be useful for understanding the normal cell physiology, but also complex pathological phenotypes [5].

In this context, data mining approaches based on association rules extraction techniques are suitable methods to reverse engineer these relational networks [6,7]. In general terms, an Association Rule (AR) establishes a causal link between two or more variables, where the semantics and the interpretation of the rule rely on the input data and the learning strategies employed to infer the association. ARs have been extensively used to extract knowledge from large datasets [8,9]. In gene expression analysis, these methods can be used to reveal biologically relevant associations among genes, at diverse environmental conditions or time point observations [10–12].

In particular, GRNCOP [13] and GRNCOP2 [10] belong to this family of AR extraction methods. GRNCOP uses combinatorial optimization and machine learning for inferring gene pairwise associations as classifiers. The most relevant feature of GRNCOP is that does not assume arbitrary nor uniform gene expression value discretization for the putative transcription factors (TFs). The thresholds are computed dynamically by using the same continuous-valued attribute discretization techniques as those applied for classification algorithms based on decision trees. Then, each pair of genes is evaluated and an AR with a particular accuracy based on an objective function is extracted. Finally, only the rules that achieved an accuracy value over a preselected threshold are reported. Later, in Gallo et al. [10], GRNCOP2 was proposed as an extension of the former method adding several improvements as adaptive discretization of target genes, inference of ARs with multiple time delay and a consensus strategy for extracting rules from multiple data sources. Gallo et al. also presented a detailed analysis of the performance of GRNCOP and GRNCORP2, where the latter achieved the most stable results. Besides, these strategies have been contrasted with other ARs extraction methods proposed for GRNs inference in Gomez-Vela et al. [14]. In that work, both algorithms were successfully compared with Soinovs and Bulashevskas approaches [15,16] using a coherence network metric computed from different biological databases, showing that GRNCOP and GRNCOP2 are competitive AR inference techniques for GRN discovery.

In this paper, the integration of GRNCOP2 as a Cytoscape [17] app is presented. Cytoscape is a software platform for the visualization and analysis of networks, specializing in biological networks. It provides a user-friendly interface which allows users with limited software programming knowledge to use complex algorithms and computational techniques. It also has a wide ecosystem of apps developed by the research community. Therefore, this contribution provides the user with the opportunity to easily execute GRNCOP2 over existing datasets and then visualize and analyze the inferred gene networks using any app from Cytoscapes app store. Furthermore, a novel case of use for GRNCOP2, crosstalk pathway analysis, is introduced in order to showcase the benets of the Cytoscape platform. Detecting the pathways that are significantly impacted by a particular condition or phenotype play a key role in understanding many complex biological phenomena [18]. These coordinated behaviors between two biological processes are usually known as pathway crosstalk and can be inferred from gene expression data [19–21]. Finally, we hope that the large user base of Cytoscape and its apps gives higher visibility to GRNCOP2 within the research community.

2. Methods

GRNCOP2 is a model-free combinatorial optimization method conceived to infer putative GRNs by extracting association rules (ARs) represented as classifiers. In order to discover the ARs, two different types of discretization are defined in this algorithm. The first one is to set the state of each target gene, and it is denoted as Target Discretization Threshold (TDT), while the second one is used for evaluating the potential interaction between each pair of genes and it is calculated in an adaptive gene-pairwise specific fashion. This last discretization is denoted as Relative Regulation Threshold (RRT).

The ARs inferred by GRNCOP2, called time-lagged rules, represent the situation in which the state of a $gene_i$ in a time-point j depends on the gene expression values of other genes in the previous t experimental condition (time-point) $j-w$, where w is a non-negative integer value representing the time-delay in the relation. The syntax of the rules is: $<symbol><gene_r>w\rightarrow<symbol><gene_i>$, where $gene_r$ and $gene_i$ stand for gene regulator and gene target respectively. The symbol $+$ $(-)$ on the left side of the rule indicates above (below) some RRT for gener w.r.t. genei, whereas the symbol $+$ $(-)$ on the right side of the rule indicates upregulated (downregulated) state, depending on the TDT for the $gene_i$. For example, the rule $+/-CLB23\rightarrow+/-CLB3$ denotes that, if gene CLB2 is above its RRT in relation to gene CLB3 in a time-point j, then CLB3 will be upregulated in the time-point j+3 and, if CLB2 is below or equal to $t_{CLB2,CLB3}$ in a sample j, then CLB3 will be downregulated in the sample j+3. In contrast with GRNCOP, this notation allows the representation of both simultaneous and time-lagged rules spanned in any unit of time-interval, which constitutes the kind of rules that GNRCOP2 is capable of inferring. GRNCOP2 infers the association rules described above by exploring the possible combinations of interactions between each pair of genes. In this sense, six particular cases are assumed, which are represented by the non-null integer numbers between -3 and 3, and a special case that indicates the absence of any relation represented by the number 0. All of these cases are described in Table 1.

Table 1
Types of rules inferred by GRNCOP2.

Rule type	Time-lagged rule associates	Visualization Color
-3	$+\ gene_r\ w\rightarrow-\ gene_i$	
-2	$-\ gene_r\ w\rightarrow+\ gene_i$	
-1	$+/-\ gene_r\ w\rightarrow-/+\ gene_i$	
0	$gene_r$ does not interact with $gene_i$	
1	$+/-\ gene_r\ w\rightarrow+/-\ gene_i$	
2	$+\ gene_r\ w\rightarrow+\ gene_i$	
3	$-\ gene_r\ w\rightarrow-\ gene_i$	

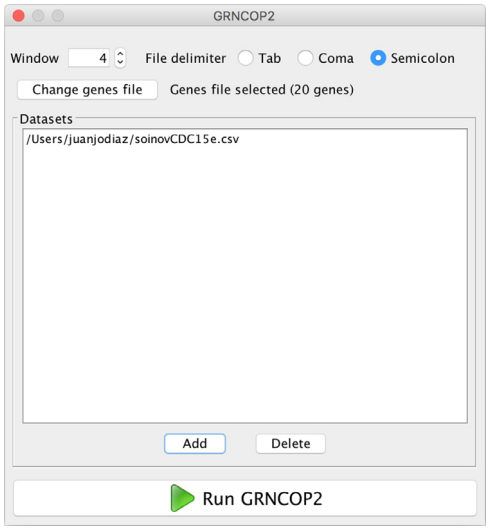


Fig. 1. GRNCOP2 configuration panel.

3. App design and benefits

GRNCOP2 App has been designed to facilitate the configuration and use of the GRNCOP2 method, the exploratory visual analysis of the inferred networks, and the integration with other tools and methods. GRNCOP2 App can be easily installed using Cytoscape's built-in app manager. Once the app is installed, a new item is displayed on the "Apps" menu on the top bar called GRNCOP2 from where the user can perform a GRNCOP2 analysis or load existing GRNCOP2 results.

As seen in Fig. 1, when running a new analysis, the user is presented with a configuration panel to load the data and configure the window to be used for time-lagged rules. The user needs to select the genes under study (as a file containing one gene per line) and an unlimited number of data files (as CSV files with the same number of lines as the genes files and an equal number of columns for all lines). The user can easily configure what should be considered the CSV separator.

Then, the app proceeds to load the data and perform the GRNCOP2 analysis leveraging multithreading for optimal performance, giving constant feedback to the user on the progress of the analysis and allowing the user to cancel the execution at any time. Once the analysis is completed, the inferred network is presented as a Cytoscape native network.

GRNCOP2 uses several thresholds to cut off low-confidence rules. Namely, the *Accuracy* acts as a cut off for rules that do not predict well based on the calculated score, the *Sample Coverage Percentage* (SCP) specifies the minimum TP (TN) count in relation to the number of samples to avoid rules with high accuracy but a small sample size, and the *Rule Consensus Accuracy* (RCA) threshold which specifies the minimum proportion of datasets in which a rule must predict well. The app uses default thresholds of 95 for all the values. As seen in Fig. 2, the user can modify them by using a panel displayed on the right-hand side of Cytoscape. It should be noted that during the initial execution the Target Discretization Threshold (TDT) and the Relative Regulation Threshold (RRT) are calculated for all the possible time lags. Thus, changing the filters only requires to recalculate the classifier inference process. This maximizes the performance and allows the user to quickly try different thresholds in order to maximize the inferred rules while reducing false positives.

One of the known limitations of GRNCOP2 is the large number of rules that are inferred, especially when working with multiple time windows. In order to reduce noise during visual analyses, the user can also use this panel to visualize only the rules corresponding to a certain time lag while hiding the others or show all the rules at once. Also, genes in the original dataset that don't belong to any rule can be hidden. Another problem is how to deal with the large amount of information inferred by GRNCOP2 in a user-friendly manner. GRNCOP2 App includes its own network visual

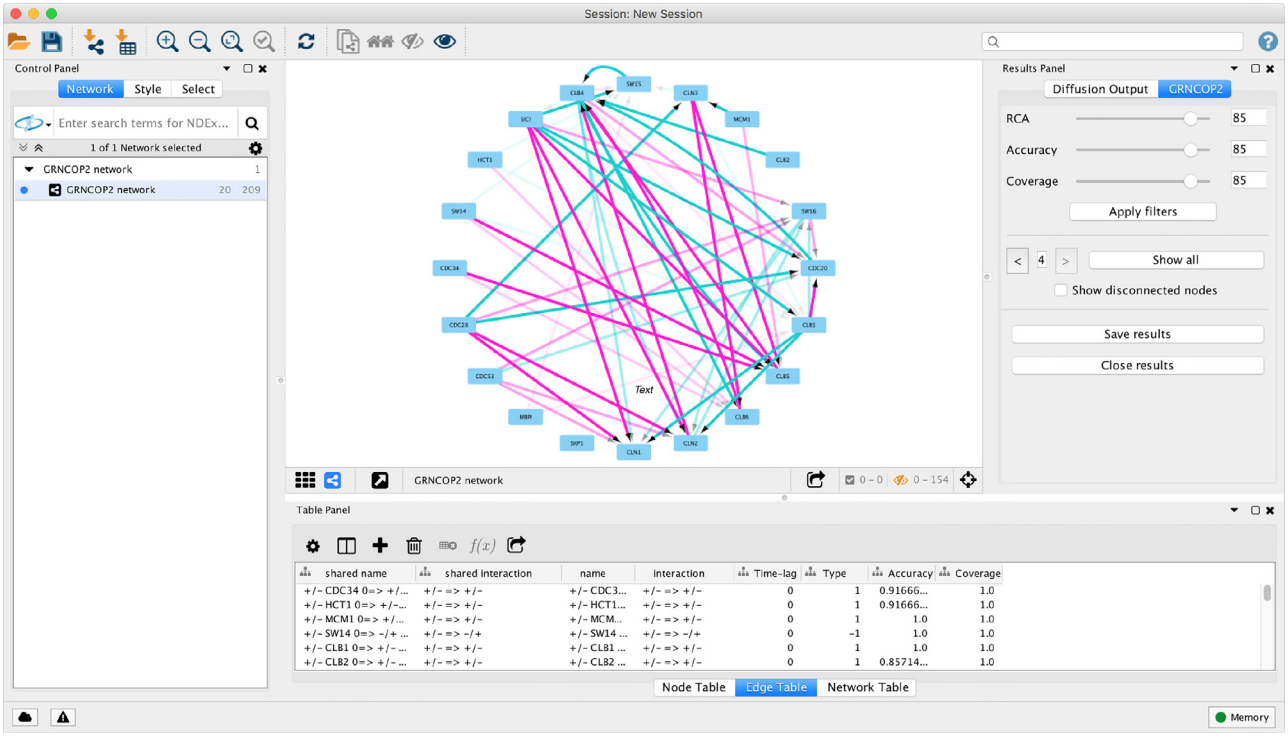


Fig. 2. GRNCOP2 results. (Center) Inferred GRN. (Right panel) GRNCOP2 control panel. (Bottom table) Detailed results.

style to facilitate the understanding and visual exploration of the resulting network (see Fig. 2). All the generated edges are directed from regulator to target gene, and their color indicates the relationship type (see Table 1). Also, the edge transparency and width indicate the accuracy and coverage of the inferred relationship respectively.

Moreover, since the inferred network is just a Cytoscape network, the user can take advantage of all the built-in functionality of Cytoscape as well as its rich app ecosystem (See results section for an example of this). Finally, to facilitate storage and sharing of GRNCOP2 results. The user can export the results using the right-hand side menu which will create a binary file containing all the information in a format specific to the GRNCOP2 App. This file can later by imported into Cytoscape using the top-bar menu in "Apps > GRNCOP2 > Load" result which will import the network and set the GRNCOP2 right-hand panel for exploratory analysis. The user can also export the network using any of the standard formats supported by Cytoscape (cif, json, xml,). However, importing those back to Cytoscape wont set up the GRNCOP2 panel.

4. Results

In this Section, the use of the GRNCOP2 App for Cytoscape is demonstrated by analyzing gene expression microarray data associated with Alzheimer's disease (AD) progression. AD is an irreversible brain disorder, which represents the largest part of dementia cases [22]. Currently, it is estimated that 47 million people suffer dementia worldwide, and it is projected that the number of cases will rise to more than 131 million by 2050 [23]. As a consequence, dementia has an enormous economic impact. During 2016, the global estimated total cost of dementia was US$818 billion, and it will grow into a trillion-dollar disease by 2018 [23]. For this reason, several gene expression analysis approaches have been applied in order to discover the hidden regulation mechanisms behind the development of this disease [24–29].

To demonstrate GRNCOP2 Apps benefits, a gene expression AD dataset extracted from Gene Expression Omnibus (GEO) [30] was analyzed. The dataset, named GSE1297 [16], contains hippocampal microarray experiments from nine control and twenty-two subjects with AD. The dataset samples are organized in four classes related to the AD progression: patients with no disease present, named control samples, patients with the disease on their early stages, named incipient samples, patients with moderate symptoms of the disease, named moderate samples, and patients on their last stages, named severe samples.

