Tesis Doctoral

Análisis de redes complejas en sistemas biomoleculares

Berenstein, Ariel José

2014-12-22

UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Física

Análisis de redes complejas en sistemas biomoleculares.

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Físicas

Lic. Ariel José Berenstein

Director de tesis: Chernomoretz Ariel
Consejero de estudios: Wisniacki Diego A.

Lugar de trabajo: Depto de Física, FCEyN, UBA

Buenos Aires, 2014

Fecha de defensa: 22/12/2014

Declaración de Autoría

Por la presente declaro que el trabajo presente ha sido escrito por mí, que es el registro del trabajo realizado durante los años 2010-2014 y que no ha sido presentada previamente en ninguna otra universidad para la obtención de un título de doctorado.

Firma:

Fecha:

Resumen

Dentro de una célula coexisten diferentes tipos de biomoléculas que participan en intrincadas redes de interacciones físicas y bioquímicas. Las mismas facilitan la supervivencia de la célula y hacen de ella un sistema extremadamente complejo. La descripción de estas interacciones bajo la perspectiva de redes complejas ofrece la posibilidad de estudiar propiedades colectivas emergentes, detectar patrones de organización y proveer una visión global de la célula como sistema. Estas redes presentan estructura modular no trivial. Numerosos trabajos han puesto esfuerzos en correlacionar esta estructura modular con grupos de biomoléculas que llevan a cabo funciones biológicas específicas. No obstante, un problema de ese enfoque es el hecho de que la estructura modular observada en una red depende de la escala adoptada así como de los algoritmos de agrupamiento utilizados.

Esta tesis hace especial énfasis en comprender cómo distintos algoritmos de agrupamiento y sus niveles de resolución implícitos pueden afectar a los análisis biológicos subsecuentes. Se consideró una red de interacción de proteínas, y distintos conjuntos de proteínas involucradas en procesos de envejecimiento celular y vías de señalización. Se emplearon dos algoritmos de agrupamiento ampliamente reconocidos, uno basado en teoría de información y otro en optimización de la modularidad de la red. Mientras que el primer tipo es capaz de detectar estructuras topológicas libres de escala característica, el segundo tiene límite de resolución bien definido. Los resultados obtenidos sugieren que si bien ambos algoritmos obtienen particiones de similar modularidad, las mismas difieren significativamente en el nivel de congruencia biológica, el grado de granularidad de las descripciones modulares y en la capacidad para detectar asociaciones estadísticamente significativas entre los conjuntos de proteínas considerados y los respectivos roles cartográficos de la red.

Por otro lado se abordó el problema de reposicionamiento de fármacos en el contexto de enfermedades tropicales desatendidas. Para ello, se construyó y caracterizó una red multicapa compuesta por fármacos, proteínas de múltiples especies y diversos tipos de relaciones estructurales, químicas, metabólicas y de bioactividad. Empleando técnicas de priorización en redes complejas se abordaron dos problemas biológicos de interés. Por un lado la priorización de potenciales blancos de droga en un organismo patógeno de interés. Por otro lado la busqueda de blancos de drogas con actividad probada sobre un organismo, pero cuyos mecanismos y blancos de acción permanecen desconocidos.

Los resultados fueron validados computacionalmente y discutidos desde un punto de vista biológico en base a bibliografía reciente. En suma los resultados indican que el enfoque adoptado resulta de gran utilidad para guiar experimentos que busquen entender los mecanismos de acción de drogas que aún permanecen desconocidos.

Palabras Clave: Redes complejas - Redes de interacción de proteínas - Redes multicapa - Reposicionamiento de Fármacos - Enfermedades Tropicales -

Abstract

Complex Network analysis in biomolecular systems

Inside a cell different types of biomolecules coexist, which are involved in intricate networks of physical and biochemical interactions. They facilitate the cell survival and make it an extremely complex system. The description of these interactions from complex network perspective allows to study emerging collective properties, to detect organization patterns and provides an overview of the cell as a system. These networks have nontrivial modular structure. Several works have concentrated on correlating this modular structure with groups of biomolecules that carry out specific biological functions. However, a concerning issue of this approach is that observed modular structure in a network depends on the adopted scale as well as the clustering algorithms used.

This thesis is focused on understanding how different clustering algorithms and their implicit resolution levels may affect subsequent biological analysis. To this end, a protein-protein interaction network was considered and examined several protein sets related to cell aging and signaling pathways. Two clustering algorithms widely recognized were considered, one based on information theory and other based on network modularity optimization. While the first one is capable of scale-free recognition of topological structure, the second one has a well-defined resolution limit. The results suggest that although both algorithms result in partitions of similar modularity, they differ significantly on its biological congruence, the granularity of modular descriptions and its ability to detect statistically significant associations between considered protein sets and network cartographic roles.

On the other hand the problem of drug repositioning is addressed in context of neglected tropical diseases. To achieve this, a multilayer network consisting of chemical compounds, proteins from multiple species and several classes of structural, chemical, and metabolic relationships as well as bioactivities between them was constructed and characterized. Taking advantage of prioritization techniques in complex networks, two interesting biological issues were addressed. Firstly, the prioritization of putative drug-targets in a given pathogenic species was tackled. Secondly, protein targets were looked for on tested active compounds whose mechanisms of action and targets remained unknown.

The results were validated computationally and discussed from a biological point of view taking into account recent literature. In summary, the results suggest that the approach is useful for guiding experiments that seek to understand drugs mechanisms of action that are still unknown.

Keywords: Complex networks - Protein Protein Interaction Networks - Multilayer Networks - Drug Repositioning - Tropical Diseases -

Agradecimientos

Gracias a vos Barby todo este camino fue más llevadero. Sos vos la que me sostuvo cuando estaba cayendo, quien me apoyó, aconsejó y permitió ver siempre la realidad desde otro punto de vista. Especialmente la realidad académica.Gracias por hacerme feliz cada día. Te amo.

Gracias a vos viejita por tu amor incondicional, por haber sabido inculcarme los principios, la moral y la cultura de trabajo. Sos vos la que me enseñó que sin esfuerzo nada que valga la pena se alcanza y lo hiciste de la manera más contundente posible: con el ejemplo. Gracias también a vos hermanita por hacerme sentir y valorar el verdadero valor que tiene la familia.

Sin duda son mis amigos y hermanos del alma Eddy, Eze, Maurito y Mati, los que estuvieron siempre a mi lado en cada momento difícil. Gracias por compartir las alegrías, las penas y sobre todo por hacerme crecer como persona. Eddy, vos sabés que nos tocó transitar juntos esta etapa con historias asombrosamente paralelas. Sin duda nuestras charlas y tus consejos han sido esenciales para afrontar las etapas más difíciles de este doctorado.

Tuve la dicha de compartir los días de trabajo junto a un grupo extraordinario de personas. La Dra Cristina Marino ha sido compañera de laboratorio durante casi 5 años. Sin duda el día a día hubiera sido muchísimo más duro sin tu presencia, tus sonrisas y tu extraordinaria calidez cotidiana. Fuiste como una madre en la oficina cada día. Voy a extrañarte. Gracias a ustedes Elin, Dieguito, Ire, Franquito, Estefi, Javi y Martín y Santiago por haber hecho de este doctorado un proceso más agradable. Gracias por los momentos dentro y fuera de la oficina, gracias por los consejos el apoyo y por sobre todo por la infinita tolerancia que han tenido para convivir conmigo a lo largo de estos años. Especialmente gracias a ustedes Elin, Dieguito y Santiago, por haberme permitido encontrar tres amigos con los que pude compartir mucho más que un doctorado.

Gracias a mi supervisor, Ariel, por haberme ofrecido todo lo que tubo a su alcance. Por brindarme las incontables horas de discusión, de análisis y profundo compromiso sobre los problemas que juntos abordamos a lo largo de estos años de doctorado. Gracias por *poner el pecho* durante todo el proceso.

Quiero agradecer a la Dra Laura Furlong, al Dr. Fernan Aguero, a la Lic. Janet Piñero, y a la Dra Paula Magariños quienes han sido parte de incontables discusiones, y esfuerzos conjuntos que han dado lugar a que esta tesis haya alcanzado el grado interdisciplinario que tiene. Gracias Janet por tu tenacidad, por no bajar nunca los brazos en un camino que nos costó muchísimo esfuerzo recorrer a ambos.

Gracias a los doctores Pablo Balenzuela y Osvaldo Podhacer por formar parte de mi comité de seguimiento de tesis, por asistir a la presentación de informes anuales, por sus críticas constructivas y por sugerir provechosos cambios a lo largo de estos años.

Gracias a todo el personal del instituto Leloir, el esfuerzo conjunto de todos los que forman parte de esta fundación hicieron que trabajar aquí haya sido tan productivo como placentero.

Prólogo del autor

Esta tesis es el resultado de muchos años de trabajo y aprendizaje. Es sólo una diminuta parte visible de un enorme esfuerzo velado.

He aquí los trabajos que según creo, proporcionan los resultados más interesantes generados a lo largo de estos años. Muchos de ellos estoy convencido que merecen ser comunicados más allá de esta tesis, mientras que otros naturalmente, fueron formulados sólo a efectos de completitud de este manuscrito o por el mero placer de generarlos. Hasta hace pocas horas, esta tesis no contaba con un dictamen internacional que *avale* la relevancia o *veracidad* de su contenido. Ahora, gran parte del capítulo 4 cuenta con este aval, mientras que la parte restante, al igual que el contenido entero del capítulo 5 esperan pacientemente su oportunidad de ser presentados a la comunidad científica internacional.

Sin embargo, a lo largo de este proceso hubieron también muchos desaciertos, trabajos inconclusos, o hasta algunos mal planteados desde un primer momento. Estoy convencido que todos esos errores y sinfonías inconclusas que en este manuscrito no se muestran fueron los principales responsables de mi proceso de formación. Ocuparon horas enteras y días de intensa discusión con mi supervisor, así como meses de desarrollo y dilatación de pupilas frente al monitor. Algunas de estas sinfonías inconclusas involucran estudios de perfiles de expresión génica, de ontologías genéticas, de métricas biológicas, o análisis estadísticos de algoritmos de priorización en redes complejas, por mencionar las más relevantes.

Esos *desaciertos* tienen para mí, el enorme valor intangible que involucra *darse cuenta* que uno estaba equivocado. Un valor que subyace en el ineludible proceso de aprendizaje que requiere advertir los propios errores. Un proceso que exige profundizar en cada detalle para encontrar a menudo que lo que creíamos entender no lo entendíamos y que aquello que creíamos un interesante logro digno de publicar era en verdad una falacia o una simple trivialidad que ya no merece la pena ser comunicada.

Esas son las historias que este manuscrito no cuenta, pero que existieron y fueron las que me permitieron adquirir las herramientas necesarias para construir la siguiente *tesis doctoral*. Espero que la disfruten.

Índice general

Abreviaturas

AUC	**Á**rea bajo una **C**urva ROC
BHI	**Í**ndice de **H**omogeneidad **B**iológica
BP	**P**rocesos **B**iológicos
CNM	Algoritmo de agrupamiento **C**lauset **N**ewman **M**oore
EC	**E**nvejecimiento **C**elular
ER	**E**rdös **R**enyi
FF	Algoritmo de **F**lujo **F**uncional
HIPPIE	**H**uman **I**ntegrated **P**rotein-**P**rotein **I**nteraction r**E**ference
NTD	Enfermedades Tropicales Desatendidas
PIN	**R**ed de **I**nteracción de **P**roteínas
ROC	**R**eceiver **O**perating **C**haracteristic
SVS	**S**ección de **V**ías de **S**eñalización
VS	e**S**quema de **V**otación

Capítulo 1

Introducción

Las células, unidades morfológicas y funcionales básicas de todo organismo vivo, para garantizar su supervivencia deben interpretar y responder a muy variados estímulos físicos y químicos, tanto externos como internos [1]. Para ello necesitan mantener un alto grado de organización que les permita llevar a cabo sus funciones vitales básicas, como nutrirse, crecer, multiplicarse, diferenciarse, registrar y transportar señales así como controlar una enorme cantidad de reacciones bioquímicas que tienen lugar en su interior. La capacidad de coordinación y control que tiene la célula sobre estos procesos y funciones, hacen de ella un sistema extremadamente complejo para su estudio que ha despertado interés en distintas disciplinas científicas, entre ellas la física.

Los múltiples fenotipos y propiedades colectivas emergentes que una célula presenta no pueden ser reducidos a meras propiedades individuales de sus componentes constituyentes. Es decir, un reduccionismo extremo no serviría para entender cabalmente el funcionamiento de la célula como sistema. En efecto sus características y funcionalidades son consecuencia de intrincados mecanismos de interacción y acoplamiento entre material genético, proteínas, enzimas, metabolitos, así como diversos mecanismos de de sensado, señalización y control. Por otro lado, la consideración simultánea de todas las posibles variables y sus combinaciones no resulta factible, por lo que es usual encontrar enfoques con distintos grados de simplificación y aproximación al problema.

Un interesante compromiso entre el reduccionismo extremo y la consideración completa de todas las posibles variables, surge al describir los sistemas biológicos con un enfoque de redes complejas. [2]. Este enfoque, ha tenido un fuerte auge en el campo de la física desde las publicaciones de Watts y Strogatz sobre redes de mundo pequeño[3] y la de Barabasi y Albert sobre el estudio de redes libres de escala[4]. En este tipo de aproximación, usualmente se

consideran componentes moleculares dentro de una célula como nodos, y las posibles interacciones (físicas, químicas, directas o indirectas) como aristas o conexiones entre ellos. Más aún, los enormes avances en tecnologías experimentales de las últimas décadas han permitido desarrollar técnicas que recopilan datos experimentales en forma masiva y facilitan la construcción de este tipo de redes a escalas antes impensadas. Tecnologías actuales como secuenciación de nueva generación, sistemas de doble híbrido en levaduras de alta eficiencia, o cromatografía de inmunoafinidad, ofrecen resultados experimentales a gran escala que permiten analizar genomas o proteomas completos con un mínimo esfuerzo. Como consecuencia directa las redes biomoleculares actuales han alcanzado escalas de organismos completos, donde los enfoques estadísticos, computacionales y desde la física resultan sumamente enriquecedores para abordar preguntas subyacentes al área de la biología celular.

Existen numerosos tipos de redes intracelulares que pueden construirse y utilizarse para analizar e interpretar procesos biológicos. Algunos paradigmáticos ejemplos los constituyen las redes metabólicas, que representan cadenas de reacciones bioquímicas, las redes de señalización celular que capturan respuestas celulares a cambios en su medio, las redes regulatorias génicas que capturan información sobre mecanismos de transcripción, o las redes gen-enfermedad que recapitulan información de información sobre mutaciones genéticas asociadas a distintas patologías complejas. Otros dos tipos de redes de especial interés en el contexto de esta tesis son las redes de interacción de proteínas, que capturan interacciones físicas entre pares de proteínas en muy diversos contextos celulares, y las redes de interacción fármaco-proteína que recapitulan información de compuestos químicos y sus respectivos blancos de acción.

Es importante tener en cuenta que casi la totalidad de las funciones biológicas que tienen lugar en una célula son llevadas a cabo por proteínas. Estas biomoléculas cumplen funciones extremadamente variadas, algunas de carácter estructural, otras de transporte de señales, de ensamblado de complejos proteicos, funciones inmunológicas, de catalización y control de reacciones químicas, por mencionar algunos pocos ejemplos. Las proteínas interactúan físicamente entre ellas y la forma en que lo hacen está relacionada a las funciones y procesos celulares específicos. Por ello, el estudio de redes de interacción de proteínas (PIN) resulta de gran importancia para el estudio de numerosos procesos celulares. Este tipo de redes presentan propiedades colectivas emergentes, como la presencia de estructura altamente modular, que las distinguen ampliamente de redes aleatorias. En efecto, las descripciones modulares de estas redes han recibido mucha atención en los últimos años [5]. En particular, se ha puesto mucho esfuerzo en establecer y aprovechar correlaciones significativas entre grupos topológicos de la red, formados a partir de nodos densamente conectados, y la idea original de Hartwell sobre *módulos de funcionalidad*

biológica, definidos como un conjunto de componentes moleculares y sus interacciones que llevan a cabo una función biológica específica [6].

De particular interés resulta el trabajo original de Guimerá [7, 8], que hace uso explícito de la estructura modular en distintas redes biomoleculares para definir dos observables topológicos de mediana escala y con ellos establecer una descripción de la red en roles cartográficos. Este tipo de enfoque, originalmente empleado en el contexto de redes metabólicas ha recibido un creciente interés, aplicándose también en numerosos estudios de PINs [9–11]. No obstante, la búsqueda de estructura modular conlleva el problema de la falta de una definición unívoca y libre de hipótesis para establecer y definir módulos en una red. De hecho la estructura modular observada depende tanto de la escala adoptada como de la definición implícita propia al algoritmo de agrupamiento empleado. Por estos motivos, en los últimos años han sido propuestos múltiples algoritmos de agrupamiento para estudiar estructura modular en redes complejas, y la vasta mayoría se basan en la optimización de funciones objetivo muy diferentes. Dos algoritmos muy aceptados y extensamente utilizados en diferentes contextos son, el algoritmo de Clauset-Newman-Moore (CNM)[12] que busca optimizar la *Modularidad* de la red (una figura de mérito ampliamente empleada para comparar la calidad de estructuras modulares) y el algoritmo Infomap [13], que busca optimizar la descripción de un paseo al azar de longitud infinita en una red mediante conceptos de teoría de información.

En este contexto, uno de los principales aportes de esta tesis que se discute en el capítulo 4, es observar como afecta la elección de uno u otro tipo de algoritmo de agrupamiento, a los patrones de conectividad que pueden detectarse y las conclusiones biológicas a las que se puede arribar en cada caso. En particular se hace especial énfasis en la necesidad evaluar la calidad de una partición no sólo por el valor de *modularidad*, sino a través de la comparación de las escalas de trabajo características, y el grado de homogeneidad biológica presente en las estructuras modulares halladas.

Otro tipo de redes biomoleculares cuya atención ha incrementado exponencialmente en los últimos años son las redes fármaco-proteína. Para entender el auge en la utilización de este tipo de redes, es preciso considerar el contexto socio-económico que subyace al desarrollo de fármacos en la actualidad. En promedio, la aprobación e inclusión de un nuevo fármaco al mercado toma un tiempo de entre 12 y 15 años (dependiendo del área terapéutica) y los costos para hacerlo ascienden al billón de dólares. Si además consideramos el hecho de que 1 de cada 24 drogas que entran en fase preclínica llega a ser aprobada, se hace evidente que el área requiere una elevada inversión de capital e involucra un alto grado de riesgo. De hecho, uno de los métodos más fructíferos y eficientes para posicionar un fármaco en mercado es comenzar

con un viejo fármaco ya existente, el cual haya pasado algunas de las fases de investigación que demandan mayor tiempo y dinero. Este recurso para la búsqueda de dianas terapéuticas se conoce como reposicionamiento (o reutilización) de fármacos. Es así como surge el creciente interés por la industria farmacéutica para incursionar en nuevas metodologías para la búsqueda de blancos y fármacos, de manera de incrementar la tasa de éxito en las distintas fases clínicas y/o reducir los costos en ellas.

Más crítico aún es el caso de las enfermedades tropicales desatendidas (NTD) que afectan principalmente a personas en condiciones de pobreza de países en desarrollo. El limitado interés comercial en el desarrollo y mejoras terapéuticas subyace principalmente en los altos costos de inversión y el bajo retorno esperado al tratarse de pacientes de bajos recursos [14]. En el área de NTD, el reposicionamiento de fármacos existentes juega un rol fundamental. En particular, mediante esfuerzos de investigación y desarrollo que provienen del área académica y de organismos gubernamentales. Aquí la estrategia consiste en hacer uso de fármacos diseñados y probados en organismos modelo y probar su eficacia en el tratamiento de NTD. Un claro ejemplo de éxito lo constituye la *eflornitina* que fue desarrollada como un compuesto contra el cáncer y se está utilizando para tratar la tripanosomiasis africana (enfermedad del sueño).

El uso de redes biomoleculares como las redes fármaco-proteína, las redes de similitud química entre fármacos o incluso las redes de interacción de proteína han sido de gran utilidad en el área. En particular estas redes han mostrado eficacia para proponer nuevos blancos de drogas que no parecieran evidentes para guiar el reposicionamiento de fármacos existentes o predecir posibles efectos secundarios de ellos, y proveer en general, distintas herramientas de síntesis conceptual y análisis integrativos para el diseño de nuevos fármacos [2]. Desafortunadamente, éstas y otras estrategias de redes moleculares e integración de datos quimiogenómicos se han centrado en su vasta mayoría en atacar el problema desde la perspectiva de enfermedades que afectan al mundo desarrollado [15–17]. En este sentido, un aporte fundamental de esta tesis es enfocar los problemas de reposicionamiento de fármacos y búsqueda de blancos desde la perspectiva de NTD y mediante el enfoque de redes biomoleculares. En el capítulo 5 se presenta la integración de una vasta cantidad y variedad de información referente a fármacos y sus blancos de acción a lo largo de más de 220 especies, incluyendo tanto organismos modelo ampliamente estudiados, como organismos patógenos causantes de NTD. La integración de datos realizada contempla información de muy variada naturaleza, como similitud estructural entre compuestos químicos, relaciones de bioactividad entre fármacos y proteínas, o incluso relaciones funcionales y estructurales ente proteínas. En referencia a este hecho, otro aporte original de la tesis es plantear el problema desde la perspectiva de redes multicapa, una generalización del formalismo

tradicional de redes que ha recibido creciente atención en los últimos años, especialmente en el campo de la física.

La tesis está organizada como sigue. En el capítulo 2 se presentan las bases de notación y conceptos básicos sobre los distintos tipos de redes que se utilizarán en los capítulos subsiguientes. En el capítulo 3 se presenta la red de interacción de proteínas en humanos utilizada (HIPPIE), el tipo de experimentos que involucra, el procesamiento y filtrado preliminar de datos realizado así como una caracterización topológica básica y búsqueda de evidencias preliminares de organización modular mediante la comparación con redes aleatorias.

En el capítulo 4 se comparan las características de dos particiones modulares sobre la PIN presentada, haciendo especial énfasis en las escalas presentes en cada caso y analizando las consecuencias en las respectivas descripciones cartográficas, en la homogeneidad biológica de las estructuras halladas y en los patrones de conectividad intra e inter modular detectados. Además se presenta en este capítulo el estudio de genes asociados a procesos de envejecimiento celular y distintas vías de señalización de proteínas, analizando en cada caso la relación con las descripciones cartográficas halladas con ambas descripciones cartográficas.

El capítulo 5 presenta la construcción de una red multicapa que condensa información sobre aproximadamente un millón y medio de fármacos y sus blancos protéicos a través de 221 especies. Se detalla información estructural entre compuestos químicos, las relaciones de bioactividad con sus blancos proteicos, y se incluyen distintos criterios de similitud entre proteínas que involucran la presencia de dominios estructurales, vías metabólicas y grupos de homología. Luego se hace énfasis en dos temáticas de gran interés biológico: la priorización de potenciales blancos de droga en una especie patógena dada, y la búsqueda de blancos para drogas *huérfanas* en organismos patógenos, es decir compuestos con actividad probada pero cuyo mecanismo de acción y blanco específico permanece desconocido. Finalmente en el capítulo 6 se resumen las conclusiones finales.

Capítulo 2

Fundamentos.

2.1. Introducción

En este capítulo se introducen y discuten conceptos fundamentales de redes complejas que serán utilizados a lo largo de toda la tesis. En primer lugar se definen observables topológicos clásicos en grafos que serán de utilidad para caracterizar estructuralmente las redes bajo estudio. Se introduce el problema de reconocimiento de estructura modular en redes complejas, detallando y discutiendo dos métodos paradigmáticos ampliamente extendidos en la comunidad de redes y de principal interés en esta tesis. Asimismo, se definen observables topológicos de la red y se presentan modelos de redes descorrelacionadas. Luego se describen dos técnicas predictivas en redes complejas y los criterios usuales para evaluar el desempeño de las mismas en el contexto de problemas de clasificación. Finalmente se extiende el concepto clásico de redes a otros tipos de grafos, tales como redes multipartitas y redes multicapas. En ambos casos se motiva la utilidad de las mismas y se presenta la notación formal necesaria para trabajar con estos tipos de redes.

2.2. Definiciones generales

Una red o grafo es una representación de un conjunto de interacciones binarias entre pares de objetos. Formalmente, un grafo $G(\mathcal{N}, \mathcal{E})$ consta de un conjunto de entidades $\mathcal{N} \neq \emptyset$, y un conjunto de interacciones $\mathcal{E}$ conformado por pares de elementos de $\mathcal{N}$. Los elementos de $\mathcal{N} := \{n_1, n_2, ..., n_N\}$ se denominan nodos o vértices del grafo y los elementos de $\mathcal{E} = \{\overline{e}_1, \overline{e}_2, ..., \overline{e}_m\}$ se denominan aristas, arcos o conexiones del mismo. El número de elementos de los conjuntos $\mathcal{N}$ y $\mathcal{E}$ los denotaremos mediante N y m respectivamente, el primero determina el número de

objetos del grafo, usualmente referido como *tamaño del grafo* o *masa*, y el segundo la cantidad de aristas o conexiones del mismo.

En lo siguiente por siplicidad, notaremos al i-ésimo nodo n_i mediante la letra i. Cada arista será denotada con una dupla de dos números identificando los nodos que la misma conecta. Por ejemplo, si la k-ésima arista $\overline{e}_k \in \mathcal{E}$ conecta los nodos i y $j \in \mathcal{N}$ la denotaremos $\overline{e}_k := e_{ij} = (i, j)$. En tal caso, diremos que los nodos i y j son nodos adyacentes o primeros vecinos.

Una arista que conecta un nodo consigo mismo (e_{ii}) se denomina *bucle*. Por otro lado, si existe más de una arista $\overline{e}_l$, $\overline{e}_k$ con $l \neq k$ conectando los nodos i y j, se dice que esos nodos tienen *aristas multiples*. Ambos, bucles y aristas múltiples no están incluidos en la definición estandar de grafo. En tal caso, estaríamos en presencia de *multigrafos* que no serán objeto de estudio en esta tesis [18, 19].

Hay dos clases bien diferenciadas de redes en función de si las conexiones entre nodos tienen o no una dirección preferencial definida. Decimos que un grafo es *no dirigido* si las conexiones carecen de una dirección preferencial. Esto implica que los elementos $e_{ij} \in \mathcal{E}$ pueden ser descriptos por pares no ordenados (i, j) y por tanto resulta $e_{ij} = e_{ji}$, $\forall e_{ij} \in \mathcal{E}$. Por el contrario un grafo se dice *dirigido* si la naturaleza de las conexiones que representa tienen una dirección u orientación preferencial. En tal caso, las aristas del mismo quedan definidas mediante pares ordenados, de manera que $e_{i,j} = (i, j)$ implica la existencia de una conexión con sentido bien definido, que va del nodo i al nodo j. En general para grafos dirigidos se tiene que $e_{ij} = (i, j) \neq e_{ji} = (j, i)$.

2.2.1. Subgrafos, Conexidad y Componente Gigante.

Dado un grafo $G(\mathcal{N}, \mathcal{E})$ es posible considerar un subconjunto de nodos y aristas del mismo para definir otro nuevo grafo G'. Diremos que $G'(\mathcal{N}', \mathcal{E}')$ es un *subgrafo* de $G(\mathcal{N}, \mathcal{E})$, si se verifica que $\mathcal{N}' = \{n_1, n_2, ...n'_N\} \subseteq \mathcal{N}$ y $\mathcal{E}' = \{\overline{e}'_1, \overline{e}'_2, ..., \overline{e}'_m\} \subseteq \mathcal{E}$. Por otro lado, si $\mathcal{E}'$ contiene todas las posibles aristas de G que unen a nodos del conjunto $\mathcal{N}'$ se dice que G' es un *subgrafo inducido* o *subgrafo completo* y se denota simplemente por $G' = G(\mathcal{N}')$. Un concepto importante en teoría de grafos es el de conexidad entre pares de nodos y del grafo completo. Un *camino* del nodo i al nodo j del grafo es una secuencia de nodos y aristas adyacentes que conducen desde el nodo i al nodo j. La longitud de un camino está dada por el número de aristas del mismo. Si cada nodo del *camino* es visitado una única vez estamos en presencia de un *camino simple*. Si el camino empieza y termina en el mismo vértice se dice que es un *ciclo*, y si además

la longitud del ciclo es unitaria decimos que estamos en presencia de un *bucle*. Decimos que dos nodos i, j son *conexos* si existe un camino que los une, caso contrario diremos que los nodos i y j son *disconexos* o *inconexos*. Decimos que un grafo es *conexo* si todos sus pares de nodos son conexos. Una *componente* de un grafo G se define como un subgrafo inducido que cumple dos condiciones: ser conexo y contener la máxima cantidad posible de aristas comunes con G. Una *componente gigante* de un grafo $G(\mathcal{N}, \mathcal{E})$, es una componente cuyo tamaño (cantidad de nodos) es del mismo orden que N [20].