In particular, this study is focused into the detection of the stage of the disease progression in which the crosstalk signal between Alzheimer's disease and Ubiquitin Mediated Proteolysis pathways, annotated in KEGG [31] as hsa05010 and hsa04120 respectively, become more intense. Biological crosstalk denotes situations in which some parts of a signal transduction pathway influence another one. The selection of this case study in order to introduce GRNCOP2 App is due to several reasons. First of all, GRNCOP2 had never been used as a methodology for crosstalk analysis between pathways. Nevertheless, the functionalities provided by the Cytoscape App proposed in this paper make possible this kind of study, providing an innovative use case for GRNCOP2 practitioners. Secondly, evidence about the association between hsa04120 and hsa05010 pathways from GSE1297 dataset analysis have been reported by several crosstalk pathway detection methods [32–34]. Therefore, the biological hypothesis explored in this analysis seems feasible and can be well-contrasted with literature recently published. Finally, an interesting characteristic of this specific putative crosstalk is the absence of shared genes between these two pathways. Because of this, any biological signal detected between hsa04120 and hsa05010 can only be a consequence of a pathway synchronization promoted by the behaviors of genes that are not shared. We thus avoid the inference of irrelevant signals that can be straightforwardly explained by the simple presence of common genes in both pathways. The exploratory analysis conducted with the GRNCOP2 App will be oriented to respond to the following biological questions:

(1) Analyzing the network-topologies extracted by GRNCOP2 App from the gene expression dataset GSE1297, is there a stage of AD progression where the synchronization between hsa04120 and hsa05010 pathways become more intense?
(2) If there is a stage of AD progression where this biological signal is significantly more intense, which are the genes that explain the synchronization between the pathways? Are pieces of evidence in the literature that confirm the putative relevance of these genes for this crosstalk signal?

The strategy to address these questions using GRNCOP2 App consists of mining GSE1297 dataset to infer the association rules that emerge among the genes that are members of hsa04120 and hsa05010 pathways. More specifically, the rule extraction procedure is focused in the inter-pathways associations, i.e., rules that represent synchronizations between genes which do not belong to the same pathway, because the intra-pathways associations are not relevant in order to detect crosstalk. Therefore, the GRNs learned during these experiments are bipartite graphs where one group represents the hsa04120 genes, whereas the other one corresponds to the hsa05010 genes.

Thanks to GRNCOP2 App integration into Cytoscape, its built-in filters can be used to separate the two subsets and remove the rules between genes in the same pathway. Furthermore, these filters can be combined with an additional Cytoscape app called setsApp [35] in order to create and manage both gene sets independently. As shown in Fig. 3, setsApp also allows the user to layout the GRN with both subsets clearly separated.

After that, a topological analysis is carried out by means of contrasting the changes in the connectivity intensity of these bigraphs between consecutive stages of the AD progression. In Table 2, these connectivity intensities are detailed, where the networks have been extracted by GRNCOP2 App using the parametrization settings defined by default and time-delay equal to cero (see Section Material and Methods).

Once again, Cytoscape's built-in NetworkAnalyzer [36] can be leveraged to provide metrics about the network topology. Such metrics are seamlessly integrated into Cytoscapes network and can be used to modify the style of the network view to include the topology information in a visual manner as shown in Fig. 4. Also as shown in Fig. 4, Cytoscape allows the user to have multiple open networks (managed by the Control Panel) and switch easily between their views or even visualize their views side by side.

5. Discussion

The analysis of the results can start by inspecting Table 2, where it is clear that the number of inter-pathways connections significantly increases (31.23%) during the transition between the moderate and severe stages. This topological result is consistent with the literature. The ubiquitin-mediated proteolysis (UMP) is a process where an enzyme system labels undesirable proteins with many molecules of the 76-amino acid residue protein ubiquitin. After that, the labelled proteins are transported to the proteasome where these proteins are degraded. Several cellular processes are regulated by ubiquitin-mediated proteolysis, including the cell cycle, DNA repair and transcription, protein quality control and the immune response. Deficiencies in this proteolysis play a causal role in numerous human diseases.

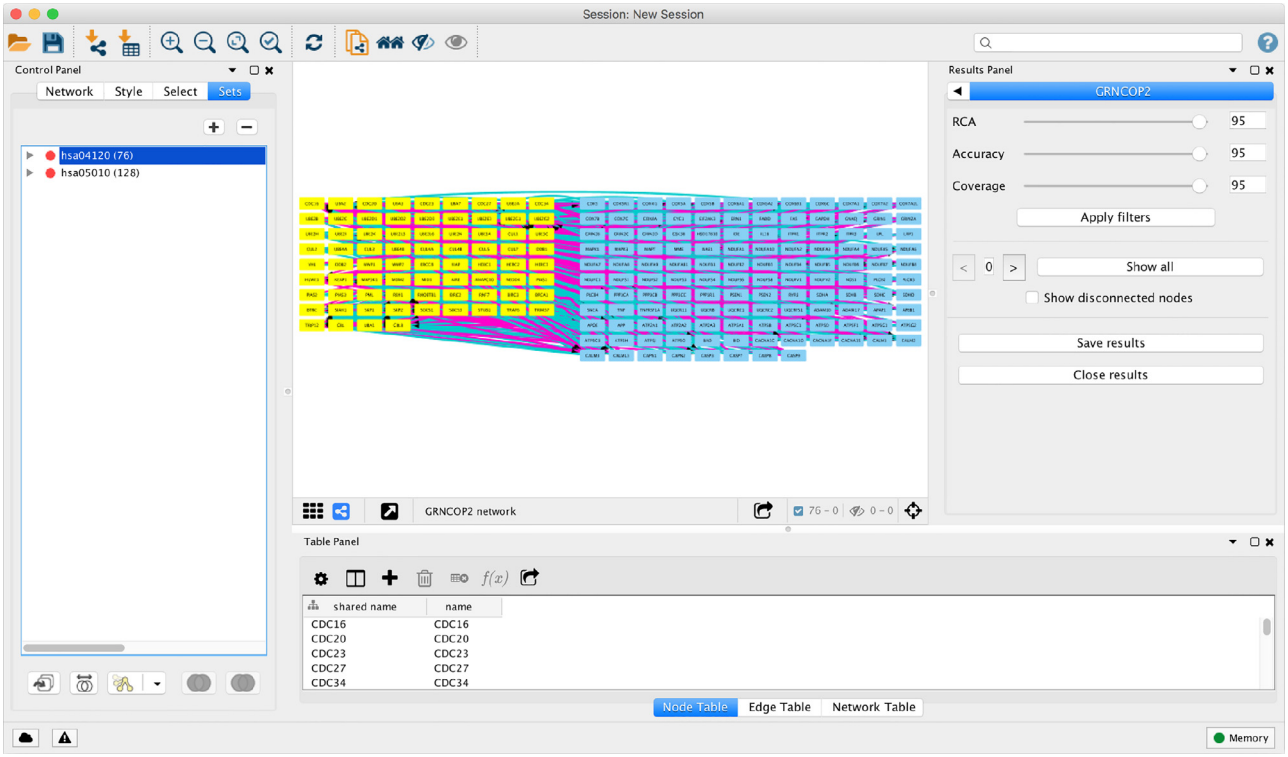


Fig. 3. GRN inferred by GRNCOP2 after its 2 sets have been identified and the intra-set rules have been removed.

Table 2
Variations of the number of node connections during the transitions between consecutive stages of AD progression.

AD Phase Transitions	#connections		
	(total)	(variation)	(ratio)
Control to Incipient	995	−75	7.54%
Incipient to Moderate	903	−92	10.19%
Moderate to Severe	1313	410	31.23%

In particular, it is well-known that the function of this pathway is perturbed in AD [37,38]. Defective proteasome activity is detected in the early phase of AD along with synaptic dysfunction, and in late AD stages it's associated with significant increments in the accumulation and aggregation of ubiquitinated (Ub)-proteins which occurs just before tangle formation [39,40]. Therefore, the observed growth in the number of inter-pathways connections during the AD progression from moderate to severe samples constitutes the expected result, which answers to the first biological question proposed in this analysis.

The next step of the study is focused on the identification of the driving genes associated with this transition using a topological criterion. For this task, the ubiquitin-mediated proteolysis genes with bigger increments in their in-out connections have been extracted by comparing the GRNs that correspond to the moderate and severe samples. The hypothesis is that these genes are signaling the changes in the transition between these progression stages. This information is summarized in Table 3, where the first five genes of the rank have been selected. The sum of the increments in the in-out connections of these 5 genes explain the 53% of the increment of association rules in the severe GRN, the other 131 genes of this pathway explain the remaining increment. Table 3. Genes of the ubiquitin-mediated proteolysis with bigger increments in their in-out connections from moderate network to severe network.

Finally, the putative relevance of these five genes in the progression of AD has been curated from a literature review. CDC16 is a component of the anaphase promoting complex/cyclosome (APC/C), which is a cell cycle-regulated E3 ubiquitin ligase. The APC/C regulates key cellular processes, including the cell cycle, by targeting a set of substrates for degradation. In the last decade, APC/C has been linked to several main functions in the nervous system, such as axon guidance, synaptic plasticity, neurogenesis, and others. Remarkably, some of the identified APC/C substrates have been related to neurodegenerative diseases. There is an accumulation of some degradation targets of APC/C in AD brains, which suggests a dysregulation of the protein complex in the progression of this disease. Additionally, evidence of inactivation of APC/C in AD has been recently provided by Fuchsberger et al. [41].

CUL1 and CUL5 belong to the Cullin-RING ubiquitin ligases (CRLs), which comprise the major known class of ubiquitin ligases. CRLs regulate a varied number of dynamic cellular processes, including the downregulation of misfolded proteins [42]. In particular, during the last two decades, neddylation has emerged as a major regulatory pathway for ubiquitination. In this process, an ubiquitin-like protein called NEDD8 is conjugated to its target proteins regulating CRLs and, recently, new pieces of evidence suggest that dysfunction of neddylation is involved in Alzheimer's disease [43]. For this reason, there is a clear relationship between these two ligases and the AD progression.

BRCA1 is a well-studied DNA repair factor. This protein works in the nucleus of cells, to mend breaks in DNA. As a consequence, the BRCA1 protein is a central actor in preserving the stability of

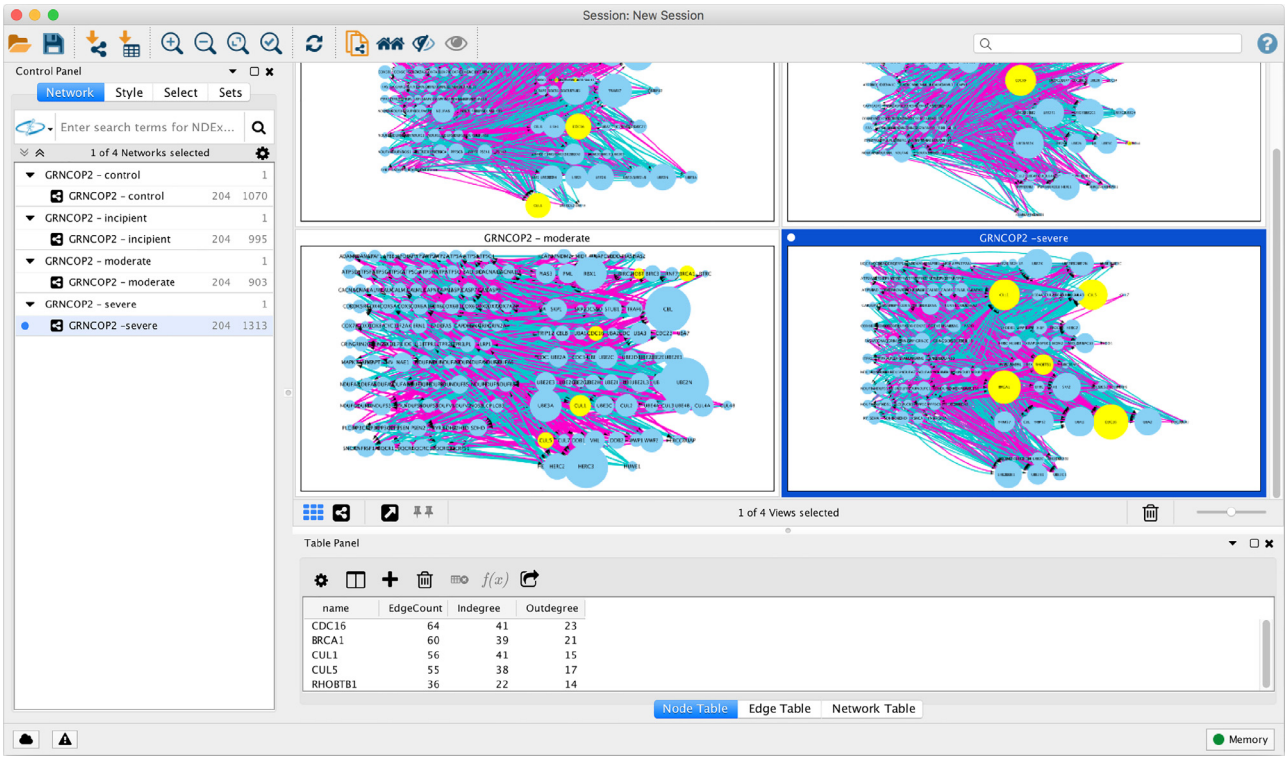


Fig. 4. GRNs inferred by GRNCOP2 for the 4 states of the study after Cytoscapes NetworkAnalyzer has been used to extract topology data and the visual style has been updated to show the size of the genes in proportion to their degree in the network.

Table 3
Genes of the ubiquitin-mediated proteolysis with bigger increments in their in-out connections from moderate network to severe network.