2.2.2. Grafos Pesados

Existen numerosos sistemas donde es posible asociar a cada interacción una magnitud o intensidad dada. Tales sistemas pueden ser descriptos más allá de un conjunto de interacciones binarias. Por ejemplo, pensemos en una red de coautorías, donde cada nodo representa un investigador y una arista entre dos investigadores denota si comparten alguna publicación en común. En este caso, resulta natural pensar en una medida de intensidad de las interacciones, la cual podría definirse proporcional al número de publicaciones que ambos autores comparten. Las redes que representan estos de sistemas resultan de especial interés para esta tesis y se denominan *grafos pesados*. Un grafo pesado puede ser dirigido o no dirigido. En general, un grafo pesado $G^W = G(\mathcal{N}, \mathcal{E}, \mathcal{W})$ consta de un conjunto de N nodos $\mathcal{N} = \{n_1, n_2, ...n_N\}$, un conjunto de m aristas $\mathcal{E} = \{\overline{e}_1, \overline{e}_2, ..., \overline{e}_m\}$ (cuyos elementos serán pares ordenados si el grafo es dirigido) y un conjunto de m pesos $\mathcal{W} : \mathcal{E} \to \Re, \quad \mathcal{W} = \{w_1, w_2, ...w_m\}$, cada uno de ellos asociado a la correspondiente arista en el conjunto $\mathcal{E}$. Usualmente el conjunto $\mathcal{W}$ toma valores positivos, pero es importante destacar que tal condición no es estrictamente necesaria (ver ejemplos [21]).

2.2.3. Matriz de adyacencia y matriz de pesos

Una representación particularmente útil para grafos es mediante notación matricial. Dado un grafo no pesado $G(\mathcal{N}, \mathcal{E})$ de N nodos, su matriz de adyacencia $\mathcal{A} \in N$ x N es una matriz cuadrada de elementos binarios $\mathcal{A} = \{a_{ij}/i, j \in 1...N\}$, de forma que $a_{ij} = 1$ si existe la correspondiente arista $\overline{e}_{i,j}$ y 0 en caso contrario. La matriz de adyacencia $\mathcal{A}$ será simétrica o asimétrica, según se trate de grafos no dirigidos o dirigidos respectivamente. Los elementos diagonoales deben ser nulos en orden de satifacer la ausencia de bucles que requiere la definición de grafo dada.

Por otro lado, en caso de tratarse de grafos pesados $G^W = G(\mathcal{N}, \mathcal{E}, \mathcal{W})$, la correspondiente representación matricial suele referirse como matriz de pesos $\mathcal{W} \in N$ x N. En este caso, el

elemento w_{ij} de la misma es el peso w del arco que conecta el nodo i con el nodo j si el mismo existe, y en caso contrario $w_{ij} = 0$. Nuevamente, la matriz de pesos $\mathcal{W}$ será simétrica sólo si el grafo es no dirigido.

2.2.4. Principales Observables Topológicos

2.2.4.1. Distribución de grado, asortatividad y disasortatividad

Consideremos un grafo no pesado y no dirigido $G(\mathcal{N}, \mathcal{E})$ con matriz de adyacencia $\mathcal{A}$. Definimos el grado k_i de un nodo como la cantidad de primeros vecinos o nodos adyacentes que posee,

$$k_i = \sum_{j \in \texttt{Nei(i)}} a_{ij} \tag{2.1}$$

donde `Nei(i)` denota el conjunto de nodos del grafo que son primeros vecinos del nodo i, y a_{ij} es el elemento (i, j) a la matriz de adyacencia $\mathcal{A}$ del grafo. En caso de grafos pesados $G^W = G(\mathcal{N}, \mathcal{E}, \mathcal{W})$, con matriz de pesos dada por $\mathcal{W} \in N$ x N se puede incluir los pesos de las aristas correspondientes, extendiendo la definición 2.1 al observable usualmente denotado como *strength* s_i de un nodo:

$$s_i = \sum_{j \in \texttt{Nei(i)}} w_{ij} \tag{2.2}$$

con $w_{ij} \in \mathcal{W}$. Estas medidas permiten establecer la caracterización más simple posible para un grafo, que es su distribución de grado $P(K)$ o bien P_k. La misma se corresponde con la probabilidad de que un nodo i tomado al azar del grafo tenga grado $k_i = K$. La distribución de grado P_k permite caracterizar por completo las propiedades estadísticas en redes no correlacionadas [20], es decir aquellas donde las conexiones entre los nodos son definidas de forma aleatoria. En contraste, en muchos casos de redes que representan sistemas reales existen correlaciones de mayor orden, en el sentido que la probabilidad que un nodo de grado k tenga una arista apuntando a otro de grado k' (que denotaremos $P(k'|k)$) depende de k. En casos reales los efectos de tamaño finito introducen ruido en el estudio directo de la probabilidad condicional $P(k'|k)$. Por tanto es útil definir un observable relacionado a esta probabilidad de mayor utilidad práctica, el grado medio de primeros vecinos. Para un nodo i esta magnitud $k_{nn,i}$ se define como

$$k_{nn,i} = \frac{1}{k_i} \sum_{j \in \texttt{Nei(i)}} k_j \tag{2.3}$$

luego, promediando sobre todos los nodos de grado k de la red puede calcularse el grado medio de primeros vecinos para nodos de grado k, $k_{nn}(k)$ la que puede ser expresada en términos de la probabilidad condicional $P(k'|k)$ según

$$k_{nn}(k) = \frac{1}{n} \sum_{k'} k' P(k'|k) \tag{2.4}$$

expresión que bajo la ausencia de correlaciones de grado, no depende de k. En general la distribución global del observable $k_{nn}(k)$ a través de la red permite definir dos posibles comportamientos. Si la función $k_{nn}(k)$ es creciente con k el grafo se dice *asortativo* mientras que si ésta es decreciente con k el grafo se dice *disortativo*. En términos más tangibles, en grafos asortativos los nodos tienden a estar conectados a otros de grado similar, mientras que en grafos disortativos los nodos de bajo grado tienden a conectarse a nodos de alto grado y viceversa.

Los conceptos de asortatividad y disasortatividad pueden ser generalizados para grafos pesados. Para ello, primero es necesario extender e interpretar la definición del grado medio pesado de primeros vecinos en un nodo i

$$k^w_{nn,i} = \sum_{j \in \texttt{Nei(i)}} \frac{w_{ij}}{s_i} k_j = \frac{1}{s_i} \sum_{j \in \texttt{Nei(i)}} w_{ij} k_j \tag{2.5}$$

donde s_i es el *strength* del nodo, `Nei(i)` su entorno de primeros vecinos. En el caso en que los arcos más pesados de las conectividades de un nodo fueran las que lo conectaran con sus vecinos de mayor grado se verificaría $k^w_{nn,i} > k_{nn,i}$, y en caso de que éstos apunten a sus vecinos menos conectados se observaría una relación de orden inversa. Por lo tanto, el grado medio pesado de primeros vecinos mide la afinidad de un nodo para conectarse con otros de alto o bajo grado en función a la distribución de pesos de sus aristas. Ahora podemos extender la noción de asortatividad y disasortatividad a grafos pesados según sea creciente o decreciente la relación $k^w_{nn}(k)$, es decir el valor medio de k^w_{nn} sobre todos los nodos de grado k.

2.2.5. Caminos cortos en un grafo: Betweenness

Tal como se definió en la sección previa, un camino entre dos nodos i y j es una sucesión de nodos y aristas que unen ambos nodos. En grafos pesados, la longitud del mismo viene dado por la cantidad de aristas que posee. Así definimos un *camino corto* entre dos nodos i, j, como aquel (o aquellos, si hubiera más de uno) de menor longitud entre los mismos. Notar que entre

dos nodos, puede haber más de un camino corto. La distancia geodésica entre dos nodos i y j se define como la longitud de los caminos cortos que unen estos nodos. Si el grafo es disconexo y los nodos no se conectan por medio de ningún camnino decimos que la distancia geodésica diverge. Es posible construir una matriz de distancias d_{ij} del grafo calculando la distancia geodésica entre todos los pares de nodos. La máxima distancia geodésica observada en la matriz d_{ij} se denomina *diámetro* del grafo, mientras que la distancia media se denomina *longitud de camino característica L*. En caso de grafos disconexos la longitud de camino característica puede calcularse en la componente gigante del grafo para evitar que diverja. Los caminos cortos pueden ser utilizados también como medida de importancia de un nodo en el grafo. Para cada nodo i del grafo podemos contabilizar cuantos caminos cortos pasan a través del mismo y definir así una medida de centralidad o importancia relativa entre nodos llamada *betweenness*. Para un nodo i, el betweenness $bet(i)$ queda definido según

$$bet(i) = \sum_{j,k \in \mathcal{N}, j \neq k} \frac{n_{jk}(i)}{n_{jk}} \tag{2.6}$$

donde $\frac{n_{jk}(i)}{n_{jk}}$ es la fracción de caminos cortos entre los nodos j y k que pasan por i. La condición $j \neq k$ significa que los ciclos no son considerados en el cálculo de betweenness. Notar que por aquellos nodos posicionados en zonas *centrales* del grafo, pasarán mayor cantidad de caminos cortos que en relación a aquellos nodos que ocupen posiciones más periféricas. Por tanto el betweenness resulta una medida topológica de centralidad muy utilizada que proporciona una medida de ”flujo”que pasa a través de un dado nodo. Notar que a diferencia del grado que sólo depende del entorno del nodo bajo estudio, el betweenness es una medida de carácter global y depende de la estructura entera del grafo.

El concepto de camino corto y betweenness de un nodo puede ser extendido a grafos pesados. En un grafo pesado la arista que conecta al nodo i con el nodo j, tiene un peso w_{ij} asociado. Como consecuencia el camino con mínima cantidad de aristas no es necesariamente el camino más corto. Dada la asignación de pesos w_{ij} para cada arco, la distancia entre nodos dependerá de la transformación utilizada. Dado que la distancia entre nodos depende de la transformación utilizada, el concepto de camino corto en un grafo pesado también está sujeto a esta elección. Por ejemplo uno puede medir la distancia $d_{ij} = \frac{1}{w_{ij}}$, aunque tal medida de distancia no respete la desigualdad triangular [22–24]. Bajo esta elección, el cálculo de caminos cortos puede interpretarse bajo una analogía eléctrica, donde w_{ij} representa la conductancia del arco y d_{ij} su resistencia. En este caso, dado que los arcos de un camino están conectados en serie, el cálculo de caminos cortos se reduce al de menor resistividad. Una vez definida la transformación

$d_{ij}(w_{ij})$, la definición de $bet(i)$ se extiende trivialmente utilizando los caminos cortos hallados.

2.2.6. Coeficiente de agrupamiento

Una medida topológica fundamental surge de cuantificar el grado en que el entorno de un nodo se encuentra conectado. La medida más básica para este fin es el coeficiente de agrupamiento local c_i de un dado nodo. Esta medida compara el número de conexiones presentes entre primeros vecinos del nodo i con el número máximo de conexiones que podrían existir entre estos. Así el coeficiente de agrupamiento local queda definido según

$$c_i = \frac{2}{k_i(k_i-1)} \sum_{j,m \in \mathtt{Nei(i)}} a_{ij} a_{jm} a_{mi} \tag{2.7}$$

Equivalentemente, este observable topológico se puede interpretar como una relación entre el número de triángulos que conforma el nodo i, y el número total de posibles triángulos que podrían producirse entre éste y sus primeros vecinos ($\frac{k_i(k_i-1)}{2}$). Esta media da una idea intuitiva de cuan conectado está el entorno de un nodo. Por ejemplo, si el subgrafo inducido G_i presenta una estructura tipo estrella (los vecinos del nodo i no se conectan entre sí sino a travez de éste último) su coeficiente de agrupamiento local será nulo ($c_i = 0$). En oposición, si todos los vecinos del nodo se encuentran completamente conectados entre sí, estaremos en presencia de un subgrafo inducido denominado *clique*, en cuyo caso tendremos $c_i = 1$. El coeficiente de agrupamiento total del grafo queda definido como la media del c_i de todos los nodos del grafo, es decir $C = \frac{1}{N} \sum_{i \in \mathcal{N}} c_i$.

En el caso de redes pesadas, una generalización posible para el coeficientede agrupamiento de un nodo i, fue propuesta en [25]

$$c_i^w = \frac{1}{s_i(k_i-1)} \sum_{j,m} \frac{w_{ij} + w_{im}}{2} a_{ij} a_{jm} a_{mi} \tag{2.8}$$

es decir, que cada triángulo se contabiliza a menos de un factor de peso dado por el promedio de los arcos del mismo que incluyen al nodo i. Notar que esta definición se reduce al caso de redes no pesadas cuando los pesos son todos uniformes. Además, el factor de normalización $s_i(k_i - 1)$ asegura que $c_i^w \in [0, 1]$ dado que sólo se consideran los pesos que involucran al nodo i. De esta manera es posible definir el coeficiente de agrupamiento total en un grafo pesado C^w como el promedio de los c_i^w y compararlo con el coeficiente de agrupamiento total que no considera los pesos w. Si $C^w > C$ significa que los triángulos del grafo están típicamente formados

por arcos de alto peso. Casos contrario, si $C^w < C$ significa que los tripletes están típicamente formados por arcos de bajo peso.

2.3. Estructura modular y calidad de una partición

Muchas redes presentan un alto grado de inhomogeneidad en sus patrones de conectividad, reflejando la presencia de un nivel de orden y organización no trivial en la red [26]. En general, la distribución de aristas del grafo no es global ni localmente uniforme, por lo que es común encontrar zonas de la red con alta densidad de aristas conectando grupos diferenciados de nodos y una baja densidad de aristas entre esos grupos. Este tipo de estructura usualmente presente en redes que representan sistemas reales, se conoce como *estructura en comunidades* o *estructura modular*. Numerosos ejemplos de comunidades en distintos tipos de redes podrían ser mencionados. En el caso de grafos sociales resulta intuitivo pensar en estructuras modulares que representen grupos familiares, grupos de amigos, laborales, etc. En redes metabólicas o de interacción de proteínas como la que presentaremos en el siguiente capítulo estos grupos modulares pueden representar y/o correlacionar con grupos funcionales.

En general, dado un grafo $G(\mathcal{N},\mathcal{E})$, una *comunidad* o *módulo* puede pensarse como un subgrafo $G'(\mathcal{N}',\mathcal{E}')$ cuyos nodos están fuertemente conectados entre sí y débilmente conectados para con otros nodos del grafo.