Gene	In–connections			Out–connections			Total
Name	Moderate	Severe	Increment	Moderate	Severe	Increment	Increment
CDC16	5	23	18	6	41	35	53
CUL5	3	17	14	3	38	35	49
BRCA1	3	21	18	9	39	30	48
CUL1	9	15	6	9	41	32	38
RHOBTB1	2	14	12	5	22	17	29

a cell's genetic information. Defective DNA repair may contribute to neurological disorders, including AD. In the last years, Suberbielle et al. [44] found reduced levels of BRCA1, but not of other DNA repair factors, in the brains of AD patients. In their experiments, the physiological neuronal activation increased BRCA1 concentrations, whereas stimulating predominantly extrasynaptic N-methyl-Daspartate receptors promoted the proteasomal degradation of BRCA1. Therefore, they conclude that BRCA1 is regulated by neuronal activity, for supporting neuronal integrity and cognitive functions. Nevertheless, the pathological accumulation of amyloid plaques, which occurs during late phases of the AD progression, depletes neuronal BRCA1 contributing to the cognitive deficits associated with this disease.

Finally, RHOBTB1 belongs to the Rho family of the small GTPase superfamily, which also integrates RhoBTB2, and RhoBTB3. They are traditionally considered tumor suppressor genes, but recent studies have reinforced their association in tumorigenesis and other pathological diseases, including AD where they play causative roles in disease progression [27,45]. Additionally, Park et al. [33] have established that RHOBTB1 gene is a strong causal mediator of AD.

In summary, it is possible to conclude that the gene expression analysis of the crosstalk hypothesis between AD and ubiquitin-mediated proteolysis pathways could be executed from the network-topologies generated by the GRNCOP2 Cytoscape App presented in this work, and the findings have been confirmed by means of literature scrutiny. The contrasting facilities and visualization tools provided for this new Cytoscape app make possible this kind of comparative topological-analysis among networks carried-out in this experiment, extending the traditional use of GRNCOP2 for studying isolated GRNs.

6. Conclusions

In this work, the implementation of a GRNCOP2 App for Cytoscape was proposed. The main goal of this development is to exploit Cytoscapes capabilities for the visualization and topological analysis of biological networks. The GRNCOP2 App is freely available at the official Cytoscape app store: http://apps.cytoscape.org/apps/grncop2.

In order to evaluate the benefits of GRNCOP2 integration to Cytoscape, a new use case for this algorithm was presented in the field of pathway analysis. In particular, the app was used to perform a topological study of the pathway crosstalk between the "Alzheimer's disease" and the "ubiquitin mediate proteolysis" pathways. A dataset that contains hippocampal microarray experiments of nine control and twenty-two subjects with AD was used for this

analysis. The dataset samples are structured in four classes related to the AD progression: patients with no disease present, called control samples, patients with the disease on their premature stages, called incipient samples, patients with moderate symptoms of the disease, called moderate samples, and patients on their last stages, named severe samples. The proposed analysis focused on assessing the stage of the AD progression where the synchronization between both pathways become more intense, and which are the genes that explain the synchronization between these pathways. Both questions were addressed by topological comparative analysis of the GRNs generated from the four classes of samples, but limited to the genes that belong to the pathways under study, where different facilities of Cytoscape were exploited. The topological findings achieved by this new App have been consistent with biological evidence reported in the scientific literature, showing the feasibility of using this new GRNCOP2 App in pathway analysis.

As future work, we plan to evaluate the extension of GRNCOP2 App with functionalities related to semantic analysis of GRNs, as a strategy to filter association rules in large networks and improve the biological interpretability of the results. Lastly, we hope that the important community of users of Cytoscape provides higher visibility and impact to GRNCOP2 method.

Declaration of Competing Interest

None.

Acknowledgements

This work is kindly supported by CONICET, grant PIP 112-2012-0100471 and UNS, grant PGI 24/N042. This work has been also partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund under the project TIN2015-64776-C3-2-R DIFERENTIAL@UPO: Massive data management, filtering and exploratory analysis. We also thank the AUIP (Asociación Universitaria Iberoamericana de Postgrado) for partially supported the visit of Dr. Ponzoni to the Pablo de Olavide University in 2017.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.05.030.

References

[1] R. Alves, D.S. Rodriguez-Baena, J.S. Aguilar-Ruiz, Gene association analysis: a survey of frequent pattern mining from gene expression data, Briefings Bioinf. 11 (2) (2009) 210–224.
[2] S.C. Madeira, M.C. Teixeira, I. Sa-Correia, A.L. Oliveira, Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm, IEEE/ACM Trans. Comput. Biol.Bioinform. (TCBB) 7 (1) (2010) 153–165.
[3] G. Agapito, M. Cannataro, P.H. Guzzi, M. Milano, Go-war: a tool for mining weighted association rules from gene ontology annotations, in: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer, 2014, pp. 3–18.
[4] P.H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Briefings Bioinf. 13 (5) (2011) 569–585.
[5] F. Emmert-Streib, M. Dehmer, B. Haibe-Kains, Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks, Front. Cell Dev. Biol. 2 (2014), doi:10.3389/fcell.2014.00038.
[6] C.A. Gallo, J.A. Carballido, I. Ponzoni, Inference of gene regulatory networks based on association rules, Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, John Wiley & Sons, Inc 80340, Hoboken, NJ, 2013.
[7] H. Jiang, T. Turki, S. Zhang, J.T. Wang, Reverse engineering gene regulatory networks using graph mining, in: International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer, 2018, pp. 335–349.
[8] C. Fernandez-Basso, M. Dolores Ruiz, M.J. Martin-Bautista, Extraction of association rules using big data technologies, Int. J. Des. Nat. Ecodyn. 11 (3) (2016) 178–185, doi:10.2495/dne-v11-n3-178-185.
[9] K. Geethanandhini, N. R., Association rule mining on big data a survey, Int. J. Eng. Res. Technol. 5 (5) (2016) 42–46, doi:10.2495/dne-v11-n3-178-185.
[10] C. Gallo, J. Carballido, I. Ponzoni, Discovering time-lagged rules from microarray data using gene profile classifiers, BMC Bioinf. 12 (1) (2011) 123+, doi:10.1186/1471-2105-12-123.
[11] Y.-C. Liu, C.-P. Cheng, V.S. Tseng, Discovering relational-based association rules with multiple minimum supports on microarray datasets, Bioinformatics 27 (22) (2011) 3142–3148, doi:10.1093/bioinformatics/btr526.
[12] S. Chen, T. Tsai, C.e.a. Chung, Dynamic association rules for gene expression data analysis, BMC Genomics. 16 (786) (2015), doi:10.1186/s12864-015-1970-x.
[13] I. Ponzoni, F. Azuaje, J. Augusto, D. Glass, Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning., IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (4) (2007) 624–634, doi:10.1109/tcbb.2007.1049.
[14] F. Gómez-Vela, J.A. Lagares, N. Díaz-Díaz, Gene network coherence based on prior knowledge using direct and indirect relationships, Comput. Biol. Chem. 56 (2015) 142–151.
[15] L. Soinov, M. Krestyaninova, A. Brazma, Towards reconstruction of gene networks from expression data by supervised learning, Genome Biol. 4 (2003) R6.
[16] E.M. Blalock, J.W. Geddes, K.C. Chen, N.M. Porter, W.R. Markesbery, P.W. Landfield, Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses, Proc.Natl. Acad. Sci. 101 (7) (2004) 2173–2178, doi:10.1073/pnas.0308512100.
[17] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.
[18] S.A. Sam, J. Teel, A.N. Tegge, A. Bharadwaj, T. Murali, XTalkDB: a database of signaling pathway crosstalk, Nucleic Acids Res. 45 (D1) (2016) D432–D439.
[19] B. Dutta, A. Wallqvist, J. Reifman, Pathnet: a tool for pathway analysis using topological information., Source Code Biol Med. 7 (1) (2012) 10+, doi:10.1186/1751-0473-7-10.
[20] J.S. Dussaut, C.A. Gallo, R.L. Cecchini, J.A. Carballido, I. Ponzoni, Crosstalk pathway inference using topological information and biclustering of gene expression data, Biosystems 150 (2016) 1–12.
[21] J.S. Dussaut, R.L. Cecchini, C.A. Gallo, I. Ponzoni, J.A. Carballido, A review of software tools for pathway crosstalk inference, Curr. Bioinf. 13 (1) (2018) 64–72.
[22] A. Burns, S. Iliffe, Alzheimers disease, BMJ (2009) 338:b158.
[23] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, M. Karagiannidou, World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future, Technical Report, Alzhemier's Disease International, 2016.
[24] P.P. Panigrahi, T.R. Singh, Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: revealed association with aging and other diseases, J. Theor. Biol. 334 (2013) 109–121, doi:10.1016/j.jtbi.2013.06.013.
[25] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A.A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M.E. MacDonald, J.R. Lamb, D.A. Bennett, C. Molony, D.J. Stone, V. Gudnason, A.J. Myers, E.E. Schadt, H. Neumann, J. Zhu, V. Emilsson, Integrated systems approach identifies genetic nodes and networks in late-Onset Alzheimer's disease, Cell 153 (3) (2013) 707–720, doi:10.1016/j.cell.2013.03.030.
[26] w. Wei Kong, X. Mou, X. Zhi, W. Zhang, Y. Yang, Dynamic regulatory network reconstruction for Alzheimer's disease based on matrix decomposition techniques, Comput. Math. Methods Med. 2014 (ID 891761) (2014) 1–10, doi:10.1155/2014/891761.
[27] W. Ji, F. Rivero, Atypical rho GTPases of the RhoBTB subfamily: roles in vesicle trafficking and tumorigenesis, Cells 5 (2) (2016), doi:10.3390/cells5020028.
[28] Y.-S. Hu, J. Xin, Y. Hu, L. Zhang, J. Wang, Analyzing the genes related to Alzheimers disease via a network and pathway-based approach, Comput. Math. Methods Med. 9 (29) (2017), doi:10.1186/s13195-017-0252-z.
[29] S. Kawalia, T. Raschka, M. Naz, R. de Matos Simoes, P. Senger, M. Hofmann-Apitius, Analytical strategy to prioritize Alzheimers disease candidate genes in gene regulatory networks using public expression data, J. Alzheimers Dis. 59 (4) (2017) 1237–1254, doi:10.3233/JAD-170011.
[30] R. Edgar, M. Domrachev, A.E. Lash, Gene expression omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (1) (2002) 207–210, doi:10.1093/nar/30.1.207.
[31] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[32] B. Dutta, A. Wallqvist, J. Reifman, PathNet: a tool for pathway analysis using topological information, Source Code Biol Med. 7 (1) (2012) 10.
[33] Y. Park, A. Sarkar, L. He, J. Davila-Velderrain, P. De Jager, M. Kellis, A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease, bioRxiv (2017), doi:10.1101/219428.
[34] J.S. Dussaut, C.A. Gallo, R.L. Cecchini, J.A. Carballido, I. Ponzoni, Crosstalk pathway inference using topological information and biclustering of gene expression data, Biosystems 150 (2016) 1–12, doi:10.1016/j.biosystems.2016.08.002.

[35] J. Morris, S. Lotia, A. Wu, N. Doncheva, M. Albrecht, A. Pico, T. Ferrin, setsApp for cytoscape: set operations for cytoscape nodes and edges [version 2; referees: 3 approved], F1000Research 3 (149) (2015), doi:10.12688/f1000research.4392.2.
[36] Y. Assenov, F. Ramírez, S.-E.E. Schelhorn, T. Lengauer, M. Albrecht, Computing topological parameters of biological networks., Bioinformatics 24 (2) (2008) 282–284, doi:10.1093/bioinformatics/btm554.
[37] B.M. Riederer, G. Leuba, A. Vernay, I.M. Riederer, The role of the ubiquitin proteasome system in Alzheimer's disease, Exp. Biol. Med. 236 (3) (2011) 268–276, doi:10.1258/ebm.2010.010327.
[38] B. Gong, M. Radulovic, M.E. Figueiredo-Pereira, C. Cardozo, The ubiquitin-proteasome system: potential therapeutic targets for Alzheimers disease and spinal cord injury, Front. Mol. Neurosci. 9 (4) (2016), doi:10.3389/fnmol.2016.00004.
[39] S. Oddo, The ubiquitin-proteasome system in Alzheimer's disease, J. Cell. Mol. Med. 12 (2) (2008) 363–373, doi:10.1111/j.1582-4934.2008.00276.x.
[40] L. Bedford, S. Paine, N. Rezvani, M. Mee, J. Lowe, R.J. Mayer, The ubiquitin-Proteasome system: potential therapeutic targets for Alzheimers disease and spinal cord injury, Autophagy 5 (2009) 224–227, doi:10.4161/auto.5.2.7389.
[41] T. Fuchsberger, A. Lloret, J. Via, New functions of APC/C ubiquitin ligase in the nervous system and its role in Alzheimer's disease, Int. J. Mol. Sci. 18 (5) (2017), doi:10.3390/ijms18051057.
[42] D.R. Bosu, E.T. Kipreos, Cullin-RING ubiquitin ligases: global regulation and activation cycles, Cell Div. 3 (2008) 7+, doi:10.1186/1747-1028-3-7.
[43] Y. Chen, R. Neve, H. Liu, Neddylation dysfunction in alzheimers disease, Journal of Cellular and Molecular Medicine 16 (11) (2012), doi:10.1111/j.1582-4934.2012.01604.x.
[44] E. Suberbielle, B. Djukic, M. Evans, D.H. Kim, P. Taneja, X. Wang, M. Finucane, J. Knox, K. Ho, N. Devidze, E. Masliah, L. Mucke, DNA Repair factor BRCA1 depletion occurs in alzheimer brains and impairs cognitive function in mice, Nature Communications 6 (2015) 8897+, doi:10.1038/ncomms9897.
[45] J. Miller, R. Woltjer, J. Goodenbour, S. Horvath, D. Geschwind, Genes and pathways underlying regional and cell type changes in Alzheimer's disease, Genome Medicine 5 (5) (2013) 48+, doi:10.1186/gm452.

Parte IV

Conclusiones

Capítulo 8

Conclusions

8.1. Proposals and results

In this document, four tools for the inference, analysis, and validation of gene network, with a strong focus on usability and accessibility, a common lack in bioinformatics tools, are presented.