Es importante destacar que sin embargo, no existe una definición formal de *comunidad* en grafos universalmente aceptada. Más aún, la definición de módulo usualmente depende del sistema y la aplicación específica que se tenga en mente [26]. Una *partición*, es una división de un grafo en estructuras modulares de manera que cada nodo del grafo pertenezca a un único módulo. En muchos problemas resulta de especial interés definir comunidades de manera tal que un dado nodo pueda pertenecer a más de una de ellas. Tal división en comunidades superpuestas se denomina usualmente *cobertura*. En esta tesis, sólo serán objeto de estudio divisiones de redes en *particiones* de módulos disjuntos. Existen numerosos algoritmos para detectar posibles *particiones* de un grafo. Cada algoritmo se basa usualmente en su propia definición de comunidad, por lo que es esperable que se obtengan *particiones* cualitativamente distintas dependiendo el algoritmo empleado. Para comparar el desempeño de distintos algoritmos y sus *particiones* resultantes, es necesario definir alguna función de calidad que permita cuantificar cuan buena es una *partición* dada. La función de calidad más popular es la *modularidad* Q de Newman y Girvan [27]. Esta medida está basada en la idea de que una red aleatoria no presenta estructura modular. Por tanto resulta factible medir la calidad de un dado módulo al comparar la densidad de arcos internos

que posee con las que se esperaría si el mismo fuera extraido de un grafo aleatorio carente de estructura modular. Es claro que tal definición depende de la elección del modelo nulo empleado, es decir del grafo carente de estructura considerado, que respeta alguna de las características estructurales del grafo bajo estudio. Dado un grafo $G(\mathcal{N}, \mathcal{E})$ la modularidad Q queda definida según

$$Q = \frac{1}{2m} \sum_{ij \in \mathcal{N}} \Big(A_{ij} - P_{ij}\Big)\delta(C_i, C_j) \tag{2.9}$$

aquí m representa el número de arcos del grafo G, A_{ij} es el elemento correspondiente de la matriz de adyacencia, y P_{ij} es la probabilidad que los nodos i y j estén conectados en el modelo nulo elegido. El módulo al cual pertenece el nodo i se denota mediante C_i y se tiene $\delta(C_i, C_j) = 1$ sólo si los nodos i, j pertenecen al mismo módulo (caso contrario $\delta(C_i, C_j) = 0$). El modelo nulo más usual para el cálculo de modularidad es el modelo configuracional [28, 29]. Este tipo de grafo aleatorio preserva el número total de aristas de cada nodo, y por consecuente preserva no sólo la cantidad total de aristas del grafo sino también la distribución de grado del mismo. En este modelo cada nodo puede conectarse a cualquier otro del grafo. Una forma de pensar la construcción de este modelo para un grafo no dirigido $G(\mathcal{N}, \mathcal{E})$ de N nodos y m aristas, es que inicialmente cada nodo tiene disponible media arista y para formar una arista del nodo i al j es necesario tomar una de las medias aristas de cada nodo. La probabilidad de unir dos nodos i, j es el producto de las probabilidades de tomar una media arista de uno y otro nodo, es decir $p_{ij} = p_i p_j = \frac{k_i}{2m}\frac{k_j}{2m}$. Por lo tanto el valor de expectación de esta probabilidad será $P_{ij} = 2m * p_{ij} = \frac{k_i k_j}{2m}$. Con ésto, las forma más usual de calcular la modularidad de una *partición* resulta

$$Q = \frac{1}{2m} \sum_{ij \in \mathcal{N}} \Big(A_{ij} - \frac{k_i k_j}{2m}\Big)\delta(C_i, C_j) \tag{2.10}$$

Esta medida de calidad de *partición* puede ser generalizada a grafos pesados e incluso dirigidos aunque este último caso no es objeto principal de estudio en esta tesis. Para grafos pesados basta con considerar el *strength* de cada nodo en lugar del grado, y modificar el factor de normalización. Siendo W la suma total de pesos del grafo bajo consideración tenemos

$$Q = \frac{1}{2W} \sum_{ij \in \mathcal{N}} \Big(w_{ij} - \frac{s_i s_j}{2W}\Big)\delta(C_i, C_j) \tag{2.11}$$

esta es la forma más general que adoptaremos en esta tesis para calcular la modularidad. En las siguientes secciones y capítulos esta medida será de gran utilidad para evaluar y comparar

el desempeño de distintos algoritmos de agrupamiento.

2.3.1. Algoritmos de agrupamiento considerados

Como se mencionó en la sección anterior, existen múltiples métodos para extraer estructura en comunidades de un grafo. En la presente tesis nos basaremos principalmente en dos metodologías ampliamente reconocidas y utilizadas que describiremos a continuación. En primer lugar se consideró el algoritmo de Clauset-Newman-Moore [27] el cual es miembro de una gran familia de algoritmos que, con distintas heurísticas, buscan *particiones* de una red optimizando directamente la función de calidad Q definida en la ecuación 2.11. Por otro lado se consideró el algoritmo Infomap [13] que hace uso de criterios de optimización completamente diferentes, basados en teoría de información. En este algoritmo los módulos quedan definidos de manera tal que se minimice la longitud media de la descripción de un proceso de paseo al azar que tiene lugar en el grafo. La idea principal es describir el paseo al azar con un sistema de etiquetas de dos niveles. Dada una *partición* P, un tipo de etiqueta es utilizada para describir las distintas comunidades de la *partición* y la otra clase de etiquetas se utiliza para identificar nodos dentro de esas comunidades. Para describir eficientemente un paseo al azar con este código de dos niveles es necesario que la *partición* refleje los patrones de flujo dentro de la red, de manera que los diferentes módulos se correspondan con zonas de alta densidad de conexiones donde el caminante del paseo al azar pase suficiente tiempo antes de pasar a recorrer otros módulos. Dada una *partición* P que consta de $P = \{P^1, P^2 ... P^s\}$ módulos, un paseo al azar de longitud infinita en la red puede ser descripto conceptualmente por dos contribuciones, una asociada a los saltos que ocurren entre distintos módulos ($P^i, P^j \quad i \neq j$) y otra asociada a los movimientos que ocurren dentro de cada uno de los módulos P^i. El algoritmo infomap cuantifica este hecho mediante la función de costo descripta según

$$L(P) = q_{inter} H(Q) + \sum_{i}^{s} p^i_{intra} H(P^i) \tag{2.12}$$

donde q_{inter} es la probabilidad de pasar de uno a otro módulo en la caminata, $H(Q)$ es la entropía de movimientos entre módulos, p^i_{intra} es la fracción de movimientos de la caminata que han ocurrido en el módulo P^i y $H(P^i)$ es la entropía de movimientos que ocurren dentro del módulo P^i. El primer término de la ecuación 2.12 da el número medio de bits necesarios para describir el movimiento entre módulos y el segundo término da el número medio de bits necesarios para describir el movimiento dentro de los distintos módulos [13]. Dar un detalle

exhaustivo del cálculo de esta función de costo escapa a los obejetivos de esta sección, pero el lector interesado puede consultar el material suplementario del trabajo original [13].

2.4. Redes aleatorias

El término redes aleatorias refiere a grafos cuya distribución de conexiones tiene naturaleza esencialmente desordenada. En general la construcción de este tipo de redes resulta útil para proveer distintos modelos nulos de referencia con los cuales se pueda comparar las propiedades topológicas de la red bajo estudio, G. En general estos modelos respetan distintas propiedades del grafo G. Los modelos más básicos permiten construir grafos aleatorios G^* que tengan la misma cantidad de nodos N y aristas m que el grafo G pero conectadas al azar (modelo de Erdös Renyi). Otros modelos más complejos generan grafos G^* respetando no sólo el número total de nodos y aristas, sino además pidiendo que cada nodo tenga la misma cantidad de conexiones (modelo configuracional ya presentado) respetando por lo tanto la distribución de grado de la red. Ambos modelos pueden complejizarse aún más exigiendo que los grafos resultantes estén libres de bucles y aristas duplicadas. Existen más modelos de redes aleatorias que buscan preservar otras propiedades globales del grafo como por ejemplo el coeficiente de agrupamiento [30], pero no serán considerados en esta tesis.

2.4.1. Erdös Renyi

Hay diferentes formas de construir una red aleatoria, siendo una de las más simples y conocidas el modelo de Erdös Renyi. En este modelo nulo ($G^{ER}_{N,K}$) se construye una red aleatoria a partir con N nodos, conectando pares elegidos al azar y omitiendo multiples conexiones entre dos mismos nodos. Este proceso se repite hasta alcanzar una cantidad total K de aristas impuesta a priori [31]. Para obtener una descripción completa de las características de $G^{ER}_{N,K}$ debería considerarse el ensamble de todas las realizaciones posibles, las que podrían ser descriptas por ejemplo con un ensamble de matrices de adyacencia de los grafos $G^{ER}_{N,K}$[32]. Otra alternativa para generar grafos ER es imponer el número total de nodos y la probabilidad de que dos nodos sean conectados $G^{ER}_{N,p}$. En este tipo de grafos la cantidad total de aristas K toma distintos valores a lo largo de todas las posibles realizaciones. Como conectar un dado par de nodos es un experimento de Bernoulli con probabilidad p, la probabilidad de que un grafo $G^{ER}_{N,p}$ con K aristas aparezca en el ensamble viene dada por $p^K(1-p)^{\frac{1}{2}N(N-1)-K}$ [31, 33, 34]. Pese a que este tipo de modelos no reproduce la mayor parte de las propiedades topológicas de las redes reales, es uno de los modelos nulos mejor estudiados [20]. En el caso de redes $G^{ER}_{N,p}$ es importante destacar

la existencia de una transición de fases de segundo orden alrededor de una probabilidad crítica $0 < p_c < 1$ donde se observa un cambio en las propiedades estructurales de la red. Para $p_c = \frac{1}{N}$ que se corresponde con un grado medio $< K >_c = 1$, Erdos y Rényi probaron que [31]:

- para $p < p_c$, con probabilidad tendiendo a 1 cuando N tiende a infinito, el grafo carece de componentes de tamaño mayor a $O(ln(N))$. Además, ninguna componente tiene más de un ciclo.
- para $p = p_c$, el grafo tiene una componente gigante de tamaño $O(N^{\frac{2}{3}})$
- para $p > p_c$, el grafo tiene una componente de tamaño $O(N)$, con un número $O(N)$ de ciclos. Además ninguna otra componente tiene tamaño mayor a $O(ln(N))$ ni más de un ciclo.

En un grafo aleatorio de ER todos los nodos son equivalentes. La probabilidad de escoger un nodo con grado k viene dada por una distribución binomial $P(k_i = k) = \binom{N-1}{k} p^k (1 - p)^{N-1-k}$[35] y por consiguiente, para valores grandes de N que preservan fijo $< k >= Np = cte$ la distribución de grado se puede aproximar correctamente por una distribución de Poison $P(k) = e^{-<k>} \frac{<k>^k}{k!}$. Por último mencionamos que el coeficiente de agrupamiento en un modelo de *ER* es exactamente $C = p = \frac{<k>}{N}$, dado que para un nodo de grado k el número de aristas entre sus primeros vecinos es $p\frac{k(k-1)}{2}$ y el máximo posible es $\frac{k(k-1)}{2}$ [3]. Muchas otras propiedades de este modelo han sido estudiadas pero escapan a los objetivos de esta sección. El lector interesado puede consultar bibliografía recomendada [3, 32, 33, 36].

2.4.2. Modelo configuracional

Una forma particularmente útil de generalizar las redes aleatorias de *ER* es pidiendo que las mismas tengan una distribución de grado arbitraria $P(k)$ [28, 29]. Tal generalización permite producir modelos nulos más realistas para el estudio y comparación de propiedades topológicas en redes reales. Denotaremos este tipo de redes mediante G_D^{conf}, donde D refiere a una dada distribución $P(k)$ que se desea a reproducir. Como se mencionó en la sección 2.3, para obtener un grafo de este tipo alcanza con inicializar cada nodo del mismo con la cantidad k_i de *semi-aristas* deseadas y escoger de a pares de éstas con igual probabilidad para establecer las conexiones finales del grafo. Es necesario por su puesto evitar conexiones múltiples y bucles.

2.5. Roles cartográficos

De particular interés para esta tesis resulta considerar dos observables topológicos adicionales definidos por Guimera[37] que tienen en cuenta explícitamente la estructura modular del grafo. Dado un grafo $G(\mathcal{N}, \mathcal{E})$ y una *partición* $S = \{C_1, C_2, ...C_{nc}\}$ del mismo, el *grado intramodular* $Z(i)$ de un nodo i perteneciente al módulo C_i es una medida del grado de conectividad del nodo i en relación a la distribución de grado propia de los nodos en C_i. El mismo queda definido según

$$Z(i) = \frac{k_i - < k_{C_i} >}{\sigma_{K_{C_i}}} \tag{2.13}$$

donde k_i es el grado del nodo i, $< k_{C_i} >$ es el grado medio de todos los nodos en el módulo C_i, y $\sigma_{K_{C_i}}$ es la desviación estándar del grado para nodos en C_i. Notar que esta medida nos permite detectar nodos cuyo nivel de conectividad puede bien no ser elevado desde un punto de vista global de la red, mas ser importante en un entorno local del nodo en cuestión. También cabe destacar que aquí el concepto de *local* no se limita estrictamente a los primeros vecinos del nodo en cuestión sino a todos los nodos en su misma comunidad. El segundo observable introducido por Guimera es el *coeficiente de participación* de un nodo. Este observable permite cuantificar la forma en que las conexiones de un nodo están distribuidas a través de los distintos módulos de la *partición*. El mismo queda definido según

$$P(i) = 1 - \sum_{c=1}^{nc} \left(\frac{k_{i_c}}{k_i}\right)^2 \tag{2.14}$$

donde k_{i_c} es la cantidad de aristas que el nodo i tiene en el módulo c. Este observable verifica $0 < P(i) \leq 1$ para todo $i \in \mathcal{N}$. Nodos que tengan todas sus aristas conectadas a nodos de su mismo módulo tendrán a un $P = 0$ mientras que nodos con escasa proporción de aristas en su mismo módulo tendrán un $P \sim 1$. También vale la pena destacar que el valor de P no sólo depende del numero de aristas dentro y fuera del módulo al cual el nodo pertenece sino de la forma en que las aristas que van hacia otros módulos están distribuidas.

Estos dos observables topológicos son sensibles a propiedades de la red de escala intermedia, es decir no son ni estrictamente locales como el grado de un nodo que sólo contempla los primeros vecinos del mismo, ni estrictamente globales como el betweenness para el cual se considera el flujo a través de todos los nodos de la red. Con estos observables Guimera definió un mapa cartográfico distinguiendo 7 categorías diferentes según el nivel de *participación*

P y *grado intramodular Z* que cada nodo posea. La Fig 2.1 ilustra las regiones correspondientes a estos roles en el plano Z-P. Según su *grado intramodular* los nodos se diferencian en dos grandes categorías según si tienen alto *grado intramodular* ($Z \geq 2.5$) o bajo *grado intramodular* ($Z < 2.5$). Este límite tiene sentido, si consideramos que Z compara el grado de un nodo con el grado de los restantes nodos de su comuna en términos de la desviación estándar que presenta la distribución de grado dentro de ese grupo. Simultáneamente, estas categorías se subdividen en otras según el nivel de participación de los nodos. Esta clasificación se describe en detalle en la tabla 2.1 y se ilustra en la figura 2.1

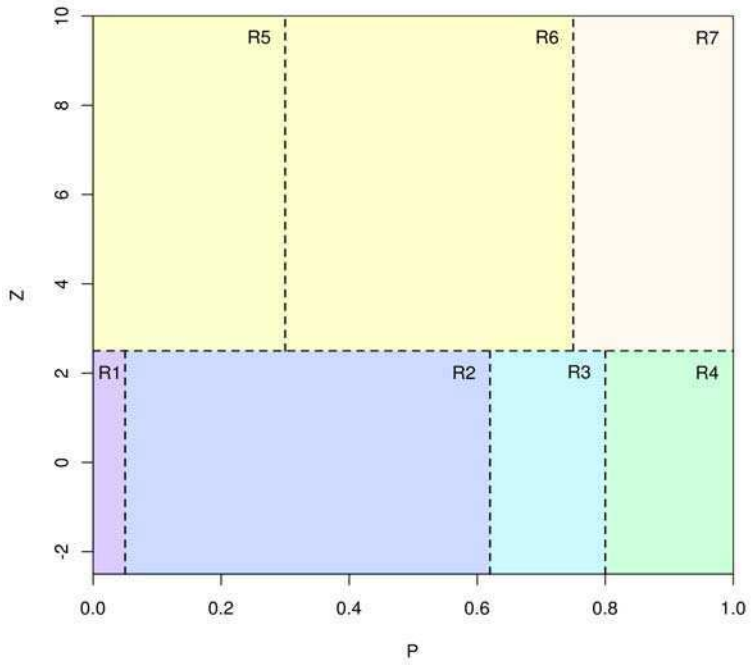


FIGURA 2.1: **Mapa Cartográfico de Roles Funcionales**: Plano cartográfico descripto por dos observables topológicos, la participación (abscisas) y el grado intramodular (ordenadas). En este plano es posible definir 7 roles cartográficos según la zona del plano que los nodos ocupen. Nodos con bajo *P* tienden a conectarse exclusivamente con nodos dentro de su comuna. Nodos con alto *P* tiendena conectarse con nodos de otras comunidades. Nodos con alto *Z* tienen un grado de conectividad *k* típicamente superior al grado medio de los miembros de su comunidad. Nodos de $Z \sim 0$ tienen un grado de conectividad que no difiere del grado medio de los nodos en su comuna.

Guimerá mostró además que estos límites son compatibles con la función de distribución de densidades en el plano $Z - P$ utilizando varias redes de sistemas reales, tales como redes metabólicas, de aeropuertos, de colaboración en publicaciones o internet.

Nodos de bajo grado intramodular ($Z < 2.5$)				
N.O.	Nm.	N.A.	*participación*	Descripción
ultra peripheral	R1	ultra periféricos	$P \sim 0$	Nodos con casi la totalidad de las conexiones dentro de su propia comuna.
peripheral	R2	periféricos	$P < 0.625$	Nodos con aproximadamente el %60 de las conexiones dentro de su propia comuna.
connectors	R3	conectores	$0.625 < P < 0.8$	Nodos con al menos la mitad de sus conexiones dentro de su propia comuna.
kinless	R4	kinless	$P > 0.80$	Nodos con menos del 35 % de las conexiones dentro de su propia comuna.
Nodos de Alto grado intramodular ($Z \geq 2.5$)				
N.O	Nm.	N.A	*Participación*	Descripción
provincial hubs	R5	conectores provinciales	$P < 0.3$	Nodos con al menos $\frac{5}{6}$ de las conexiones dentro de su propia comuna.
connector hubs	R6	conectores satélites	$0.3 < P < 0.75$	Nodos con al menos la mitad de las conexiones dentro de su propia comuna.
kinless hubs	R7	kinless hubs	$P > 0.75$	Menos de la mitad de las conexiones dentro de su propia comuna.

Cuadro 2.1: Definición de roles cartográfico según los niveles de *participación* y *grado intramodular* Abrebiaturas: N.O: nombre original (Guimera [37]),Nm: nomenclatura, N.A: nombre alternativo.

2.6. Métodos predictivos

2.6.1. Algoritmos de Priorizacion en Redes complejas

Los algoritmos de priorización en redes complejas son esencialmente modelos predictivos. Supongamos que tenemos un grafo $G = G(\mathcal{N}, \mathcal{E})$, donde cada nodo del conjunto $\mathcal{N} = \{n_1, n_2, ...n_N\}$ representa un individuo y las aristas $e_{ij} \in \mathcal{E}$ representan relaciones entre pares de éstos, ya sea de tipo laboral, familiar, de amistad, etc. Supongamos además que tenemos información concreta que un subconjunto $N_a = \{n_1, n_2, ...n_k\}$ ha comprado un producto P, y ninguno de los restantes indiviudos $N_b = \{n_{k+1}, n_{k+2}, ...n_j...n_N\}$ ha adquirido aún este producto. La idea básica de algoritmo de priorización en redes complejas es utilizar la información del conjunto N_a que denominaremos *semillas*, y la información embebida en los patrones de conectividad de la red para inferir nuevos potenciales compradores del producto P. El resultado típico de un algoritmo de priorización de esta naturaleza es una lista de nodos L ordenada según el grado de confianza otorgado a cada nodo $n_j \in N_b$ como potencial comprador de P.

Existe una amplia variedad de algoritmos de priorización en redes complejas cada uno basado en ideas y principios muy variados (para una revisión exhaustiva ver [38]). En esta tesis presentaremos y utilizaremos dos algoritmos de priorización diferentes. El primero y el más simple, es

una analogía de un esquema de votaciones VS donde cada nodo del conjunto N_a puede transmitir información a sus primeros vecinos. En el segundo algoritmo, basdo en una analogía de flujos, la información de cada nodo en el cojunto N_a puede propagarse a distancias geodésicas más grandes. A continuación se presenta una breve descripción de cada uno de ellos.

2.6.1.1. Esquema de Votación

Sea $G = G(\mathcal{N}, \mathcal{E}, \mathcal{W})$ un grafo pesado y $N_a = \{n_1, n_2, ...n_k\}$ un subconjunto de nodos del mismo que se sabe a partir de información externa a la red, asociados a una categoría o clase P. Para los restantes nodos del grafo $N_b = \{n_{k+1}, n_{k+2}, ...n_j...n_N\}$ se desea inferir aquellos con mayor potencial de pertenecer a la misma categoría P del conjunto N_a.
La estrategia más simple que se plantea en este trabajo está basada en un esquema de votación (VS), es decir, una suma pesada sobre primeros vecinos del conjunto de nodos utilizados como semillas N_a. Es un método simple, pero presenta una buena tasa de éxito, comparable a algoritmos de priorización más complejos [39, 40] con el beneficio extra de ser extremadamente eficiente. En un esquema VS, dado un grafo pesado $G = G(\mathcal{N}, \mathcal{E}, \mathcal{W})$ representado por su matriz de pesos M_P (w_{ij}), y el subconjunto de semillas $N_a = \{n_1, n_2, ...n_k\}$, el algoritmo VS prioriza los restantes nodos de la red $N_b = \{n_{k+1}, n_{k+2}, ...n_j...n_N\}$ a partir de la función de asignación de puntaje $f_P(n_j)$

$$f_P(n_j) = \sum_{l \in \texttt{Nei}(n_j), l \in N_a} w_{jl} \qquad \forall n_j \in N_b \tag{2.15}$$

donde la suma recorre sólo proteínas que estén simultáneamente en el conjunto de primeros vecinos de n_j, es decir $\texttt{Nei}(n_j)$, y el conjunto de semillas utilizado N_a. Los pesos de la suma w_{jl} son los pesos de las conexiones en la matriz M_P. Como resultado se obtiene una lista ordenada L donde los nodos que obtengan mayor puntaje (*score*) en la ecuación 2.15 serán inferidos como potenciales candidatos a pertenecer a la categoría P. Notar que todo nodo $n_j \in N_b$ que no sea vecino directo de algún nodo en el conjuto de semillas N_a obtendrá en la ecuación 2.15 un puntaje nulo.

2.6.1.2. Flujo Funcional

Otra estrategia de priorización considerada se basa en una analogía de dispersión de flujos en redes [41]. En este algoritmo que llamaremos *flujo funcional* o *Functional Flow* (FF), cada nodo es considerado como reservorio de flujo que puede transmitir su *caudal* sólo a vecinos con menor nivel de flujo que éste. El algoritmo se itera actualizando a cada paso la cantidad de flujo

neta de todos los nodos de la red, en base a la cantidad de flujo entrante y saliente de los mismos. En particular, el algoritmo se inicializa considerando un flujo infinito para todo nodo $n_i \in N_a$, y cada uno de éstos transmite su fljo en forma proporcional a los pesos w_{ij} de cada una de sus conexiones. La cantidad de flujo neta de los nodos en N_a se considera infinita a lo largo de toda la simulación y a cada paso transmiten su flujo según el peso de sus conexiones. Por defecto, el algoritmo considera 5 iteraciones (ver trabajo original [41]).

Formalmente, para propagar el grado de asociación de una clase funcional P, puede definirse para todo nodo de la red, una función $R_t^P(n_i)$ que determina la cantidad de flujo neto que el nodo n_i tiene a tiempo t para la propagación de la clase funcional P. Se define además la variable $h_t^P(n_i, n_j)$ que representa el flujo asociado a la función P que el nodo n_i pasa al nodo n_j en el tiempo t. El algoritmo se itera d veces, inicializando el proceso según

$$R_{t=0}^P(n_i) = \begin{cases} \infty & \text{si } n_i \in N_a \\ 0 & \text{en otro caso} \end{cases} \tag{2.16}$$

y actualizando en cada paso temporal la función $R_t^P(n_i)$ según,

$$R_t^P(n_i) = R_{t-1}^P(n_i) + \sum_{j \in \texttt{Nei}(i)} [h_t^P(n_j, n_i) - h_t^P(n_i, n_j)] \tag{2.17}$$

de manera que la cantidad de flujo transferida es proporcional al peso de las conexiones w_{ij}

$$h_t^P(n_i, n_j) = \begin{cases} 0 & \text{si } R_{t-1}^P(n_i) \leq R_{t-1}^P(n_j) \\ min\{w_{ij}, \frac{w_{ij}}{\sum_{k \in \texttt{Nei}(i)} w_{ik}}\} & \text{en otro caso} \end{cases} \tag{2.18}$$

El puntaje final que se asigna a cada nodo en relación a la función P después de d iteraciones, es la cantidad total de flujo entrante que ha recibido cada nodo dada por

$$f_P(n_i) = \sum_{t=1}^{d} \sum_{j \in \texttt{Nei}(i)} h_t^P(n_j, n_i) \tag{2.19}$$

Notar que la información que contiene cualquier semilla del conjunto N_a no puede ser propagada a otros nodos de distancia geodésica mayor a d.

2.6.2. Métricas de desempeño: Curvas ROC

Como se mencionó anteriormente, los algoritmos de priorización son esencialmente modelos predictivos. Se cuenta con un grafo G y una clase funcional P que involucra al menos a un subconjunto N_a de nodos en G. El problema que nos confiere aquí es el de predecir cuáles de los restantes nodos de la red N_b pertenecen a P y cuáles no. Es decir, estamos en presencia de un problema de clasificación binaria (pertenecer o no, a la clase P). Los algoritmos presentados en secciones precedentes dan como resultado una lista de nodos $L = \{n_i, \quad / n_i \in N_b\}$ ordenada según una magnitud escalar $f_P(n_i)$ (ver ecuaciones 2.15 y 2.19). Es decir, los primeros nodos de la lista serán aquellos con mayor valor de f_P. Además, se espera que éstos se encuentren asociados a la clase funcional P con mayor nivel de confianza que nodos con menores valores de f_P. Esta lista puede responder al problema de clasificación planteado mediante la definición de un umbral $f_P(n_i) = u$, de manera que todo nodo en N_b será clasificado según verifique

$$L(n_i, u) = \begin{cases} n_i \in P & \text{si } f_P(n_i) \geq u \\ n_i \notin P & \text{si} f_P(n_i) < u \end{cases} \tag{2.20}$$

La función $L(n_i, u)$ representa un clasificador binario. Este clasificador depende del umbral de corte u elegido en la lista L que provee cada algoritmo de priorización. Para evaluar la capacidad predictiva de un clasificador de este tipo es necesario disponer entre los elementos de L algún subconjunto de referencia que se sepa a priori asociado a P. Teniendo este conjunto de referencia, es posible contabilizar la cantidad de aciertos y fallas que el clasificador $L(n_i, u)$ comete, las que permiten a su vez definir distintas métricas de desempeño.