First, a new web tool to analyze genes affected by mutations and classify them by integrating existing prioritization tools, WGPA, is presented. WGPA assess different aspects of genetic pathogenicity using sequence data at the population level. In addition, to explore the polygenic contribution of mutations to disease, WGPA implements gene set enrichment analysis to prioritize disease-causing genes and gene interaction networks, thus providing a comprehensive annotation of personal genome data in diseases.

Expanding on the use of external information a novel methodology for the analysis of gene networks through semantic similarity, GFD-Net, is proposed. This novel methodology has been tested as a method for gene network validation, surpassing existing measures. It has also been tested as an analysis tool allowing the correct identification of the function of each gene in a given gene network related to yeast. Finally, it has been tested for the study of human diseases, being able to associate a given gene network with the disease with which it is related. GFD-Net has been implemented as an accessible Cytoscape App easily usable without specialized prior knowledge.

Along the same line of gene network validation and also focusing on usability and accessibility, GNC App is presented as an evaluation tool that measures the goodness of a network based on its structural similarity with a known network used as a gold-standard. This implementation significantly reduces the computational cost of the methodology, allowing the analysis of large networks and offering all the benefits of being integrated into the Cytoscape ecosystem, such as better visualizations and integration with external data sources or other analysis tools. The usefulness of the GNC App has been tested by analyzing a gene network related to the yeast cell cycle.

Finally, a graphic tool is presented for the inference of gene regulatory networks based on the generation of rules by means of machine learning. This solution reduces the computational cost of the original proposal and includes new functionalities for visualization of the inferred networks, dynamic user interactions, and integration with other apps for topological analysis of the networks. To demonstrate the usefulness of integrating GRNCOP2 into Cytoscape, the new App was used to address a novel use case for GRNCOP2: the analysis of cross-talk between pathways, specifically between Alzheimer's disease and ubiquitin-mediated proteolysis.

Thus, it can be concluded that a new methodology for the validation and analysis of gene networks, a novel tool for the analysis of gene sets and networks affected

by mutations, and two tools that improve the usability and accessibility of existing methodologies and apply them to novel use cases have been presented. In addition, these four tools have been demonstrated to be useful in the study of human diseases and in other species such as yeast.

8.2. Future proposals

8.2.1. Improve GRNCOP-2 performance by using external information for rule filtering

It would be possible to improve GRNCOP2 as a network inference method by filtering false positives using external biological information. One of the great problems that GRNCOP2 presents is the large number of rules or interactions that arise between genes. This causes the user to have to parameterize multiple filters to eliminate false positives without losing information. In this sense, GFD-Net could be used to pre-filter the results obtained by GRNCOP2 and thus avoid this problem. Furthermore, GFD-Net would provide additional information on the relationships between genes inferred by GRNCOP2.

Since GRNCOP2 and GFD-Net are available as Cytoscape apps, the new approach would not require a re-implementation of both, just to consume them in a more suitable way.

8.2.2. Creation of a new network inference methodology based on similarity measures

A new gene network inference methodology could be developed to infer the relationships between a given group of genes. This way, it would be possible to find relationships between genes identified by other studies. To achieve this, a multi-objective evolutionary algorithm could be created that looks for the relationships in a set of genes and searches for the largest possible number of relationships between the genes in a way that maximizes the semantic similarity of the resulting network. To measure this similarity, GFD-Net could be used.

To maximize its use and interpretability, this new approach could be implemented within the Cytoscape ecosystem, either as an extension of the existing App for GFD-Net or as a new app.

Bibliografía

[1] D. Eisenberg, E. M. Marcotte, I. Xenarios y T. O. Yeates, «Protein function in the post-genomic era.,» *Nature*, vol. 405(6788), págs. 823-6. 2000.

[2] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From data mining to knowledge discovery in databases,» *AI magazine*, vol. 17, n.º 3, págs. 37-37, 1996.

[3] W. Ni, S. Zhang, B. Jiang y col., «Identification of cancer-related gene network in hepatocellular carcinoma by combined bioinformatic approach and experimental validation,» *Pathology-Research and Practice*, vol. 215, n.º 6, pág. 152 428, 2019.

[4] C. Cava, G. Bertoli, A. Colaprico, G. Bontempi, G. Mauri e I. Castiglioni, «In-Silico Integration Approach to Identify a Key miRNA Regulating a Gene Network in Aggressive Prostate Cancer,» *International journal of molecular sciences*, vol. 19, n.º 3, pág. 910, 2018.

[5] O. E. Ogundijo, A. Elmas y X. Wang, «Reverse engineering gene regulatory networks from measurement with missing values,» *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, n.º 1, págs. 1-11, 2016.

[6] F. M. Delgado y F. Gómez-Vela, «Computational methods for gene regulatory networks reconstruction and analysis: a review,» *Artificial intelligence in medicine*, vol. 95, págs. 133-145, 2019.

[7] K. Laukens, S. Naulaerts y W. V. Berghe, «Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis,» *Proteomics*, vol. 15, n.º 5-6, págs. 981-996, 2015.

[8] E. R. Dougherty, «Validation of inference procedures for gene regulatory networks,» *Current genomics*, vol. 8, n.º 6, págs. 351-359, 2007.

[9] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber y Z. Nikoloski, «Gene regulatory network inference using fused LASSO on multiple data sets,» *Scientific reports*, vol. 6, 2016.

[10] S. Mangul, L. S. Martin, E. Eskin y R. Blekhman, *Improving the usability and archival stability of bioinformatics software*, 2019.

[11] P. O. Brown y D. Botstein, «Exploring the new world of the genome with DNA microarrays,» *Nature genetics*, vol. 21, n.º 1, págs. 33-37, 1999.

[12] Z. Wang, M. Gerstein y M. Snyder, «RNA-Seq: a revolutionary tool for transcriptomics,» *Nature reviews genetics*, vol. 10, n.º 1, págs. 57-63, 2009.

[13] C. Evans, J. Hardin y D. M. Stoebel, «Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions,» *Briefings in bioinformatics*, vol. 19, n.º 5, págs. 776-792, 2018.

[14] P. J. Park, «ChIP–seq: advantages and challenges of a maturing technology,» *Nature reviews genetics*, vol. 10, n.º 10, págs. 669-680, 2009.

[15] T. Barrett, S. E. Wilhite, P. Ledoux y col., «NCBI GEO: archive for functional genomics data sets—update,» *Nucleic acids research*, vol. 41, n.º D1, págs. D991-D995, 2012.

[16] B. Zhang y S. Horvath, «A general framework for weighted gene co-expression network analysis,» *Statistical applications in genetics and molecular biology*, vol. 4, n.º 1, 2005.

[17] A.-C. Haury, F. Mordelet, P. Vera-Licona y J.-P. Vert, «TIGRESS: trustful inference of gene regulation using stability selection,» *BMC systems biology*, vol. 6, n.º 1, págs. 1-17, 2012.

[18] S. Liang, S. Fuhrman, R. Somogyi y col., «Reveal, a general reverse engineering algorithm for inference of genetic network architectures,» en *Pacific symposium on biocomputing*, Citeseer, vol. 3, 1998, págs. 18-29.

[19] A. V. Werhli y D. Husmeier, «Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge,» *Statistical applications in genetics and molecular biology*, vol. 6, n.º 1, 2007.

[20] E. Sakamoto y H. Iba, «Inferring a system of differential equations for a gene regulatory network by using genetic programming,» en *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, IEEE, vol. 1, 2001, págs. 720-726.

[21] B. Ma, M. Fang y X. Jiao, «Inference of gene regulatory networks based on nonlinear ordinary differential equations,» *Bioinformatics*, vol. 36, n.º 19, págs. 4885-4893, 2020.

[22] A. Niculescu-Mizil y R. Caruana, «Inductive transfer for Bayesian network structure learning,» en *Artificial intelligence and statistics*, PMLR, 2007, págs. 339-346.

[23] J. Chiquet, Y. Grandvalet y C. Ambroise, «Inferring multiple graphical structures,» *Statistics and Computing*, vol. 21, n.º 4, págs. 537-553, 2011.

[24] X. Li, S. Rao, W. Jiang y col., «Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling,» *BMC bioinformatics*, vol. 7, n.º 1, pág. 26, 2006.

[25] Y. Ichihashi, J. A. Aguilar-Martínez, M. Farhi y col., «Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape,» *Proceedings of the National Academy of Sciences*, vol. 111, n.º 25, E2616-E2621, 2014.

[26] S. Tabe-Bordbar, A. Emad, S. D. Zhao y S. Sinha, «A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models,» *Scientific reports*, vol. 8, 2018.

[27] R. M. Piro, S. Wiesberg, G. Schramm y col., «Network topology-based detection of differential gene regulation and regulatory switches in cell metabolism and signaling,» *BMC systems biology*, vol. 8, n.º 1, págs. 1-10, 2014.

[28] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos y col., «Using graph theory to analyze biological networks,» *BioData mining*, vol. 4, n.º 1, págs. 1-27, 2011.

[29] F. Gómez-Vela, J. A. Lagares y N. Díaz-Díaz, «Gene network coherence based on prior knowledge using direct and indirect relationships,» *Computational biology and chemistry*, vol. 56, págs. 142-151, 2015.

[30] J. B. Bard y S. Y. Rhee, «Ontologies in biology: design, applications and future challenges,» *nature reviews genetics*, vol. 5, n.º 3, págs. 213-222, 2004.

[31] B. Smith, M. Ashburner, C. Rosse y col., «The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration,» *Nature biotechnology*, vol. 25, n.º 11, págs. 1251-1255, 2007.

[32] A. Subramanian, P. Tamayo, V. K. Mootha y col., «Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,» *Proceedings of the National Academy of Sciences*, vol. 102, n.º 43, págs. 15 545-15 550, 2005.

[33] D. W. Huang, B. T. Sherman y R. A. Lempicki, «Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,» *Nucleic acids research*, vol. 37, n.º 1, págs. 1-13, 2009.

[34] C. Bettembourg, C. Diot y O. Dameron, «Optimal threshold determination for interpreting semantic similarity and particularity: application to the comparison of gene sets and metabolic pathways using GO and ChEBI,» *PloS one*, vol. 10, n.º 7, e0133579, 2015.

[35] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider y A. Valencia, «EnrichNet: network-based gene set enrichment analysis,» *Bioinformatics*, vol. 28, n.º 18, págs. i451-i457, 2012.

[36] O. Frings, J. E. Mank, A. Alexeyenko y E. L. Sonnhammer, «Network analysis of functional genomics data: application to avian sex-biased gene expression,» *The Scientific World Journal*, vol. 2012, 2012.

[37] G. K. Mazandu, E. R. Chimusa y N. J. Mulder, «Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery,» *Briefings in bioinformatics*, vol. 18, n.º 5, págs. 886-901, 2017.

[38] C. Pesquita, D. Faria, A. O. Falcao, P. Lord y F. M. Couto, «Semantic similarity in biomedical ontologies,» *PLoS computational biology*, vol. 5, n.º 7, e1000443, 2009.

[39] J. J. Díaz-Montana, O. J. Rackham, N. Díaz-Díaz y E. Petretto, «Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data,» *Bioinformatics*, vol. 32, n.º 4, págs. 635-637, oct. de 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv598. dirección: https://doi.org/10.1093/bioinformatics/btv598.

[40] J. J. Díaz-Montaña, N. Díaz-Díaz y F. Gómez-Vela, «GFD-Net: A novel semantic similarity methodology for the analysis of gene networks,» *Journal of Biomedical Informatics*, vol. 68, págs. 71-82, 2017, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2017.02.013. dirección: https://www.sciencedirect.com/science/article/pii/S1532046417300382.

[41] J. J. Dí-Montaña, F. Gómez-Vela y N. Díaz-Díaz, «GNC-app: A new Cytoscape app to rate gene networks biological coherence using gene,Äìgene indirect relationships,» *Biosystems*, vol. 166, págs. 61-65, 2018, ISSN: 0303-2647. DOI: https://doi.org/10.1016/j.biosystems.2018.01.007. dirección: https://www.sciencedirect.com/science/article/pii/S0303264717303258.

[42] J. J. Díaz-Montaña, N. Díaz-Díaz, C. D. Barranco e I. Ponzoni, «Development and use of a Cytoscape app for GRNCOP2,» *Computer Methods and Programs in Biomedicine*, vol. 177, págs. 211-218, 2019, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2019.05.030. dirección: https://www.sciencedirect.com/science/article/pii/S0169260718313592.

[43] M. List, P. Ebert y F. Albrecht, *Ten simple rules for developing usable software in computational biology*, 2017.

[44] S. Mangul, T. Mosqueiro, R. J. Abdill y col., «Challenges and recommendations to improve the installability and archival stability of omics computational tools,» *PLoS biology*, vol. 17, n.º 6, e3000333, 2019.

[45] P. Shannon, A. Markiel, O. Ozier y col., «Cytoscape: a software environment for integrated models of biomolecular interaction networks,» *Genome research*, vol. 13, n.º 11, págs. 2498-2504, 2003.

[46] R. Saito, M. E. Smoot, K. Ono y col., «A travel guide to Cytoscape plugins,» *Nature methods*, vol. 9, n.º 11, págs. 1069-1076, 2012.

[47] R. C. Jiménez, M. Kuzak, M. Alhamdoosh y col., «Four simple recommendations to encourage best practices in research software,» *F1000Research*, vol. 6, n.º 876, pág. 876, 2017.

[48] S. Altschul, B. Demchak, R. Durbin y col., «The anatomy of successful computational biology software.,» *Nature Biotechnology*, vol. 31, n.º 10, págs. 894-897, 2013.

[49] J. J. Díaz-Montaña, «Cytoscape: guía de iniciación al desarrollo,» *MoleQla: revista de Ciencias de la Universidad Pablo de Olavide*, n.º 14, págs. 8-5, 2014.

[50] J. J. Diaz-Montana y N. Diaz-Diaz, «Development and use of the Cytoscape app GFD-Net for measuring semantic dissimilarity of gene networks,» *F1000Research*, vol. 3, 2014.