En la práctica, para realizar la evaluación de un clasificador binario es usual dividir al conjunto N_a en dos subcojuntos N_a^T, N_a^E, de manera que se verifique $N_a = N_a^T \bigcup N_a^E, N_a^T \bigcap N_a^E = \emptyset$. El de mayor tamaño N_a^T (supongamos un 90 % de los nodos en N_a) se denomina conjunto de entrenamiento mientras que N_a^E (el 10 % restante) se conoce como conjunto de evaluación o referencia. Los nodos en N_a^T serán utilizados como semillas del algoritmo de priorización, mientras que los nodos en N_a^E se sumarán a la lista cuya clase se desea inferir y permitirán evaluar la capacidad predictiva del clasificador. Notar que ahora el resultado de un algoritmo de priorización es una lista L que contiene elementos de N_a^E y de N_b asignando a cada elemento un observable $f_P(n_i)$. Para un umbral de corte o discriminación u, podemos calcular la tasa de aciertos del clasificador $L(n_i, u)$ en la predicción de elementos del conjunto N_a^E, es decir la *fracción de verdaderos positivos* (TPR) o *sensibilidad* del predictor

$$TPR(u) = \frac{1}{|N_a^E|} \sum_{n_i \in L} \delta_i^{TP} \qquad \delta_i^{TP} = \begin{cases} 1 & \text{si} \quad f_P(n_i) > u \quad \wedge \quad n_i \in N_a^E \\ 0 & \text{en otro caso} \end{cases} \tag{2.21}$$

donde $|N_a^E|$ refiere al número de nodos en el conjunto de evaluación. Por otro lado, a tasa de fallos, es decir la *fracción de falsos positivos* (FPR) viene dada por

$$FPR(u) = \frac{1}{|L| - |N_a^E|} \sum_{n_j \in L} \delta_i^{FP} \qquad \delta_i^{FP} = \begin{cases} 1 & \text{si} \quad f_P(n_i) > u \quad \wedge \quad n_i \notin N_a^E \\ 0 & \text{en otro caso} \end{cases} \tag{2.22}$$

donde $|L| - |N_a^E|$ es el número de elementos de la lista L que no pertenecen al conjunto de evaluación N_a^E. Notar que ambas cantidades $FPR(u)$ y $TPR(u)$ están normalizadas en el intervalo $[0, 1]$. La tasa FPR puede ser también expresada alternativamente en términos de la *especificidad* del predictor mediante $FPR = 1 - especificidad$. Notar además que tanto FPR como TPR son funciones monótonas crecientes de u y se verifica que para $u_{min} = min(f_P(u))$ se tiene $TPR(u_{min}) = FPR(u_{min}) = 1$. Intuitivamente se espera que un "*buen clasificador*"pueda inferir nodos de la clase funcional P con una alta tasa verdaderos positivos (TPR) a expensas de una baja tasa de falsos positivos (FPR). Esto es, se espera que el predictor tenga simultáneamente alta especificidad y sensibilidad. La figura 2.2 consigna los observables $TPR(u)$ y $FPR(u)$ al variar el umbral de discriminación u desde el máximo al mínimo valor de f_P. Este tipo de gráfica se denomina curva ROC (Receiver Operating Characteristic) y son de gran utilidad para comparar el desempeño de distintos clasificadores binarios. En particular, el área bajo esta curva denotada mediante AUC (*Area Under Curve*) se utiliza como medida de desempeño del clasificador bajo estudio. Analicemos dos ejemplos extremos para ganar intuición sobre la interpretación de los valores de AUC. Por un lado, un *clasificador ideal* debería poder asignar para todo nodo en N_a^E un valor más alto de f_P que para cualquier otro nodo de la lista L, esto es

$$f_P(n_i) > f_P(n_j), \qquad \forall n_i \in N_a^E, n_j \in N_b \tag{2.23}$$

de lo que se deduce la existencia de un $u*$ que verifica ($TPR(u*) = 1$ y $FPR(u*) = 0$). Por lo tanto, la curva ROC asociada a un *clasificador ideal* encierra un área unitaria $AUC = 1$ (ver figura 2.2). En contraste, en un *clasificador aleatorio* la magnitud de $f_P(n_i)$ se encuentra completamente descorrelacionada con la clase a la que cada nodo n_i pertenece ($n_i \in P, \quad o \quad n_i \notin$

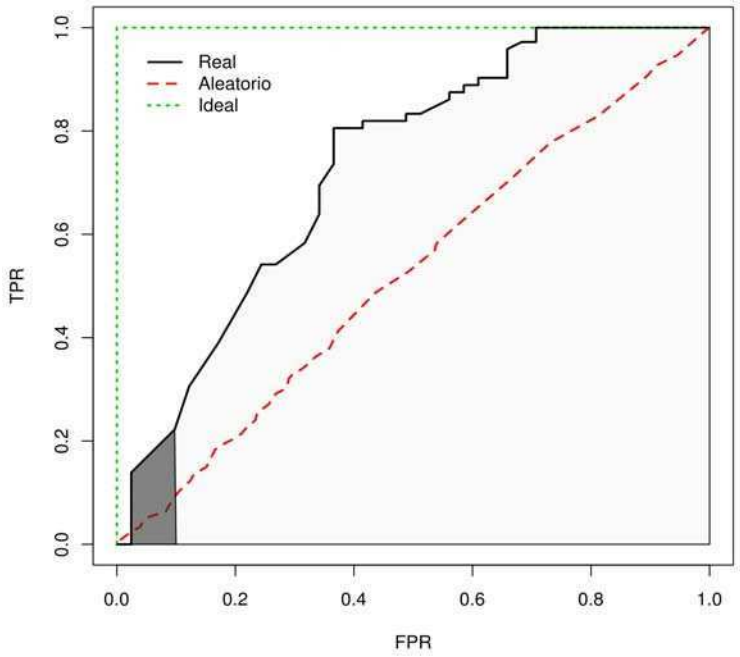


FIGURA 2.2: **Curvas ROC**: tasa de verdaderos positivos TPR de un predictor binario (o sensibilidad) en función de la tasa de falsos positivos FPR (1-especificidad del predictor). La curva de puntos verdes corresponde a un predictor ideal, cuya área encerrada es unitaria (AUC=1). La curva roja de trazos representa un predictor aleatorio cuya gráfica oscila sobre la recta identidad y presenta por tanto un AUC$\sim$ 0.5. La línea negra continua ilustra el caso de un predictor real. En el caso ilustrado, el área total $AUC = 0.73$ y se sombrea en gris tenue a modo ilustrativo. La superficie sombreada en gris oscuro corresponde al área bajo la curva limitada al 10 % de FPR, $AUC_{0.1} = 0.0137$. Bajo la corrección de McClish (ver ec.2.25), se tiene $AUC^c_{0.1} = 0.546$.

P). En tal caso, independientemente del umbral de discriminación u seleccionado se espera que la tasa aciertos y fallos en el clasificador sean del mismo orden. Por lo tanto, la curva ROC asociada a un clasificador aleatorio debe aproximarse a una recta de pendiente unitaria y el área asociada es $AUC \sim \frac{1}{2}$. En general un algoritmo predictivo obtiene valores de AUC en el rango $(\frac{1}{2}, 1)$. Dentro de este rango, a mayor AUC, mejor será el desempeño del algoritmo bajo estudio.

En la práctica sin embargo, no resulta demasiado útil comparar dos algoritmos en base a la totalidad de la lista L. En contraste resulta más apropiado comparar algoritmos considerando los elementos mejor puntuados en sus respectivas listas (es decir, con mayor valor de f_P). Con esta idea en mente se puede definir una medida muy útil para comparar algoritmos predictivos, el AUC-01, definida como el área bajo la curva ROC en el intervalo $FP \in [0, 0.1]$. Esto es, limitando el análisis, a lo que ocurra para una tasa de falsos positivos igual al 10 %. Si se tiene $|N_a^E| << |L|$, el AUC-01 equivale a considerar aproximadamente el 10 % de los nodos con mayor valor de f_P en la lista L. Notemos que en el caso de AUC-01 un predictor aleatorio presenta

un AUC-01=0.005, mientras que un predictor ideal presenta un área AUC-01=0.1. Resulta útil entonces considerar algún tipo de normalización del AUC-01 para llevarla al intervalo [0.5,1]. En esta tesis se utilizó para este fin la corrección de McClish [42] que se expresa según.

$$AUC^c_\alpha = \frac{1}{2}\left(1 + \frac{AUC_\alpha - AUC^{aleat}_\alpha}{AUC^{max}_\alpha - AUC^{aleat}_\alpha}\right) \tag{2.24}$$

$$AUC^c_{0.1} = \frac{1}{2}\left(1 + \frac{AUC_{01} - 0.005}{0.1 - 0.005}\right) \tag{2.25}$$

donde α es el valor máximo de FPR considerado, AUC^c_α el área renormalizada, AUC^{aleat}_α el área correspondiente a un predictor aleatorio, y AUC^{max}_α el área correspondiente a un predictor ideal. En la ec.2.25 se consideró el caso de interés en esta tesis, $\alpha = 0.1$.

2.7. Otros tipos de redes

Hasta aquí se han presentado definiciones elementales referidas a redes que se conocen usualmente bajo el nombre de redes monopartitas, es decir, redes con una única clase de nodos y arcos. Existen otros tipos de redes más complejas donde los nodos y arcos pueden clasificarse en categorías bien distinguidas, dando lugar a diferentes tipos de grafos. A continuación describiremos aquellos que resultan de interés en el contexto de esta tesis.

2.7.1. Redes Bipartitas

Una red $G(\mathcal{N}, \mathcal{E})$ se dice *bipartita* si existe una *partición* de $\mathcal{N}$, $(\mathcal{N}_1, \mathcal{N}_2)$ que verifica $\mathcal{N}_1 \cup \mathcal{N}_2 = \mathcal{N}$, $\mathcal{N}_1 \cap \mathcal{N}_2 = \phi$, y además ningún arco $\overline{e}_i \in \mathcal{E}$ une nodos de un mismo conjunto $\mathcal{N}_1$ ó $\mathcal{N}_2$. Hay muchos casos concretos de redes bipartitas. Por ejemplo, las redes de colaboración o coautorías tienen dos tipos de nodos bien distinguidos: autores y sus publicaciones. Un ejemplo biológico típico son las redes metabólicas [43] donde los nodos pueden clasificarse en compuestos químicos y reacciones químicas. Otro caso de particular interés para esta tesis son las redes de afiliación, donde se identifican objetos como una clase de nodos y características comunes a ellos como otra clase. Un ejemplo de red de afiliación sería considerar actores como objetos y las películas donde participaron como características. Otro caso posible que se abordará en el capítulo 5, consiste en tomar proteínas de diferentes especies como objetos y grupos funcionales o estructurales de las mismas como conjunto de características.

Una extensión natural del concepto de redes bipartitas es introducir la idea de redes *multipartitas*. En estas últimas existe una *partición* $\mathcal{N}_1, \mathcal{N}_2, ...\mathcal{N}_m$ que verifica las condiciones $\bigcup_i \mathcal{N}_i = \mathcal{N}$, y para cualquier par $i, j \in \{1...m\}$ se cumple $\mathcal{N}_i \bigcap \mathcal{N}_j = \phi$. Además, ningún par de nodos de una misma clase $\mathcal{N}_i$ se encuentra conectado. Un ejemplo de red multipartita de tres tipos de nodos (tripartita) son las llamadas, *folksonomías* término que refiere a métodos de indexación social [44, 45], donde usuarios, etiquetas y recursos *online* son los tres tipos de nodos de la red. Por ejemplo *flick.com* es un sitio web donde usuarios pueden asignar etiquetas a distintas fotografías, o bien *CiteUlike.com* es otro sitio donde usuarios pueden asignar etiquetas a referencias de publicaciones.

2.7.1.1. Proyección de Redes bipartitas en redes monopartitas

Un mecanismo usual para calcular la similitud estructural entre nodos de una misma clase en una red bipartita, es proyectando ésta en una red de tipo monopartita que contiene uno solo de los dos tipos de nodos. En este tipo de red proyectada, dos nodos comparten una conexión sólo si ambos están conectados al menos a un nodo común en la red bipartita original.

Sea un grafo bipartito con dos conjuntos de nodos $X = \{x_1, x_2, ...x_n\}, Y = \{y_1, y_2, ...y_m\}$ y sus conexiones dadas por el conjunto de aristas $\mathcal{E}$, de manera que $e_{ij} \in \mathcal{E}$ con $i \in X, j \in Y$. Denotaremos a este grafo $G^{bip}(\mathcal{N} = \{X, Y\}, \mathcal{E})$. El mismo puede ser representado por la matriz de adyacencia $A = (a_{ij})^{n \mathbf{x} m}$.

$$a_{ij} = \begin{cases} 1 & \text{si } e_{ij} \in \mathcal{E} \\ 0 & \text{en otro caso} \end{cases} \tag{2.26}$$

Recordemos que G^{bip} es bipartito por lo cual ningún elemento en $\mathcal{E}$ conecta dos nodos del conjunto X o dos nodos del conjunto Y. Una posible proyección bipartita del grafo G^{bip} sobre nodos X, fue definida por Zhou [46]. Esta proyección resulta en un grafo monopartito, pesado y dirigido $G_x(X = \{x_1, x_2, ...x_n\}, \mathcal{E}_x, \mathcal{W}_x)$ de nodos $X = \{x_1, x_2, ...x_n\}$, aristas $\mathcal{E}_x = e_{i,j}$ con $i, j \in \{1, 2...n\}$ que tienen pesos asociados $W_x = w_{i,j}$. Cada arista $e_{i,j} \in \mathcal{E}_x$ toma valores

$$e_{ij} = \begin{cases} 1 & \text{si } w_{i,j} \neq 0 \\ 0 & \text{si } w_{i,j} = 0 \end{cases} \tag{2.27}$$

y los pesos $w_{ij} \in \mathcal{W}_x$ se definen según

$$w_{ij} = \frac{1}{k_{x_j}} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k_{y_l}} \tag{2.28}$$

donde la suma corre sobre todos los nodos $y_l, l = \{1, 2, ...m\}$. Notemos que si los nodos x_i, x_j no tienen ningún vecino común y_l, entonces la suma 2.28 será nula y no habrá conexión entre estos nodos en G_x. La expresión 2.28 da los elementos de la matriz de pesos $\mathcal{W}$. Notemos que en general en este grafo proyectado $w_{ij} \neq w_{ji}$.

Esta matriz de pesos $\mathcal{W}$ puede ser obtenida matricialmente. Dado G^{bip} con matriz de adyacencia A, definimos la operación de normalización por columnas como la división de cada elemento a_{ij} por la suma de los elementos de la j-ésima columna y lo notaremos $\widehat{A}$. Es decir que para cada elemento $\widehat{a_{ij}} \in \widehat{A}$ tenemos

$$\widehat{a_{ij}} = \frac{a_{ij}}{\sum_j a_{ij}} \tag{2.29}$$

Notar que el denominador de la ec. 2.29 es el grado del j-ésimo nodo k_{y_j}. Entonces podemos reescribir la ecuación 2.28 como

$$w_{ij} = \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k_{y_l} k_{x_j}} = \sum_{l=1}^{m} \frac{a_{il}}{k_{y_l}} \frac{a_{jl}}{k_{x_j}} \tag{2.30}$$

$$= \sum_{l=1}^{m} \widehat{a_{il}} \frac{a_{jl}}{k_{x_j}} = \sum_{l=1}^{m} \widehat{a_{il}} \frac{(a_{lj})^t}{k_{x_j}} \tag{2.31}$$

$$w_{ij} = \sum_{l=1}^{m} \widehat{a_{il}} \widehat{(a_{lj})^t} \tag{2.32}$$

$$\mathcal{W} = \widehat{A} \quad \widehat{A^t} \tag{2.33}$$

donde el exponente $(a_{ij})^t$ en la ec. 2.31 indica la operación de trasposición. La expresión matricial de la ec. 2.33 da una relación directa entre la matriz de pesos del grafo monopartito proyectado como función de la matriz de adyacencia A de la red bipartita G^{bip}. Por otro lado, la expresión 2.33 puede ser trivialmente extendida a la expresión

$$\mathcal{W} = \widehat{A} \quad I \quad \widehat{A^t} \tag{2.34}$$

donde I representa la matriz identidad $I \in m\text{x}m$. La ec. 2.34 aporta una expresión alternativa a la ecuación 2.28 que será de particular utilidad en el capítulo 5 para extender el concepto de proyección bipartita propuesto por Zhou y colaboradores.

2.7.2. Redes Multicapa

En muchos sistemas reales, la utilización de redes como las que hemos hasta aquí definido puede resultar en una sobresimplificación del problema bajo estudio. En particular, el hecho de pensar que las interacciones entre objetos ocurren siempre a un mismo nivel de importancia resulta inapropiado para muchos casos e incluso puede conducir a conclusiones incorrectas en el estudio de la dinámica del sistema bajo estudio [47]. Una generalización de la teoría clásica de redes denominada *redes multicapa*, consiste en pensar a los sistemas compuestos por un conjunto de redes en distintos planos o capas de abstracción interconectadas entre sí, con aristas de distinta naturaleza y nivel de relevancia.

Consideremos a modo de ejemplo, el paradigma clásico e histórico de redes complejas: los sistemas sociales. Pensemos en una red social como *Facebook*, donde los nodos representen usuarios y las aristas representan conexiones entre éstos. Un usuario suele tener conexiones de muy diversa naturaleza, puede estar conectado a otros usuarios por relaciones laborales, familiares, de amistad, compañeros de determinadas actividades deportivas o culturales, etc. En este sentido puede resultar apropiado, pensar que vínculos de diferente índole o naturaleza estén situados en diferentes planos de abstracción, en lugar de ser tratados todos a un mismo nivel de jerarquía. Si uno quiere estudiar por ejemplo la propagación de un rumor en esta red social, es lógico que cada usuario no disemine el rumor de manera uniforme a lo largo de todos sus vínculos, sino que lo haga en principio con mayor probabilidad hacia usuarios potencialmente interesados en el rumor particular. Más aún, puede no esparcirlo a contactos de un determinado ámbito. Este ejemplo sería particularmente propicio para tratarse con redes multicapa, donde cada capa de la red puede contener un tipo específico de relaciones y los usuarios pueden estar simultáneamente en distintas capas, de manera que, la probabilidad de que un usuario disemine el rumor a sus vecinos depende de la capa o naturaleza de la conexión que tenga con estos vecinos.

Otro ejemplo que concierne más al eje temático de esta tesis, es el de redes de coexpresión génicas. Un enfoque clásico para éstas, es pensar el conjunto de genes de un dado organismo y trazar conexiones entre dos genes si existe algún tipo de correlación en el nivel de expresión de los mismos en un dado experimento. No obstante, estos experimentos pueden ser de muy variada índole, incluso puede tratase de experimentos realizados en diferentes tejidos, o bajo diferentes condiciones experimentales. Un hecho aceptado actualmente en literatura es que el tratamiento de este tipo de sistemas considerando todas las interacciones simultáneamente puede resultar en modelos ruidosos, ya que las interacciones pueden darse en contextos muy dispares.

Es usual llevar a cabo la construcción de estas redes limitando las interacciones a un tejido de interés, o a un conjunto dado de condiciones experimentales. Desde el punto de vista de redes multicapa, este tipo de sistemas es particularmente apropiado para pensar a cada tejido o condición experimental en una capa diferente, de manera que cada gen pueda pertenecer a más de una capa y de hecho tener diferentes entornos y niveles de conectividad en cada una de ellas.

Otro caso de particular interés para esta tesis que ampliaremos en el capítulo 5, es el de redes de proteínas y compuestos químicos empleadas para la búsqueda, *priorización* y reposicionamiento de fármacos. Estas redes pueden pensarse como capas compuestas por nodos de naturaleza diferente, tales como compuestos químicos, proteínas, procesos metabólicos, dominios funcionales propios de proteínas etc. Las conexiones entre nodos tienen también naturaleza muy diversa, pudiendo representar similitud estructural entre compuestos, evidencias de actividad de un dado compuesto sobre un blanco protéico de acción, pertenencia de dos proteínas a un mismo dominio funcional o una misma vía metabólica, etc. Este ejemplo será ampliado con mayor detalle en el capítulo 5, donde se llevará a cabo la construcción de una red de estas características.

2.7.2.1. Notación

Las redes multicapa (*multilayer networks*) son esencialmente una generalización de la tradicional teoría de redes. Una red multicapa puede pensarse como un conjunto de redes en diferentes niveles o capas relacionados entre sí. Formalmente, una red multicapa puede representarse mediante un par $\mathcal{M} = (\mathcal{G}, C)$, donde $\mathcal{G} = \{G_\alpha, \alpha \in \{1, 2, ...M\}\}$ es una familia de grafos $G_\alpha = G(\mathcal{X}_\alpha, \mathcal{E}_\alpha)$ (dirigidos, no dirigidos, pesados o no pesados) que denominaremos *capas* de $\mathcal{M}$, y $C = \{\mathcal{E}_{\alpha,\beta} \subseteq \mathcal{X}_\alpha \text{ x } \mathcal{X}_\beta; \alpha, \beta \in \{1, 2, ...M\}, \alpha \neq \beta\}$ es el conjunto de conexiones entre diferentes capas $G_\alpha, G_\beta, \alpha \neq \beta$. Los elementos de C se denominan *capas cruzadas o transversales* [47]. Cada capa G_α contiene el conjunto de nodos $\mathcal{X}_\alpha = \{x_1^\alpha, x_2^\alpha, ..., x_{N\alpha}^\alpha\}$ y sus conexiones se pueden representar con una matrz de adyacencia $A^{[\alpha]} = (a_{ij}^\alpha) \in \Re^{N_\alpha \mathbf{x} N_\alpha}$ cuyos elementos se definen

$$a_{ij}^\alpha = \begin{cases} 1 & \text{si } (x_i^\alpha, x_j^\alpha) \in \mathcal{E}_\alpha \\ 0 & \text{en otro caso} \end{cases} \tag{2.35}$$

Por otro lado, las capas transversales pueden representarse también por su matriz de adyacencia $A^{[\alpha,\beta]} = (a_{ij}^{\alpha,\beta}) \in \Re^{N_\alpha \mathbf{x} N_\beta}$, cuyos elementos se definen mediante

$$a_{ij}^{\alpha,\beta} = \begin{cases} 1 & \text{si } (x_i^\alpha, x_j^\beta) \in \mathcal{E}_{\alpha,\beta} \\ 0 & \text{en otro caso} \end{cases} \tag{2.36}$$

En suma, cada nodo $x_i^\alpha \in \mathcal{X}_\alpha$ *vive* una capa α y puede conectarse a través de dos clases de aristas, unas denominadas *conexiones intra-capa* $\mathcal{E}_\alpha$ que las unen a nodos de su misma capa, y otras denominadas *conexiones inter-capa* $\mathcal{E}_{\alpha,\beta}$ las cuales lo unen a nodos ubicados en otras capas.

Cabe destacar que muchas otras clases de redes pueden ser representadas como redes multicapa. Por ejemplo, las redes temporales (cuyos nodos y aristas dependen del tiempo) pueden mapearse a redes multicapa interpretando cada paso temporal como una nueva capa. Los hipergrafos son otra generalización de redes monopartitas, donde las conexiones no son sólo entre pares de nodos, sino que sus hiper-aristas conectan grupos de éstos. Un hipergrafo puede ser representado con un formalismo de redes multicapa reinterpretando cada hiper-arista como una capa del mismo e identificando las superposiciones entre hiper-aristas como conexiones en las capas transversales [47]. Otro ejemplo que resulta de particular interés para esta tesis son las redes multipartitas definidas en la sección anterior. Las mismas pueden ser representadas como un caso particular de redes multicapa, donde hay tantas capas $\alpha \in \{1, 2, ...M\}$ como clases de nodos de la red multipartita, y además se verifica que el conjunto de conexiones $E_\alpha = \emptyset, \forall \alpha \in \{1, 2, ...M\}$ dado que nodos de una misma clase (o capa en este caso) no están conectados entre sí. En esta representación, las aristas de la red multipartita se corresponden a las conexiones inter-capa $E_{\alpha,\beta}$, es decir a las capas transversales.

Capítulo 3

Redes de Interacción Proteína-Proteína.

3.1. Introducción

Las proteínas son macromoléculas formadas por cadenas lineales de aminoácidos, usualmente denominadas *cadenas polipeptídicas*. Las mismas se ensamblan a partir de la información contenida en los genes que las codifican. La secuencia específica de aminoácidos de una proteína se conoce como estructura primaria. Luego estas cadenas se pliegan en pequeños plegamientos locales y regulares (estructura secundaria) que pueden tomar formas de hélice alfa, hojas plegadas beta o giros beta que a menudo conectan estructuras alfa y beta. La distribución tridimencional final que adopta la cadena polipeptídicas se conoce como estructura terciaria. Esta estructura es esencial para determinar las propiedades biológicas y funcionales de la proteína, ya que condiciona su capacidad de interacción con otros grupos funcionales.

Es importante destacar que las proteínas llevan a cabo casi la totalidad de las funciones que tienen lugar dentro y fuera de la célula. Transmitir señales, catalizar reacciones químicas, cumplir roles estructurales, de transporte, de control de procesos o de regulación transcripcional son algunas de las innumerables funciones que tienen. Estas macromoléculas raramente llevan a cabo sus funciones de forma independiente, sino que en general interactúan físicamente entre sí y forman complejos proteicos que llevan a cabo las funciones específicas. El registro global de estas interacciones físicas entre proteínas conforma lo que se denomina **red de interacción de proteínas** o **PIN**. En particular, si la red contempla la totalidad de las proteínas en una especie dada, la PIN correspondiente suele referirse como **interactoma completo**. La consideración de PINs para el estudio de procesos biológicos tiene por sí misma una gran relevancia que subyace

principalmente en el activo rol que tienen las proteínas en la ejecución de procesos y funciones celulares. No obstante, resulta de suma importancia tener en cuenta el tipo de evidencias experimentales que conlleva a la construcción de una PIN. El conjunto de interacciones en una PIN tiene generalmente orígenes en una amplia variedad de técnicas experimentales y cada una de ellas presenta distintos niveles de precisión y tasas de error. La calidad de cualquier análisis y conclusiones biológicas que puedan extraerse de un análisis basado en PINs dependerá por su puesto de la calidad de la red de interacciones utilizada.

En este capítulo se presenta y caracteriza la red de interacciones correspondiente al organismo humano que será extensamente utilizada en el capítulo siguiente. Se detallan los criterios de filtrado utilizados, y las características topológicas de la red resultante. Se realiza también un análisis comparativo con distintos modelos nulos de redes aleatorias que dan evidencia preliminar de la presencia de patrones de conectividad no triviales y estructura modular en la red. Se analiza además en que medida la cantidad y calidad de ensayos experimentales que soportan la evidencia de interacciones de la red, correlacionan con la formación de patrones de conectividad no triviales.

3.2. Red de Interacción de Proteínas

3.2.1. Generalidades

A lo largo de esta tesis se consideró la red de interacción de proteínas generada por Schaefer y colaboradores, conocida como Human Integrated Protein-Protein Interaction rEference ,**HIPPIE** [48]. Esta red releva interacciones entre proteínas en el organismo humano. Las interacciones presentes en HIPPIE provienen de diferentes bases de datos de dominio público que se detallan en la tabla 3.1. Adicionalmente los autores incluyeron curaciones manuales extraídas de trabajos seleccionados [49–58]. Cabe destacar también que en esta red no se consideran formas de *splicing alternativo*, por lo que cada nodo representa indistintamente una proteína o su gen codificante. En total, HIPPIE reúne información de 11836 proteínas y 72917 interacciones. El 78 % de las proteínas reportadas en HIPPIE participan en al menos uno de tres experimentos predominantes que se detallan a continuación:

- **Inmunoprecipitación de complejos protéicos por anticuerpos** (Coprep): esta técnica captura a la proteína de interés con un anticuerpo específico y luego mediante un *Western blot* (una técnica de detección de proteínas específicas) se identifican las moléculas que

interaccionan con la proteína en estudio. Una de las principales ventajas de esta técnica es que la proteína de interés puede estudiarse en estado endógeno, es decir originada dentro de la célula o tejido específico.

- **Purificación por afinidad y espectrometría de masa** (MS). El método inicialmente purifica una proteína marcada y sus proteínas interactuantes. Luego las interacciones son cuantificadas mediante espectrometría de masas. El proceso de purificación de proteínas utilizado en este caso es mediante TAP (Tandem affinity purification), una técnica que captura la proteína *in vivo* mediante otra proteína de fusión, y luego la purifica mediante un doble proceso de lavado.

- **Sistema doble híbrido en levaduras** (Y2H): esta técnica estudia interacciones entre proteínas de fusión artificiales en el interior del núcleo celular de levaduras. La técnica se basa en que usualmente los factores de transcripción eucarióticos para promover la transcripción de un gen reportero necesitan dos dominios protéicos que deben estar próximos entre sí (uno de reconocimiento y otro de activación). La técnica Y2H toma un factor de transcripción y lo separa en dos fragmentos, cada uno conteniendo uno de estos dominios. Luego, cada fragmento se fusiona a una de las proteínas cuya interacción se desea analizar. Si las proteínas forman un complejo entre sí, los dos fragmentos del factor de transcripción se encontrarán próximos y se observará la transcripción del gen reportero correspondiente.

Notar que, si bien estos tres experimentos involucran al 78 % de las proteínas, los mismos abarcan sólo el 50 % del total de las interacciones. Es decir, la red HIPPIE contiene interacciones basadas en muchos otros tipos de experimentos adicionales. Un reporte completo de éstos se detalla en la tabla 3.2.

Cada experimento fue descripto con un vocabulario formal provisto por la ontología PSI-MI (Proteomics Standard Initiative - Molecular Interactions) [66] y más importante aún, la calidad de cada tipo de experimento fue cuantificada por Schaefer et. al. utilizando un rango de pesos en el intervalo $q \in [0, 10]$. Los valores de calidad adjudican máxima confianza ($q \sim 10$) a técnicas experimentales de alta fiabilidad como cristalografía de rayos X, valores de calidad media ($q \sim 5$) para ensayos de afinidad o de complementación de fragmentos de proteínas y los valores de calidad más bajos en casos de técnicas que no proveen evidencia directa de interacción como por ejemplo colocalización de proteínas (ver tabla 3.2).