[51] N. Diaz-Diaz y J. J. Diaz-Montana, «GFD-Net: a novel apporach for analyzing the functional dissimilarity of gene networks,» en *6th Argentinian Conference on Bioinformatics and Computational Biology*, A2B2C, 2015.

[52] F. H. Crick, «On protein synthesis,» en *Symp Soc Exp Biol*, vol. 12, 1958, pág. 8.

[53] Y. Zheng, S. Josefowicz, A. Chaudhry, X. P. Peng, K. Forbush y A. Y. Rudensky, «Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate,» *Nature*, vol. 463, n.º 7282, pág. 808, 2010.

[54] C. M. Dobson, «Protein folding and misfolding,» *Nature*, vol. 426, n.º 6968, pág. 884, 2003.

[55] F. U. Hartl, «Molecular chaperones in cellular protein folding,» *Nature*, vol. 381, n.º 6583, pág. 571, 1996.

[56] C. J. Epstein, R. F. Goldberger y C. B. Anfinsen, «The genetic control of tertiary protein structure: studies with model systems,» en *Cold Spring Harbor symposia on quantitative biology*, Cold Spring Harbor Laboratory Press, vol. 28, 1963, págs. 439-449.

[57] T. W. Myers y D. H. Gelfand, «Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase,» *Biochemistry*, vol. 30, n.º 31, págs. 7661-7666, 1991.

[58] C. P. Paul, P. D. Good, I. Winer y D. R. Engelke, «Effective expression of small interfering RNA in human cells,» *Nature biotechnology*, vol. 20, n.º 5, pág. 505, 2002.

[59] P. Ahlquist, «RNA-dependent RNA polymerases, viruses, and RNA silencing,» *Science*, vol. 296, n.º 5571, págs. 1270-1273, 2002.

[60] N. M. Luscombe, D. Greenbaum, M. Gerstein y col., «What is bioinformatics? An introduction and overview,» *Yearbook of medical informatics*, vol. 1, n.º 83-100, pág. 2, 2001.

[61] L. Perezleo Solórzano, R. Arencibia Jorge, C. Conill González, G. Achón Veloz y J. A. Araújo Ruiz, «Impacto de la bioinformática en las ciencias biomédicas,» *Acimed*, vol. 11, n.º 4, págs. 0-0, 2003.

[62] M. Huerta, G. Downing, F. Haseltine, B. Seto e Y. Liu, «NIH working definition of bioinformatics and computational biology,» *US National Institute of Health*, 2000.

[63] P. Edman y col., «Method for determination of the amino acid sequence in peptides,» *Acta chem. scand*, vol. 4, n.º 7, págs. 283-293, 1950.

[64] S. B. Needleman y C. D. Wunsch, «A general method applicable to the search for similarities in the amino acid sequence of two proteins,» *Journal of molecular biology*, vol. 48, n.º 3, págs. 443-453, 1970.

[65] F. Sanger, S. Nicklen y A. R. Coulson, «DNA sequencing with chain-terminating inhibitors,» *Proceedings of the national academy of sciences*, vol. 74, n.º 12, págs. 5463-5467, 1977.

[66] R Staden, «Sequence data handling by computer,» *Nucleic Acids Research*, vol. 4, n.º 11, págs. 4037-4052, 1977.

[67] F. Sanger, A. Coulson, T Friedmann y col., «The nucleotide sequence of bacteriophage φX174,» *Journal of molecular biology*, vol. 125, n.º 2, págs. 225-246, 1978.

[68] K. Cravedi, «GenBank Celebrates 25 Years of Service with Two-Day Conference. Leading Scientists Will Discuss the DNA Database at April 7–8 Meeting,» *NIH News*, 2008.

[69] E. E. Abola, N. O. Manning, J. Prilusky, D. R. Stampf y J. L. Sussman, «The Protein Data Bank: current status and future challenges,» *Journal of research of the National Institute of Standards and Technology*, vol. 101, n.º 3, pág. 231, 1996.

[70] R. A. Harper, «EMBnet: an institute without walls.,» *Trends in biochemical sciences*, vol. 21, n.º 4, págs. 150-152, 1996.

[71] G. H. Hamm y G. N. Cameron, «The EMBL data library,» *Nucleic acids research*, vol. 14, n.º 1, págs. 5-9, 1986.

[72] D. L. Wheeler, T. Barrett, D. A. Benson y col., «Database resources of the national center for biotechnology information,» *Nucleic acids research*, vol. 33, n.º suppl_1, págs. D39-D45, 2005.

[73] T. J. Berners-Lee y R. Cailliau, «World-wide web,» 1992.

[74] M. P. Sawicki, G. Samara, M. Hurwitz y E. Passaro, «Human genome project,» *The American journal of surgery*, vol. 165, n.º 2, págs. 258-264, 1993.

[75] F. Abascal Sebastián de Erice, «Análisis de genomas: métodos para la predicción y anotación de la función de las proteínas,» 2003.

[76] G. M. Rubin, M. D. Yandell, J. R. Wortman y col., «Comparative genomics of the eukaryotes,» *Science*, vol. 287, n.º 5461, págs. 2204-2215, 2000.

[77] C. B. Anfinsen, «The formation and stabilization of protein structure.,» *Biochemical Journal*, vol. 128, n.º 4, pág. 737, 1972.

[78] Y. Zhang, «Progress and challenges in protein structure prediction,» *Current opinion in structural biology*, vol. 18, n.º 3, págs. 342-348, 2008.

[79] R. A. Laskowski, «Protein structure databases,» *Molecular biotechnology*, vol. 48, n.º 2, págs. 183-198, 2011.

[80] V. Wirta, «Mining the Transcriptome-methods and Applications,» Tesis doct., KTH, 2006.

[81] O. Poetz, J. M. Schwenk, S. Kramer, D. Stoll, M. F. Templin y T. O. Joos, «Protein microarrays: catching the proteome,» *Mechanisms of ageing and development*, vol. 126, n.º 1, págs. 161-170, 2005.

[82] S. Cristoni y L. R. Bernardi, «Bioinformatics in mass spectrometry data analysis for proteomics studies,» *Expert review of proteomics*, vol. 1, n.º 4, págs. 469-483, 2004.

[83] I. G. Romero, I. Ruvinsky e Y. Gilad, «Comparative studies of gene expression and the evolution of gene regulation,» *Nature Reviews Genetics*, vol. 13, n.º 7, pág. 505, 2012.

[84] M. Oti y H. G. Brunner, «The modular nature of genetic diseases,» *Clinical genetics*, vol. 71, n.º 1, págs. 1-11, 2007.

[85] D. Latchman, *Gene regulation*. Taylor & Francis, 2007.

[86] H. Kitano, «Systems biology: a brief overview,» *Science*, vol. 295, n.º 5560, págs. 1662-1664, 2002.

[87] L. Xiaoli, N. See-kiong y W. J. TL, *Biological data mining and its applications in healthcare*. World scientific, 2013, vol. 8.

[88] A. M. Mabu, R. Prasad, R. Yadav y S. S. Jauro, «A Review of Data Mining Methods in Bioinformatics,» en *2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS)*, IEEE, 2018, págs. 1-6.

[89] M. J. Zaki, G. Karypis y J. Yang, *Data mining in bioinformatics (BIOKDD)*, 2007.

[90] S. Ahmed, M. U. Ali, J. Ferzund, M. A. Sarwar, A. Rehman y A. Mehmood, «Modern Data Formats for Big Bioinformatics Data Analytics,» *arXiv preprint arXiv:1707.05364*, 2017.

[91] D. J. Rigden y X. M. Fernández, «The 2021 Nucleic Acids Research database issue and the online molecular biology database collection,» *Nucleic acids research*, vol. 49, n.º D1, págs. D1-D9, 2021.

[92] D. M. Bolser, P.-Y. Chibon, N. Palopoli y col., «MetaBase—the wiki-database of biological databases,» *Nucleic acids research*, vol. 40, n.º D1, págs. D1250-D1254, 2011.

[93] M. D. Brazas, D. S. Yim, J. T. Yamada y B. F. Ouellette, «The 2011 bioinformatics links directory update: more resources, tools and databases and features to empower the bioinformatics community,» *Nucleic acids research*, vol. 39, n.º suppl_2, W3-W7, 2011.

[94] J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.

[95] A. Kozomara y S. Griffiths-Jones, «miRBase: annotating high confidence microRNAs using deep sequencing data,» *Nucleic acids research*, vol. 42, n.º D1, págs. D68-D73, 2013.

[96] I. Kalvari, J. Argasinska, N. Quinones-Olvera y col., «Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families,» *Nucleic acids research*, vol. 46, n.º D1, págs. D335-D342, 2017.

[97] Y. Xin y W. K. Olson, «BPS: a database of RNA base-pair structures,» *Nucleic acids research*, vol. 37, n.º suppl_1, págs. D83-D88, 2008.

[98] R. Consortium, «RNAcentral: a comprehensive database of non-coding RNA sequences,» *Nucleic acids research*, gkw1008, 2016.

[99] T. Hubbard, D. Barker, E. Birney y col., «The Ensembl genome database project,» *Nucleic acids research*, vol. 30, n.º 1, págs. 38-41, 2002.

[100] D. R. Zerbino, P. Achuthan, W. Akanni y col., «Ensembl 2018,» *Nucleic acids research*, vol. 46, n.º D1, págs. D754-D761, 2017.

[101] J. M. Cherry, E. L. Hong, C. Amundsen y col., «Saccharomyces Genome Database: the genomics resource of budding yeast,» *Nucleic acids research*, vol. 40, n.º D1, págs. D700-D705, 2011.

[102] L. S. Gramates, S. J. Marygold, G. d. Santos y col., «FlyBase at 25: looking to the future,» *Nucleic acids research*, gkw1016, 2016.

[103] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson y M. G. D. Group, «The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease,» *Nucleic acids research*, vol. 43, n.º D1, págs. D726-D736, 2014.

[104] M. Shimoyama, J. De Pons, G. T. Hayman y col., «The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease,» *Nucleic acids research*, vol. 43, n.º D1, págs. D743-D750, 2014.

[105] R. Y. N. Lee, K. L. Howe, T. W. Harris y col., «WormBase 2017: molting into a new stage,» *Nucleic acids research*, vol. 46, n.º D1, págs. D869-D874, 2017.

[106] P. Lamesch, T. Z. Berardini, D. Li y col., «The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools,» *Nucleic acids research*, vol. 40, n.º D1, págs. D1202-D1210, 2011.

[107] C. James-Zorn, V. G. Ponferrada, K. A. Burns y col., «Xenbase: Core features, data acquisition, and data processing,» *Genesis*, vol. 53, n.º 8, págs. 486-497, 2015.

[108] J. Sprague, L. Bayraktaroglu, D. Clements y col., «The Zebrafish Information Network: the zebrafish model organism database,» *Nucleic acids research*, vol. 34, n.º suppl_1, págs. D581-D585, 2006.

[109] W. C. Barker, J. S. Garavelli, H. Huang y col., «The protein information resource (PIR),» *Nucleic acids research*, vol. 28, n.º 1, págs. 41-44, 2000.

[110] B. Boeckmann, A. Bairoch, R. Apweiler y col., «The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003,» *Nucleic acids research*, vol. 31, n.º 1, págs. 365-370, 2003.

[111] U. Consortium, «UniProt: a hub for protein information,» *Nucleic acids research*, vol. 43, n.º D1, págs. D204-D212, 2014.

[112] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez y R. Apweiler, «UniProt archive,» *Bioinformatics*, vol. 20, n.º 17, págs. 3236-3237, 2004.

[113] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu y U. Consortium, «UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches,» *Bioinformatics*, vol. 31, n.º 6, págs. 926-932, 2014.

[114] H. Berman, K. Henrick, H. Nakamura y J. L. Markley, «The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data,» *Nucleic acids research*, vol. 35, n.º suppl_1, págs. D301-D303, 2006.

[115] S. Velankar, Y. Alhroub, A. Alili y col., «PDBe: protein data bank in Europe,» *Nucleic acids research*, vol. 39, n.º suppl_1, págs. D402-D410, 2010.

[116] P. D. Bank, «Research Collaboratory for Structural Bioinformatics,» *http://www.tcsb. org/pdb/*, 2005.

[117] A. G. Murzin, S. E. Brenner, T. Hubbard y C. Chothia, «SCOP: a structural classification of proteins database for the investigation of sequences and structures,» *Journal of molecular biology*, vol. 247, n.º 4, págs. 536-540, 1995.

[118] N. L. Dawson, T. E. Lewis, S. Das y col., «CATH: an expanded resource to predict protein function through structure and sequence,» *Nucleic acids research*, vol. 45, n.º D1, págs. D289-D295, 2016.

[119] N. K. Fox, S. E. Brenner y J.-M. Chandonia, «SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures,» *Nucleic acids research*, vol. 42, n.º D1, págs. D304-D309, 2013.

[120] R. D. Finn, P. Coggill, R. Y. Eberhardt y col., «The Pfam protein families database: towards a more sustainable future,» *Nucleic acids research*, vol. 44, n.º D1, págs. D279-D285, 2016.

[121] C. J. Sigrist, E. De Castro, L. Cerutti y col., «New and continuing developments at PROSITE,» *Nucleic acids research*, vol. 41, n.º D1, págs. D344-D347, 2012.

[122] C. H. Wu, A. Nikolskaya, H. Huang y col., «PIRSF: family classification system at the Protein Information Resource,» *Nucleic acids research*, vol. 32, n.º suppl_1, págs. D112-D114, 2004.

[123] M. Ashburner, C. A. Ball, J. A. Blake y col., «Gene Ontology: tool for the unification of biology,» *Nature genetics*, vol. 25, n.º 1, pág. 25, 2000.

[124] G. O. Consortium, «Gene ontology consortium: going forward,» *Nucleic acids research*, vol. 43, n.º D1, págs. D1049-D1056, 2014.

[125] P. Gaudet, M. S. Livstone, S. E. Lewis y P. D. Thomas, «Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium,» *Briefings in bioinformatics*, vol. 12, n.º 5, págs. 449-462, 2011.

[126] R. P. Huntley y R. C. Lovering, «Annotation extensions,» en *The Gene Ontology Handbook*, Springer, 2017, págs. 233-243.

[127] D. Gkika, L. Lemonnier, G. Shapovalov y col., «Trp channel–associated factors are a novel protein family that regulates trpm8 trafficking and activityidentification of trpm8 partner proteins,» *The Journal of cell biology*, vol. 208, n.º 1, págs. 89-107, 2015.

[128] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie y D. Eisenberg, «The database of interacting proteins: 2004 update,» *Nucleic acids research*, vol. 32, n.º suppl_1, págs. D449-D451, 2004.

[129] G. D. Bader, D. Betel y C. W. Hogue, «BIND: the biomolecular interaction network database,» *Nucleic acids research*, vol. 31, n.º 1, págs. 248-250, 2003.

[130] T. Keshava Prasad, R. Goel, K. Kandasamy y col., «Human protein reference database—2009 update,» *Nucleic acids research*, vol. 37, n.º suppl_1, págs. D767-D772, 2008.

[131] S. Kerrien, B. Aranda, L. Breuza y col., «The IntAct molecular interaction database in 2012,» *Nucleic acids research*, vol. 40, n.º D1, págs. D841-D846, 2011.

[132] L. Licata, L. Briganti, D. Peluso y col., «MINT, the molecular interaction database: 2012 update,» *Nucleic acids research*, vol. 40, n.º D1, págs. D857-D861, 2011.

[133] P. Pagel, S. Kovac, M. Oesterheld y col., «The MIPS mammalian protein–protein interaction database,» *Bioinformatics*, vol. 21, n.º 6, págs. 832-834, 2004.

[134] A. Chatr-Aryamontri, R. Oughtred, L. Boucher y col., «The BioGRID interaction database: 2017 update,» *Nucleic acids research*, vol. 45, n.º D1, págs. D369-D379, 2017.

[135] D. S. Wishart, Y. D. Feunang, A. C. Guo y col., «DrugBank 5.0: a major update to the DrugBank database for 2018,» *Nucleic acids research*, vol. 46, n.º D1, págs. D1074-D1082, 2017.

[136] S. Orchard, L. Salwinski, S. Kerrien y col., «The minimum information required for reporting a molecular interaction experiment (MIMIx),» *Nature biotechnology*, vol. 25, n.º 8, pág. 894, 2007.

[137] S. Orchard, S. Kerrien, S. Abbani y col., «Protein interaction data curation: the International Molecular Exchange (IMEx) consortium,» *Nature methods*, vol. 9, n.º 4, pág. 345, 2012.

[138] D. Szklarczyk, J. H. Morris, H. Cook y col., «The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible,» *Nucleic acids research*, gkw937, 2016.

[139] K. Zuberi, M. Franz, H. Rodriguez y col., «GeneMANIA prediction server 2013 update,» *Nucleic acids research*, vol. 41, n.º W1, W115-W122, 2013.

[140] H. Kim, J. Shin, E. Kim y col., «YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae,» *Nucleic acids research*, vol. 42, n.º D1, págs. D731-D736, 2013.

[141] G. Zhu, A. Wu, X.-J. Xu y col., «PPIM: a protein-protein interaction database for maize,» *Plant physiology*, pp-01 821, 2015.

[142] T. Murali, S. Pacifico, J. Yu, S. Guest, G. G. Roberts y R. L. Finley, «DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila,» *Nucleic acids research*, vol. 39, n.º suppl_1, págs. D736-D743, 2010.

[143] J. Shin, S. Yang, E. Kim y col., «FlyNet: a versatile network prioritization server for the Drosophila community,» *Nucleic acids research*, vol. 43, n.º W1, W91-W97, 2015.

[144] M. Hoebeke, H. Chiapello, P. Noirot y P. Bessieres, «SPiD: a subtilis protein interaction database,» *Bioinformatics*, vol. 17, n.º 12, págs. 1209-1212, 2001.

[145] Q. Lv, Y. Lan, Y. Shi y col., «AtPID: a genome-scale resource for genotype–phenotype associations in Arabidopsis,» *Nucleic acids research*, vol. 45, n.º D1, págs. D1060-D1063, 2016.

[146] M. Kanehisa y S. Goto, «KEGG: kyoto encyclopedia of genes and genomes,» *Nucleic acids research*, vol. 28, n.º 1, págs. 27-30, 2000.

[147] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs y col., «Expansion of the BioCyc collection of pathway/genome databases to 160 genomes,» *Nucleic acids research*, vol. 33, n.º 19, págs. 6083-6089, 2005.

[148] A. Fabregat, S. Jupe, L. Matthews y col., «The reactome pathway knowledgebase,» *Nucleic acids research*, vol. 46, n.º D1, págs. D649-D655, 2017.

[149] C. F. Schaefer, K. Anthony, S. Krupa y col., «PID: the pathway interaction database,» *Nucleic acids research*, vol. 37, n.º suppl_1, págs. D674-D679, 2009.

[150] D. Alonso-Lopez, M. A. Gutiérrez, K. P. Lopes, C. Prieto, R. Santamaría y J. De Las Rivas, «APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks,» *Nucleic acids research*, vol. 44, n.º W1, W529-W535, 2016.

[151] J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb y P. Uetz, «MPIDB: the microbial protein interaction database,» *Bioinformatics*, vol. 24, n.º 15, págs. 1743-1744, 2008.

[152] M. J. Cowley, M. Pinese, K. S. Kassahn y col., «PINA v2. 0: mining interactome modules,» *Nucleic acids research*, vol. 40, n.º D1, págs. D862-D865, 2011.

[153] N. Orii y M. K. Ganapathiraju, «Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function,» *PloS one*, vol. 7, n.º 11, e49029, 2012.

[154] A. Kamburov, U. Stelzl, H. Lehrach y R. Herwig, «The ConsensusPathDB interaction database: 2013 update,» *Nucleic acids research*, vol. 41, n.º D1, págs. D793-D800, 2012.

[155] E. G. Cerami, B. E. Gross, E. Demir y col., «Pathway Commons, a web resource for biological pathway data,» *Nucleic acids research*, vol. 39, n.º suppl_1, págs. D685-D690, 2010.

[156] T. F. Smith y M. S. Waterman, «Comparison of biosequences,» *Advances in applied mathematics*, vol. 2, n.º 4, págs. 482-489, 1981.

[157] D. J. Lipman y W. R. Pearson, «Rapid and sensitive protein similarity searches,» *Science*, vol. 227, n.º 4693, págs. 1435-1441, 1985.

[158] S. F. Altschul, W. Gish, W. Miller, E. W. Myers y D. J. Lipman, «Basic local alignment search tool,» *Journal of molecular biology*, vol. 215, n.º 3, págs. 403-410, 1990.

[159] R. Chenna, H. Sugawara, T. Koike y col., «Multiple sequence alignment with the Clustal series of programs,» *Nucleic acids research*, vol. 31, n.º 13, págs. 3497-3500, 2003.

[160] C. Notredame, D. G. Higgins y J. Heringa, «T-coffee: a novel method for fast and accurate multiple sequence alignment1,» *Journal of molecular biology*, vol. 302, n.º 1, págs. 205-217, 2000.

[161] W. Huber, V. J. Carey, R. Gentleman y col., «Orchestrating high-throughput genomic analysis with Bioconductor,» *Nature methods*, vol. 12, n.º 2, pág. 115, 2015.

[162] J. E. Stajich, D. Block, K. Boulez y col., «The Bioperl toolkit: Perl modules for the life sciences,» *Genome research*, vol. 12, n.º 10, págs. 1611-1618, 2002.

[163] P. J. Cock, T. Antao, J. T. Chang y col., «Biopython: freely available Python tools for computational molecular biology and bioinformatics,» *Bioinformatics*, vol. 25, n.º 11, págs. 1422-1423, 2009.

[164] A. Prlić, A. Yates, S. E. Bliven y col., «BioJava: an open-source framework for bioinformatics in 2012,» *Bioinformatics*, vol. 28, n.º 20, págs. 2693-2695, 2012.

[165] G. Yachdav, T. Goldberg, S. Wilzbach y col., «Cutting edge: anatomy of BioJS, an open source community for the life sciences,» *Elife*, vol. 4, e07009, 2015.

[166] N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts y T. Katayama, «BioRuby: bioinformatics software for the Ruby programming language,» *Bioinformatics*, vol. 26, n.º 20, págs. 2617-2619, 2010.

[167] O. Spjuth, J. Alvarsson, A. Berg y col., «Bioclipse 2: A scriptable integration platform for the life sciences,» *BMC bioinformatics*, vol. 10, n.º 1, pág. 397, 2009.

[168] F. Madeira, Y. M. Park, J. Lee y col., «The EMBL-EBI search and sequence analysis tools APIs in 2019,» *Nucleic acids research*, vol. 47, n.º W1, W636-W641, 2019.

[169] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis y T. L. Madden, «NCBI BLAST: a better web interface,» *Nucleic acids research*, vol. 36, n.º suppl_2, W5-W9, 2008.

[170] K.-B. Li, «ClustalW-MPI: ClustalW analysis using distributed and parallel computing,» *Bioinformatics*, vol. 19, n.º 12, págs. 1585-1586, 2003.

[171] H. Mi, A. Muruganujan, J. T. Casagrande y P. D. Thomas, «Large-scale gene function analysis with the PANTHER classification system,» *Nature protocols*, vol. 8, n.º 8, pág. 1551, 2013.

[172] G. Su, J. H. Morris, B. Demchak y G. D. Bader, «Biological network exploration with cytoscape 3,» *Current protocols in bioinformatics*, vol. 47, n.º 1, págs. 8-13, 2014.

[173] D. Otasek, J. H. Morris, J. Bouças, A. R. Pico y B. Demchak, «Cytoscape automation: empowering workflow-based network analysis,» *Genome biology*, vol. 20, n.º 1, págs. 1-15, 2019.

[174] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer y G. D. Bader, «Cytoscape. js: a graph theory library for visualisation and analysis,» *Bioinformatics*, vol. 32, n.º 2, págs. 309-311, 2016.

[175] S. Lotia, J. Montojo, Y. Dong, G. D. Bader y A. R. Pico, «Cytoscape app store,» *Bioinformatics*, vol. 29, n.º 10, págs. 1350-1351, 2013.

[176] M. Kutmon, S. Lotia, C. T. Evelo y A. R. Pico, «WikiPathways App for Cytoscape: making biological pathways amenable to network analysis and visualization,» *F1000Research*, vol. 3, 2014.

[177] S. Maere, K. Heymans y M. Kuiper, «BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks,» *Bioinformatics*, vol. 21, n.º 16, págs. 3448-3449, 2005.

[178] J. H. Morris, L. Apeltsin, A. M. Newman y col., «clusterMaker: a multi-algorithm clustering plugin for Cytoscape,» *BMC bioinformatics*, vol. 12, n.º 1, pág. 436, 2011.

[179] M. Kucera, R. Isserlin, A. Arkhangorodsky y G. D. Bader, «AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations,» *F1000Research*, vol. 5, 2016.

[180] K. Faust y J. Raes, «CoNet app: inference of biological association networks using Cytoscape,» *F1000Research*, vol. 5, 2016.

[181] N. Al-Ageel, A. Al-Wabil, G. Badr y N. AlOmar, «Human Factors in the Design and Evaluation of Bioinformatics Tools,» *Procedia Manufacturing*, vol. 3, págs. 2003-2010, 2015.

[182] F. da Veiga Leprevost, V. C. Barbosa, E. L. Francisco, Y. Perez-Riverol y P. C. Carvalho, «On best practices in the development of bioinformatics software.,» *Frontiers in Genetics*, vol. 5, págs. 199-199, 2014.

[183] D. Bolchini, A. Finkelstein, V. Perrone y S. Nagl, «Better bioinformatics through usability analysis,» *Bioinformatics*, vol. 25, n.º 3, págs. 406-412, 2009.

[184] R. Shamir y R. Sharan, «Algorithmic approaches to clustering gene expression data,» en *In*, Citeseer, 2002.

[185] D. Jiang, C. Tang y A. Zhang, «Cluster analysis for gene expression data: a survey,» *IEEE Transactions on knowledge and data engineering*, vol. 16, n.º 11, págs. 1370-1386, 2004.

[186] M. Binelli, S. C. Scolari, G. Pugliesi y col., «The transcriptome signature of the receptive bovine uterus determined at early gestation,» *PLoS One*, vol. 10, n.º 4, e0122874, 2015.

[187] S. Nagi, D. K. Bhattacharyya y J. K. Kalita, «Gene expression data clustering analysis: A survey,» en *2011 2nd National Conference on Emerging Trends and Applications in Computer Science*, IEEE, 2011, págs. 1-12.

[188] G. Kerr, H. J. Ruskin, M. Crane y P. Doolan, «Techniques for clustering gene expression data,» *Computers in biology and medicine*, vol. 38, n.º 3, págs. 283-293, 2008.

[189] G. H. Ball y D. J. Hall, «A clustering technique for summarizing multivariate data,» *Behavioral science*, vol. 12, n.º 2, págs. 153-155, 1967.

[190] J. MacQueen y col., «Some methods for classification and analysis of multivariate observations,» en *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, págs. 281-297.

[191] W. S. Sarle, *Algorithms for clustering data*, 1990.

[192] L. Kaufman y P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[193] H. Yin, «The self-organizing maps: background, theories, extensions and applications,» en *Computational intelligence: A compendium*, Springer, 2008, págs. 715-762.

[194] G. P. Zhang, «Neural networks for classification: a survey,» *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, n.º 4, págs. 451-462, 2000.

[195] A. Ben-Dor, R. Shamir y Z. Yakhini, «Clustering gene expression patterns,» *Journal of computational biology*, vol. 6, n.º 3-4, págs. 281-297, 1999.

[196] A. Tanay, R. Sharan y R. Shamir, «Biclustering algorithms: A survey,» *Handbook of computational molecular biology*, vol. 9, n.º 1-20, págs. 122-124, 2005.

[197] J. Colcombet y H. Hirt, «Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes,» *Biochemical Journal*, vol. 413, n.º 2, págs. 217-226, 2008.

[198] S. C. Madeira y A. L. Oliveira, «Biclustering algorithms for biological data analysis: a survey,» *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 1, n.º 1, págs. 24-45, 2004.

[199] Y Cheng y G. Church, «Biclustering of expression data. In Proceedings of Intelligent Systems for Molecular Biology,» *2000*, 2000.

[200] K. Y. Yip, D. W. Cheung y M. K. Ng, «Harp: A practical projected clustering algorithm,» *IEEE Transactions on knowledge and data engineering*, vol. 16, n.º 11, págs. 1387-1397, 2004.

[201] T. Yun y G.-S. Yi, «Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion,» *BMC genomics*, vol. 14, n.º 1, págs. 1-15, 2013.

[202] J. Yang, H. Wang, W. Wang y P. S. Yu, «An improved biclustering method for analyzing gene expression profiles,» *International Journal on Artificial Intelligence Tools*, vol. 14, n.º 05, págs. 771-789, 2005.

[203] F. Angiulli, E. Cesario y C. Pizzuti, «Random walk biclustering for microarray data,» *Information Sciences*, vol. 178, n.º 6, págs. 1479-1497, 2008.

[204] S. Dharan y A. S. Nair, «Biclustering of gene expression data using reactive greedy randomized adaptive search procedure,» *BMC bioinformatics*, vol. 10, n.º 1, págs. 1-10, 2009.

[205] W. Ayadi, M. Elloumi y J.-K. Hao, «Pattern-driven neighborhood search for biclustering of microarray data,» en *BMC bioinformatics*, Springer, vol. 13, 2012, págs. 1-11.

[206] S. Kirkpatrick, C. D. Gelatt y M. P. Vecchi, «Optimization by simulated annealing,» *science*, vol. 220, n.º 4598, págs. 671-680, 1983.

[207] J. Liu, Z. Li, X. Hu e Y. Chen, «Biclustering of microarray data with MOSPO based on crowding distance,» en *BMC bioinformatics*, BioMed Central, vol. 10, 2009, págs. 1-10.

[208] G. P. Coelho, F. O. de França y F. J. Von Zuben, «Multi-objective biclustering: When non-dominated solutions are not enough,» *Journal of Mathematical Modelling and Algorithms*, vol. 8, n.º 2, págs. 175-202, 2009.

[209] S. Bleuler, A. Prelic y E. Zitzler, «An EA framework for biclustering of gene expression data,» en *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753)*, IEEE, vol. 1, 2004, págs. 166-173.

[210] S. Mitra y H. Banka, «Multi-objective evolutionary biclustering of gene expression data,» *Pattern Recognition*, vol. 39, n.º 12, págs. 2464-2477, 2006.

[211] C. Cano, L Adarve, J López y A. Blanco, «Possibilistic approach for biclustering microarray data,» *Computers in biology and medicine*, vol. 37, n.º 10, págs. 1426-1436, 2007.

[212] W.-H. Yang, D.-Q. Dai y H. Yan, «Finding correlated biclusters from gene expression data,» *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, n.º 4, págs. 568-584, 2010.

[213] D. Yan y J. Wang, «Biclustering of gene expression data based on related genes and conditions extraction,» *Pattern recognition*, vol. 46, n.º 4, págs. 1170-1182, 2013.

[214] G. Getz, E. Levine y E. Domany, «Coupled two-way clustering analysis of gene microarray data,» *Proceedings of the National Academy of Sciences*, vol. 97, n.º 22, págs. 12 079-12 084, 2000.

[215] C. Tang y A. Zhang, «Interrelated two-way clustering and its application on gene expression data,» *International Journal on Artificial Intelligence Tools*, vol. 14, n.º 04, págs. 577-597, 2005.

[216] A. Tanay, R. Sharan y R. Shamir, «Discovering statistically significant biclusters in gene expression data,» *Bioinformatics*, vol. 18, n.º suppl_1, S136-S144, 2002.

[217] L. Zhao y M. J. Zaki, «Microcluster: Efficient deterministic biclustering of microarray data,» *IEEE Intelligent Systems*, vol. 20, n.º 6, págs. 40-49, 2005.

[218] G. Li, Q. Ma, H. Tang, A. H. Paterson e Y. Xu, «QUBIC: a qualitative biclustering algorithm for analyses of gene expression data,» *Nucleic acids research*, vol. 37, n.º 15, e101-e101, 2009.

[219] S. Roy, D. K. Bhattacharyya y J. K. Kalita, «Cobi: pattern based co-regulated biclustering of gene expression data,» *Pattern Recognition Letters*, vol. 34, n.º 14, págs. 1669-1678, 2013.

[220] L. Lazzeroni y A. Owen, «Plaid models for gene expression data,» *Statistica sinica*, págs. 61-86, 2002.

[221] E. Segal, B. Taskar, A. Gasch, N. Friedman y D. Koller, «Rich probabilistic models for gene expression,» *Bioinformatics*, vol. 17, n.º suppl_1, S243-S252, 2001.

[222] Q. Sheng, Y. Moreau y B. De Moor, «Biclustering microarray data by Gibbs sampling,» *Bioinformatics*, vol. 19, n.º suppl_2, págs. ii196-ii205, 2003.

[223] T. Chekouo y A. Murua, «The penalized biclustering model and related algorithms,» *Journal of Applied Statistics*, vol. 42, n.º 6, págs. 1255-1277, 2015.

[224] Y. Kluger, R. Basri, J. T. Chang y M. Gerstein, «Spectral biclustering of microarray data: coclustering genes and conditions,» *Genome research*, vol. 13, n.º 4, págs. 703-716, 2003.

[225] S. Bergmann, J. Ihmels y N. Barkai, «Iterative signature algorithm for the analysis of large-scale gene expression data,» *Physical review E*, vol. 67, n.º 3, pág. 031 902, 2003.

[226] R. Henriques y S. C. Madeira, «BicPAM: Pattern-based biclustering for biomedical data analysis,» *Algorithms for Molecular Biology*, vol. 9, n.º 1, págs. 1-30, 2014.

[227] A. Ben-Dor, B. Chor, R. Karp y Z. Yakhini, «Discovering local structure in gene expression data: the order-preserving submatrix problem,» en *Proceedings of the sixth annual international conference on Computational biology*, 2002, págs. 49-57.

[228] M. Harrell, J. Xia y Z. Zhao, «Network analysis of gene fusions in human cancer,» en *BMC bioinformatics*, BioMed Central, vol. 14, 2013, A13.

[229] B. Ristevski, «A survey of models for inference of gene regulatory networks,» *Nonlinear Anal Model Control*, vol. 18, n.º 4, págs. 444-65, 2013.

[230] D. J. Watts y S. H. Strogatz, «Collective dynamics of 'small-world'networks,» *nature*, vol. 393, n.º 6684, págs. 440-442, 1998.

[231] A. D. Broido y A. Clauset, «Scale-free networks are rare,» *Nature communications*, vol. 10, n.º 1, págs. 1-10, 2019.

[232] K. Raman, N. Damaraju y G. K. Joshi, «The organisational structure of protein networks: revisiting the centrality–lethality hypothesis,» *Systems and Synthetic Biology*, vol. 8, n.º 1, págs. 73-81, 2014.

[233] V. A. Huynh-Thu y G. Sanguinetti, «Gene regulatory network inference: an introductory survey,» en *Gene Regulatory Networks*, Springer, 2019, págs. 1-23.

[234] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris y Z. Zakaria, «A review on the computational approaches for gene regulatory network construction,» *Computers in biology and medicine*, vol. 48, págs. 55-65, 2014.

[235] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren y R. Guthke, «Gene regulatory network inference: Data integration in dynamic models – A review,» *Biosystems*, vol. 96, n.º 1, págs. 86-103, 2009.

[236] F. Borelli, R. de Camargo, D. Martins y L. Rozante, «Gene regulatory networks inference using a multi-GPU exhaustive search algorithm,» *BMC Bioinformatics*, vol. 14, n.º Suppl 18, S5, 2013.

[237] F. Emmert-Streib, M. Dehmer y B. Haibe-Kains, «Untangling statistical and biological models to understand network inference: the need for a genomics network ontology,» *Frontiers in genetics*, vol. 5, pág. 299, 2014.

[238] B. R. Borate, E. J. Chesler, M. A. Langston, A. M. Saxton y B. H. Voy, «Comparison of threshold selection methods for microarray gene co-expression matrices,» *BMC research notes*, vol. 2, n.º 1, págs. 1-6, 2009.

[239] A. J. Butte e I. S. Kohane, «Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements,» en *Biocomputing 2000*, World Scientific, 1999, págs. 418-429.

[240] A. A. Margolin, I. Nemenman, K. Basso y col., «ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,» en *BMC bioinformatics*, Springer, vol. 7, 2006, págs. 1-15.

[241] R. A. C. Montes, G. Coello, K. L. González-Aguilera, N. Marsch-Martínez, S. de Folter y E. R. Alvarez-Buylla, «ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks,» *BMC plant biology*, vol. 14, n.º 1, págs. 1-14, 2014.

[242] A. Lachmann, F. M. Giorgi, G. Lopez y A. Califano, «ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information,» *Bioinformatics*, vol. 32, n.º 14, págs. 2233-2235, 2016.

[243] A. Madar, A. Greenfield, E. Vanden-Eijnden y R. Bonneau, «DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator,» *PloS one*, vol. 5, n.º 3, e9803, 2010.

[244] C. Olsen, P. E. Meyer y G. Bontempi, «On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information,» *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, págs. 1-9, 2008.

[245] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel y P. Geurts, «Inferring regulatory networks from expression data using tree-based methods,» *PloS one*, vol. 5, n.º 9, e12776, 2010.

[246] V. A. Huynh-Thu y G. Sanguinetti, «Combining tree-based and dynamical systems for the inference of gene regulatory networks,» *Bioinformatics*, vol. 31, n.º 10, págs. 1614-1622, 2015.

[247] P. Geurts y col., «dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data,» *Scientific reports*, vol. 8, n.º 1, págs. 1-12, 2018.

[248] G. Michailidis y F. d'Alché Buc, «Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues,» *Mathematical biosciences*, vol. 246, n.º 2, págs. 326-334, 2013.

[249] J. Schäfer y K. Strimmer, «An empirical Bayes approach to inferring large-scale gene association networks,» *Bioinformatics*, vol. 21, n.º 6, págs. 754-764, 2005.

[250] S. Ma, Q. Gong y H. J. Bohnert, «An Arabidopsis gene network based on the graphical Gaussian model,» *Genome research*, vol. 17, n.º 11, págs. 1614-1625, 2007.

[251] A. Wille, P. Zimmermann, E. Vranová y col., «Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana,» *Genome biology*, vol. 5, n.º 11, págs. 1-13, 2004.

[252] S. Bornholdt, «Boolean network models of cellular regulation: prospects and limitations,» *Journal of the Royal Society Interface*, vol. 5, n.º suppl_1, S85-S94, 2008.

[253] R. Thomas, «Boolean formalization of genetic control circuits,» *Journal of theoretical biology*, vol. 42, n.º 3, págs. 563-585, 1973.

[254] N. Maheshri y E. K. O'Shea, «Living with noisy genes: how cells function reliably with inherent variability in gene expression,» *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, págs. 413-434, 2007.

[255] A. A. Melkman, X. Cheng, W.-K. Ching y T. Akutsu, «Identifying a probabilistic Boolean threshold network from samples,» *IEEE transactions on neural networks and learning systems*, vol. 29, n.º 4, págs. 869-881, 2017.

[256] I. Shmulevich, E. R. Dougherty, S. Kim y W. Zhang, «Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks,» *Bioinformatics*, vol. 18, n.º 2, págs. 261-274, 2002.

[257] A. Larjo, I. Shmulevich y H. Lähdesmäki, «Structure learning for Bayesian networks as models of biological networks,» en *Data Mining for Systems Biology*, Springer, 2013, págs. 35-45.

[258] E. Acerbi, T. Zelante, V. Narang y F. Stella, «Gene network inference using continuous time Bayesian networks: a comparative study and application to Th17 cell differentiation,» *BMC bioinformatics*, vol. 15, n.º 1, págs. 1-27, 2014.

[259] T. Chekouo, F. C. Stingo, J. D. Doecke y K.-A. Do, «miRNA–target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer,» *Biometrics*, vol. 71, n.º 2, págs. 428-438, 2015.

[260] D. Chudasama, V. Bo, M. Hall y col., «Identification of novel cancer biomarkers of prognostic value using specific gene regulatory networks (GRN): a novel role of RAD51AP1 for ovarian and lung cancers,» *Carcinogenesis*, 2017.

[261] S. M. Hill, Y. Lu, J. Molina y col., «Bayesian inference of signaling network topology in a cancer cell line,» *Bioinformatics*, vol. 28, n.º 21, págs. 2804-2810, 2012.

[262] E. S. Adabor y G. K. Acquaah-Mensah, «Restricted-derestricted dynamic Bayesian Network inference of transcriptional regulatory relationships among genes in cancer,» *Computational biology and chemistry*, vol. 79, págs. 155-164, 2019.

[263] S. Y. Kim, S. Imoto y S. Miyano, «Inferring gene networks from time series microarray data using dynamic Bayesian networks,» *Briefings in bioinformatics*, vol. 4, n.º 3, págs. 228-235, 2003.

[264] Z. Li, P. Li, A. Krishnan y J. Liu, «Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis,» *Bioinformatics*, vol. 27, n.º 19, págs. 2686-2691, 2011.

[265] E. R. Morrissey, M. A. Juárez, K. J. Denby y N. J. Burroughs, «On reverse engineering of gene interaction networks using time course data with repeated measurements,» *Bioinformatics*, vol. 26, n.º 18, págs. 2305-2312, 2010.

[266] T. Äijö y H. Lähdesmäki, «Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics,» *Bioinformatics*, vol. 25, n.º 22, págs. 2937-2944, 2009.

[267] M. Grzegorczyk y D. Husmeier, «Non-homogeneous dynamic Bayesian networks for continuous data,» *Machine Learning*, vol. 83, n.º 3, págs. 355-419, 2011.

[268] M. Bansal, G. D. Gatta y D. Di Bernardo, «Inference of gene regulatory networks and compound mode of action from time course gene expression profiles,» *Bioinformatics*, vol. 22, n.º 7, págs. 815-822, 2006.

[269] E. O. Voit, *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

[270] R. Xu, X. Hu y D. C. Wunsch, «Inference of genetic regulatory networks with recurrent neural network models,» en *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, vol. 2, 2004, págs. 2905-2908.

[271] D. Zafeiris, S. Rutella y G. R. Ball, «An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study,» *Computational and structural biotechnology journal*, vol. 16, págs. 77-87, 2018.

[272] Y. Yuan y Z. Bar-Joseph, «Deep learning for inferring gene relationships from single-cell expression data,» *Proceedings of the National Academy of Sciences*, vol. 116, n.º 52, págs. 27 151-27 158, 2019.

[273] D. L. Tong, D. J. Boocock, G. K. R. Dhondalay, C. Lemetre y G. R. Ball, «Artificial neural network inference (ANNI): a study on gene-gene interaction for biomarkers in childhood sarcomas,» *PLoS One*, vol. 9, n.º 7, e102483, 2014.

[274] A. Blanco, M. Delgado y M. Pegalajar, «A genetic algorithm to obtain the optimal recurrent neural network,» *International Journal of Approximate Reasoning*, vol. 23, n.º 1, págs. 67-83, 2000.

[275] C. A. Penfold, V. Buchanan-Wollaston, K. J. Denby y D. L. Wild, «Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks,» *Bioinformatics*, vol. 28, n.º 12, págs. i233-i241, 2012.

[276] A. Ahmed y E. P. Xing, «Recovering time-varying networks of dependencies in social and biological studies,» *Proceedings of the National Academy of Sciences*, vol. 106, n.º 29, págs. 11 878-11 883, 2009.

[277] J. W. Robinson, A. J. Hartemink y Z. Ghahramani, «Learning Non-Stationary Dynamic Bayesian Networks.,» *Journal of Machine Learning Research*, vol. 11, n.º 12, 2010.

[278] T. Thorne y M. P. Stumpf, «Inference of temporally varying Bayesian networks,» *Bioinformatics*, vol. 28, n.º 24, págs. 3298-3305, 2012.

[279] F. Dondelinger, S. Lèbre y D. Husmeier, «Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure,» *Machine Learning*, vol. 90, n.º 2, págs. 191-230, 2013.

[280] A. Ceglar y J. F. Roddick, «Association mining,» *ACM Computing Surveys (CSUR)*, vol. 38, n.º 2, 5-es, 2006.

[281] C. Creighton y S. Hanash, «Mining gene expression databases for association rules,» *Bioinformatics*, vol. 19, n.º 1, págs. 79-86, 2003.

[282] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo y A. Pascual-Montano, «Integrated analysis of gene expression by association rules discovery,» *BMC bioinformatics*, vol. 7, n.º 1, págs. 1-16, 2006.

[283] C. A. Gallo, J. A. Carballido e I. Ponzoni, «Discovering time-lagged rules from microarray data using gene profile classifiers,» *BMC bioinformatics*, vol. 12, n.º 1, pág. 123, 2011.

[284] B. Haibe-Kains y F. Emmert-Streib, «Quantitative assessment and validation of network inference methods in bioinformatics,» *Frontiers in genetics*, vol. 5, pág. 221, 2014.

[285] N. X. Vinh, M. Chetty, R. Coppel y P. P. Wangikar, «Issues impacting genetic network reverse engineering algorithm validation using small networks,» *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1824, n.º 12, págs. 1434-1441, 2012.

[286] E. R. Dougherty, «Validation of gene regulatory networks: scientific and inferential.,» *Briefings in Bioinformatics*, vol. 12, n.º 3, págs. 245-252, 2011.

[287] B. Haibe-Kains y F. Emmert-Streib, *Quantitative assessment and validation of network inference methods in bioinformatics*. Frontiers Media SA, 2015.

[288] H. Xu, Y.-S. Ang, A. Sevilla, I. R. Lemischka y A. Ma'ayan, «Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells,» *PLoS Comput Biol*, vol. 10, n.º 8, e1003777, 2014.

[289] C. Olsen, K. Fleming, N. Prendergast y col., «Inference and validation of predictive gene networks from biomedical literature and gene expression data,» *Genomics*, vol. 103, n.º 5, págs. 329-336, 2014.

[290] X. Wu, R. Jiang, M. Q. Zhang y S. Li, «Network-based global inference of human disease genes,» *Molecular systems biology*, vol. 4, n.º 1, pág. 189, 2008.

[291] T. Van den Bulcke, K. Van Leemput, B. Naudts y col., «SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms,» *BMC bioinformatics*, vol. 7, n.º 1, págs. 1-12, 2006.

[292] S. Ganscha, V. Fortuin, M. Horn, E. Arvaniti y M. Claassen, «Supervised learning on synthetic data for reverse engineering gene regulatory networks from experimental time-series,» *bioRxiv*, pág. 356 477, 2018.

[293] W. Winterbach, P. V. Mieghem, M. Reinders, H. Wang y D. d. Ridder, «Topology of molecular interaction networks,» *BMC systems biology*, vol. 7, n.º 1, págs. 1-15, 2013.

[294] J. Sławek y T. Arodź, «ENNET: inferring large gene regulatory networks from expression data using gradient boosting,» *BMC systems biology*, vol. 7, n.º 1, pág. 1, 2013.

[295] L. Franke, H. Van Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen y C. Wijmenga, «Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes,» *The American Journal of Human Genetics*, vol. 78, n.º 6, 1011omranian1025, 2006.

[296] U. Mansmann y V. Jurinovic, «Biological feature validation of estimated gene interaction networks from microarray data: a case study on MYC in lymphomas,» *Briefings in bioinformatics*, vol. 12, n.º 3, págs. 230-244, 2011.

[297] M. C. Teixeira, P. T. Monteiro, J. F. Guerreiro y col., «The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae,» *Nucleic acids research*, gkt1015, 2013.

[298] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta y col., «RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond,» *Nucleic acids research*, gkv1156, 2015.

[299] I. M. Keseler, J. Collado-Vides, S. Gama-Castro y col., «EcoCyc: a comprehensive database resource for Escherichia coli,» *Nucleic acids research*, vol. 33, n.º suppl 1, págs. D334-D337, 2005.

[300] G. Altay, N. Altay y D. Neal, «Global assessment of network inference algorithms based on available literature of gene/protein interactions,» *Turkish Journal of Biology*, vol. 37, n.º 5, págs. 547-555, 2013.

[301] X. Lin, M. Liu y X.-w. Chen, «Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms,» en *BMC bioinformatics*, BioMed Central, vol. 10, 2009, págs. 1-14.

[302] E. R Dougherty e I. Shmulevich, «On the limitations of biological knowledge,» *Current genomics*, vol. 13, n.º 7, págs. 574-587, 2012.

[303] E. Dougherty y X. Qian, «Validation of gene regulatory network inference based on controllability,» *Frontiers in genetics*, vol. 4, pág. 272, 2013.

[304] E. S. H. Cheow, W. C. Cheng, C. N. Lee, D. De Kleijn, V. Sorokin y S. K. Sze, «Plasma-derived extracellular vesicles contain predictive biomarkers and potential therapeutic targets for myocardial ischemic (MI) injury,» *Molecular & Cellular Proteomics*, vol. 15, n.º 8, págs. 2628-2640, 2016.

[305] X. Dong, A. Yambartsev, S. A. Ramsey, L. D. Thomas, N. Shulzhenko y A. Morgun, «Reverse enGENEering of regulatory networks from big data: a roadmap for biologists,» *Bioinformatics and biology insights*, vol. 9, BBI-S12467, 2015.

[306] A. Alexeyenko, W. Lee, M. Pernemalm y col., «Network enrichment analysis: extension of gene-set enrichment analysis to gene networks,» *BMC bioinformatics*, vol. 13, n.º 1, págs. 1-11, 2012.

[307] T. McCormack, O. Frings, A. Alexeyenko y E. L. Sonnhammer, «Statistical assessment of crosstalk enrichment between gene groups in biological networks,» *PloS one*, vol. 8, n.º 1, e54945, 2013.

[308] M. Pathan, S. Keerthikumar, C.-S. Ang y col., «FunRich: An open access standalone functional enrichment and interaction network analysis tool,» *Proteomics*, vol. 15, n.º 15, págs. 2597-2601, 2015.

[309] M. Gan, X. Dou y R. Jiang, «From ontology to semantic similarity: calculation of ontology-based semantic similarity,» *The Scientific World Journal*, vol. 2013, 2013.

[310] P. H. Guzzi, M. Mina, C. Guerra y M. Cannataro, «Semantic similarity analysis of protein data: assessment with biological features and issues,» *Briefings in bioinformatics*, vol. 13, n.º 5, págs. 569-585, 2012.

[311] P. Resnik, «Using information content to evaluate semantic similarity in a taxonomy,» *arXiv preprint cmp-lg/9511007*, 1995.

[312] D. Lin y col., «An information-theoretic definition of similarity.,» en *Icml*, vol. 98, 1998, págs. 296-304.

[313] J. J. Jiang y D. W. Conrath, «Semantic similarity based on corpus statistics and lexical taxonomy,» *arXiv preprint cmp-lg/9709008*, 1997.

[314] P. W. Lord, R. D. Stevens, A. Brass y C. A. Goble, «Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation,» *Bioinformatics*, vol. 19, n.º 10, págs. 1275-1283, 2003.

[315] P. W. Lord, R. D. Stevens, A. Brass y C. A. Goble, «Semantic similarity measures as tools for exploring the gene ontology,» en *Biocomputing 2003*, World Scientific, 2002, págs. 601-612.

[316] A. Schlicker, F. S. Domingues, J. Rahnenführer y T. Lengauer, «A new measure for functional similarity of gene products based on Gene Ontology,» *BMC bioinformatics*, vol. 7, n.º 1, págs. 1-16, 2006.

[317] F. M. Couto, M. J. Silva y P. M. Coutinho, «Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors,» en *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, págs. 343-344.

[318] O. Bodenreider, M. Aubry y A. Burgun, «Non-lexical approaches to identifying associative relations in the gene ontology,» en *Biocomputing 2005*, World Scientific, 2005, págs. 91-102.

[319] Y. Xu, M. Guo, W. Shi, X. Liu y C. Wang, «A novel insight into Gene Ontology semantic similarity,» *Genomics*, vol. 101, n.º 6, págs. 368-375, 2013.

[320] G. K. Mazandu y N. J. Mulder, «A topology-based metric for measuring term similarity in the gene ontology,» *Advances in bioinformatics*, vol. 2012, 2012.

[321] M. A. Alvarez y C. Yan, «A graph-based semantic similarity measure for the gene ontology,» *Journal of bioinformatics and computational biology*, vol. 9, n.º 06, págs. 681-695, 2011.

[322] Q. Hu, Z. Wang y Z. Zhang, «FSim: a novel functional similarity search algorithm and tool for discovering functionally related gene products,» *BioMed research international*, vol. 2014, 2014.

[323] V. Pekar y S. Staab, «Taxonomy learning-factoring the structure of a taxonomy into a semantic classification decision,» en *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[324] J. Cheng, M. Cline, J. Martin y col., «A knowledge-based clustering algorithm driven by gene ontology,» *Journal of biopharmaceutical statistics*, vol. 14, n.º 3, págs. 687-700, 2004.

[325] H. Wu, Z. Su, F. Mao, V. Olman e Y. Xu, «Prediction of functional modules based on comparative genome analysis and Gene Ontology application,» *Nucleic acids research*, vol. 33, n.º 9, págs. 2822-2837, 2005.

[326] X. Wu, L. Zhu, J. Guo, D.-Y. Zhang y K. Lin, «Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations,» *Nucleic acids research*, vol. 34, n.º 7, págs. 2137-2150, 2006.

[327] A. Nagar y H. Al-Mubaid, «A hybrid semantic similarity measure for gene ontology based on offspring and path length,» en *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, 2015, págs. 1-7.

[328] J. Peng, H. Li, Q. Jiang, Y. Wang y J. Chen, «An integrative approach for measuring semantic similarities using gene ontology,» *BMC systems biology*, vol. 8, n.º 5, págs. 1-12, 2014.

[329] S.-B. Zhang y Q.-R. Tang, «Protein–protein interaction inference based on semantic similarity of gene ontology terms,» *Journal of theoretical biology*, vol. 401, págs. 30-37, 2016.

[330] S.-B. Zhang y J.-H. Lai, «A hybrid measure for the semantic similarity of gene ontology terms,» en *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*, IEEE, 2014, págs. 911-916.

[331] C. Bettembourg, C. Diot y O. Dameron, «Semantic particularity measure for functional characterization of gene sets using gene ontology,» *PloS one*, vol. 9, n.º 1, e86525, 2014.

[332] B. Liu, M. Jin y P. Zeng, «Prioritization of candidate disease genes by combining topological similarity and semantic similarity,» *Journal of biomedical informatics*, vol. 57, págs. 1-5, 2015.

[333] J. C. Jeong y X. Chen, «A new semantic functional similarity over gene ontology,» *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 12, n.º 2, págs. 322-334, 2014.

[334] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang y P. Xuan, «Measuring gene functional similarity based on group-wise comparison of GO terms,» *Bioinformatics*, vol. 29, n.º 11, págs. 1424-1432, 2013.

[335] M. Mistry y P. Pavlidis, «Gene Ontology term overlap as a measure of gene functional similarity,» *BMC bioinformatics*, vol. 9, n.º 1, págs. 1-11, 2008.

[336] M. Zhao, W. He, J. Tang, Q. Zou y F. Guo, «A comprehensive overview and critical evaluation of gene regulatory network inference technologies,» *Briefings in Bioinformatics*, 2021.