Los autores de HIPPIE utilizaron estos valores de calidad q, para asignar un peso normalizado S a cada una de las interacciones. Este peso $S \in [0, 1]$ permite comparar la fiabilidad de

Base de datos	Tamaño	Referencia
HPRD	40110	[59]
BioGRID	30027	[60]
IntAct	28073	[61]
MINT	14094	[62]
Rual09	6946	[56]
Lim06	5579	[54]
Bell09	3300	[50]
Stelzl05	3232	[51]
DIP	1618	[63]
BIND	1415	[64]
Colland04	882	[57]
Lehner04	385	[53]
Albers05	290	[49]
MIPS	252	[65]
Venkatesan09	239	[58]
Kaltenbach07	227	[52]
Nakayama02	84	[55]
HIPPIE	72916	[48]

CUADRO 3.1: Bases de datos originales de interacciones de proteína utilizadas por HIPPIE.

las distintas interacciones. El mismo se calcula como la suma pesada de tres puntajes parciales: s_t que es función del tipo de experimentos que dan origen a la interacción, s_s que es función de la cantidad de experimentos y s_o que es función de la evidencia adicional existente en otros organismos (es decir el número de especies adicionales donde el par de proteínas ortólogas interactúan). Estos puntajes parciales se estiman con una forma funcional dada por

$$s_i(n) = \frac{2}{1 + e^{-a_i.n}} - 1 \qquad i = \{s, t, o\} \tag{3.1}$$

de manera que $s_i(0) = 0, \quad s_i(n \to \infty) = 1$. Para $i = t$, n representa la suma de los valores de calidad q de cada experimento donde se observa interacción entre las proteínas bajo estudio. Para $i = s$, n representa el número de experimentos donde se observa interacción y para $i = o$, n representa el número de especies donde proteínas ortólogas interactúan. Los parámtros a_i controlan la escala característica de cada transformación. Con estos puntajes parciales se calcula el peso total de cada interacción en HIPPIE mediante una suma pesada

$$S = w_s.s_s + w_t.s_t + w_o.s_o \qquad w_s + w_t + w_o = 1 \tag{3.2}$$

Los 6 parámetros libres $(w_s, w_t, w_o, a_s, a_t, a_o)$ fueron fijados en HIPPIE vía validación cruzada, donde el conjunto de evaluación fue generado sacando provecho de la existencia de

Ensayo	q	Ensayo	q
colocalization/visualisation technologies	1	phosphatase assay	7.5
comigration in gel electrophoresis	3	protease assay	7.5
comigration in non denaturing gel electrophoresis	3	protein array	5
comigration in sds page	3	protein complementation assay	5
competition binding	5	protein cross-linking with a bifunctional reagent	5
confocal microscopy	1	protein kinase assay	7.5
copurification	2	Protein-peptide	5
cosedimentation	2	protein tri hybrid	5
cosedimentation in solution	2	pull down	2.5
cosedimentation through density gradient	2	pull-down/mass spectrometry	5
cross-linking study	5	Reconstituted Complex	10
deacetylase assay	7.5	reverse phase chromatography	1
demethylase assay	7.5	reverse two hybrid	5
dihydrofolate reductase reconstruction	6	ribonuclease assay	7.5
dynamic light scattering	9	saturation binding	7.5
electron microscopy	5	scintillation proximity assay	7.5
electron tomography	9	solid phase assay	1
electrophoretic mobility shift assay	2	surface plasmon resonance	10
electrophoretic mobility supershift assay	2	t7 phage display	6
enzymatic study	1	tandem affinity purification	5
enzyme linked immunosorbent assay	5	transcriptional complementation assay	5
experimental interaction detection	1	Two-hybrid	5
far western blotting	5	two hybrid fragment pooling approach	5
filamentous phage display	6	ubiquitin reconstruction	5
filter binding	5	x-ray crystallography	10
fluorescence-activated cell sorting	1	x ray scattering	9
fluorescence correlation spectroscopy	10	yeast display	5

CUADRO 3.2: Ensayos experimentales que soportan la evidencia de interacción de proteínas en la red HIPPIE y sus respectivos valores de calidad ($q \in [0, 10]$) asigados orignialmente. Los nombres de los ensayos se corresponden con una denominación formal provista por la ontología PSI-MI [66].

interacciones con múltiple evidencia. Los mismos fueron fijados con valores $w_s = 0.6, w_t = 0.3, w_o = 0.1, a_s = 2.3, a_o = 1.6, a_t = 0.2$. Dar un detalle exhaustivo de los criterios y funciones objetivo de optimización de estas transformaciones no es objetivo de esta sección, el lector interesado puede consultar la referencia [48]. En la práctica resulta extremadamente útil contar con una forma sistemática para comparar la calidad de cada interacción en una PIN, dado que las mismas están sujetas a falsos positivos de las distintas técnicas experimentales que dan origen a estas interacciones. De hecho, Schaefer y colaboradores definen en su manuscrito original una PIN de *alta calidad* donde consideran sólo aquellas interacciones con un peso superior al tercer cuantil de la distribución $S \geq 0.73$. En esta tesis, hemos adoptado como red de interacción de proteínas el uso de esta red de *alta calidad* de interacciones, que puede pensarse como un subgrafo de la red HIPPIE completa. En adelante nos referiremos a éste como $G'_{HC}(\mathcal{N}', \mathcal{E}', \mathcal{W}' = S)$ (o

abreviando G'_{HC}), done $\mathcal{N}'$ representa el conjunto de proteínas, $\mathcal{E}'$ el conjunto de interacciones y $\mathcal{W}' = S$ el vector de pesos asociado a cada interacción dado por la ecuación 3.2. El conjunto $\mathcal{N}' = \{n_1, n_2, ..., n_N\}$ consta de un total de N=8277 proteínas, el conjunto $\mathcal{E}' = \{e_1, e_2, ..., e_M\}$ consta de $M' = 32321$ aristas cada una con un peso $S \geq 0.73$ asociado.

Este grafo tiene bucles, aristas duplicadas y además no es conexo. El mismo tiene una componente gigante de 8000 proteínas, y los restantes nodos distribuidos en su gran mayoría en subgrafos inconexos de 2 y 3 proteínas. En adelante, trabajaremos sólo con la componente gigante de este grafo, que denotaremos $G_{HC}(\mathcal{N}, \mathcal{E}, \mathcal{W} = S)$. La misma cuenta, luego de remover bucles y aristas duplicadas con $N = 8000$ nodos y $M = 30835$ conexiones. En esta red, cada arista es soportada por evidencia que proviene de varios experimentos. En el grafo G_{HC}, las 30835 aristas están soportadas por evidencias de 106671 ensayos sobre 112 tipos de experimentos diferentes. A fin de caracterizar la composición de la red G_{HC} considerada, se consigna en la figura 3.1 las clases de experimentos de mayor frecuencia que abarcan en su conjunto el 94.7 % de los ensayos presentes en esta red. Se reporta además, el porcentaje de ensayos que abarca y la calidad q asignada en HIPPIE para cada caso. Notamos que, exceptuando experimentos de tipo “in vivo”, “in vitro ”o “pull down ”que abarcan en conjunto el 31.2 % de los ensayos, la mayor parte de los experimentos de red involucran experimentos de media o alta calidad.

Una característica notable de la red G_{HC} es que el 99.6 % de sus aristas $e \in \mathcal{E}$ están soportadas por al menos un ensayo de calidad $q >= 5$. Además, cada arista posee un promedio de 3.4 ensayos que dan origen a la misma. Por lo tanto, si bien en la figura 3.1 se evidencia que hay aproximadamente un 35 % de los ensayos involucrados en G_{HC} con baja calidad ($q \lesssim 3$), los mismos son en su basta mayoría complementarios a otros existentes de mayor calidad y en raros casos definen interacciones por sí mismos.

3.3. Caracterización topológica

3.3.1. Evidencias preliminares de patrones de conectividad no triviales y estructura modular

En primer lugar, se analizará el comportamiento de las principales características topológicas de la red G_{HC} sin considerar los pesos relativos de las aristas asignados en HIPPIE. Se comenzará estudiando la distribución de grado, el betweenness y el coeficiente de agrupamiento de los nodos de la red. La figura 3.2 consigna la distribución de grado de la red G_{HC} (círculos grises). La misma presenta una cola pesada que refleja grandes fluctuaciones en la conectividad

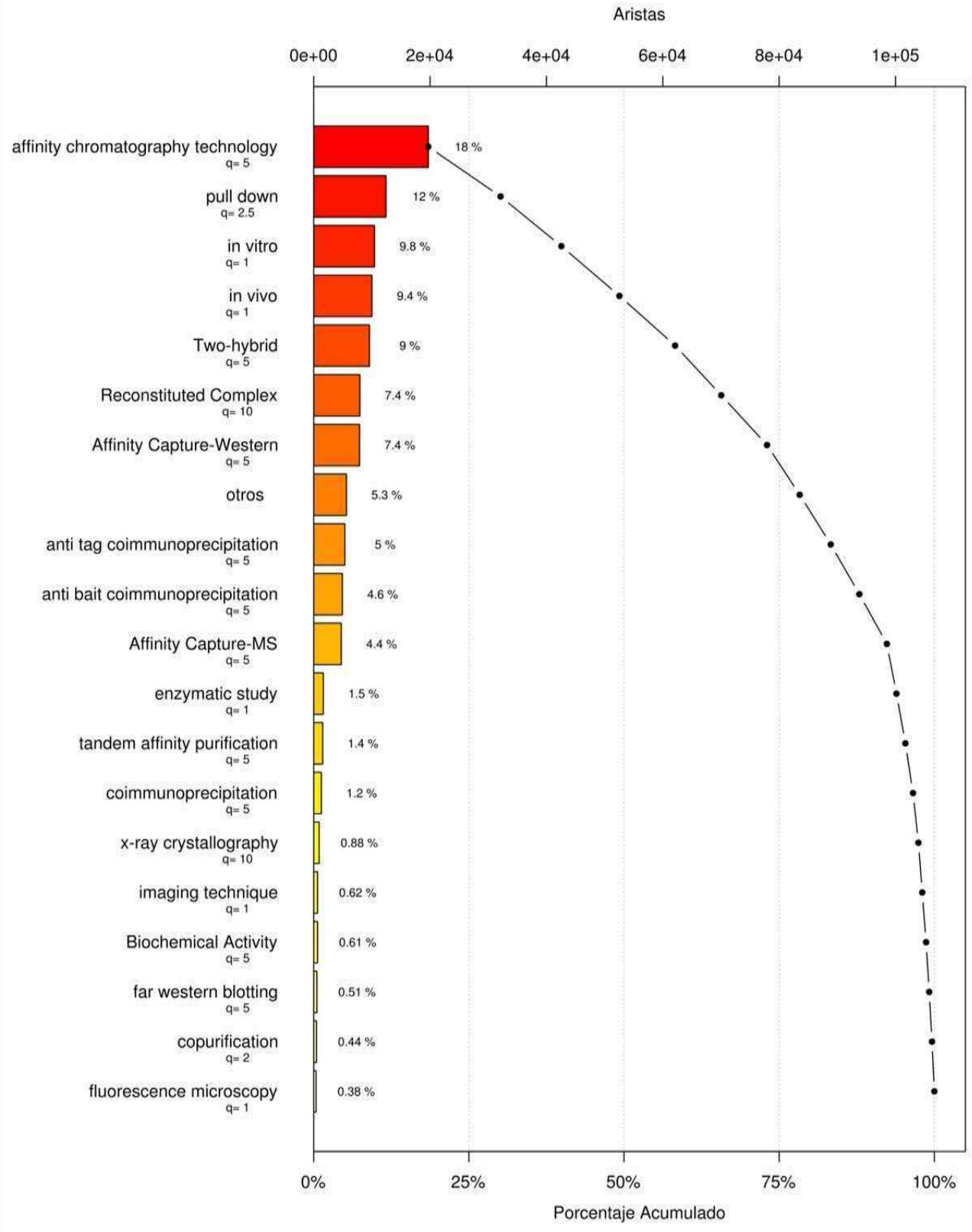


FIGURA 3.1: Evidencias experimentales predominantes en la PIN de alta calidad considerada G_{HC}. Las técnicas que se describen en este gráfico representan el 94.7 % de los 106.671 ensayos involucrados en la red G_{HC}. En la escala superior se representa el número de ensayos que representa cada experimento, mientras que a escala inferior se representa el porcentaje acumulado. En cada caso se denota además, el valor de calidad q asignado en HIPPIE a la técnica correspondiente y el porcentaje de experimentos que involucra. Exceptuando casos de técnicas experimentales *pull down*, *in vitro* e *in vivo* que abarcan en conjunto un 31.2 % de los ensayos, la vasta mayoría de los experimentos involucrados son de calidad media o alta ($q \geq 5$).

de la red, en un rango que abarca nodos con $k \in [1, 492]$, con un valor medio $\bar{k} = 7.7$. En la misma figura, se consigna la distribución de grado de un modelo nulo de Erdos Renyi, (G_{ER} triángulos amarillos), que conservan la cantidad total de nodos y aristas de la red (ver sección 2.4). La distribución de grado correspondiente al modelo G_{ER} sigue, tal como se mencionó en el capítulo anterior, una distribución de Poison de parámetro $\lambda = \bar{k} = 7,7$, y presenta por tanto una escala natural para la conectividad de sus vértices. En suma, se observa una discrepancia fundamental en las distribuciones de grado de la red G_{HC} respecto al modelo nulo G_{ER} considerado, poniendo de relieve la existencia de patrones de correlación no triviales en la conectividad de los vértices que presenta G_{HC}. Cabe mencionar también, que la heterogeneidad observada en el grado de la red implica la existencia de nodos de alta conectividad que pueden actuar como vías de *atajo*, acortando en general las distancias geodésicas en la red.

Por otro lado, en términos de la capacidad de difusión de información global en la red, también pueden existir nodos relevantes más allá del grado de conectividad de los mismos. La medida de centralidad betweenness realza la importancia relativa de los nodos de la red desde este punto de vista (ver sección 2.2.5). La figura Figura 3.3 presenta la relación entre el betweenness (*bet*) y el grado k de los nodos de la red G_{HC} (círculos grises), para el grafo G_{ER} (triángulos amarillos), y para el modelo configuracional presentado en la sección 2.4 G_{CF} (rombos rojos) que respeta la distribución de grado exacta de la red G_{HC}. En los tres casos se observa un crecimiento monótono lo cual denota una tendencia general evidente de una correlación positiva entre estas dos variables topológicas. Sin embargo, cabe destacar que para un dado grado fijo k, la red biomolecular G_{HC} presenta una mayor dispersión en los niveles de betweenness. Además, en esta red existe una mayor fracción de nodos de alto betweenness y bajo grado respecto a los modelos nulos analizados. Si sólo se consideraran nodos de bajo grado, (por ejemplo $k < \bar{k} = 7.7$), se encontrarían 49 nodos en la red G_{HC} por encima del percentil 90 % de la distribución de betweenness de todos los nodos de la red. En contraste, para el modelo configuracional G_{CF} sólo se hallarían 4 nodos, y ninguno en el caso del modelo G_{ER}. Estos nodos de bajo grado y alto betweenness, usualmente denominados *cuellos de botella*, se encuentran sobrerrepresentados en la red real G_{HC}, constituyendo un conjunto de proteínas interesante *per-se*, ya que podrían desempeñar un rol central como intermediarios en procesos de transmisión de información que tiene lugar en toda la red [67, 68].

Para estudiar los patrones de conectividad locales resulta útil analizar el coeficiente de agrupamiento c_i que cuantifica cuan densamente conectada se encuentra la vecindad de un dado nodo (ver sección 2.2.6). La figura 3.4 muestra el coeficiente de agrupamiento de los nodos de la red como función del grado, para la red G_{HC} y los dos modelos nulos considerados. Se observa

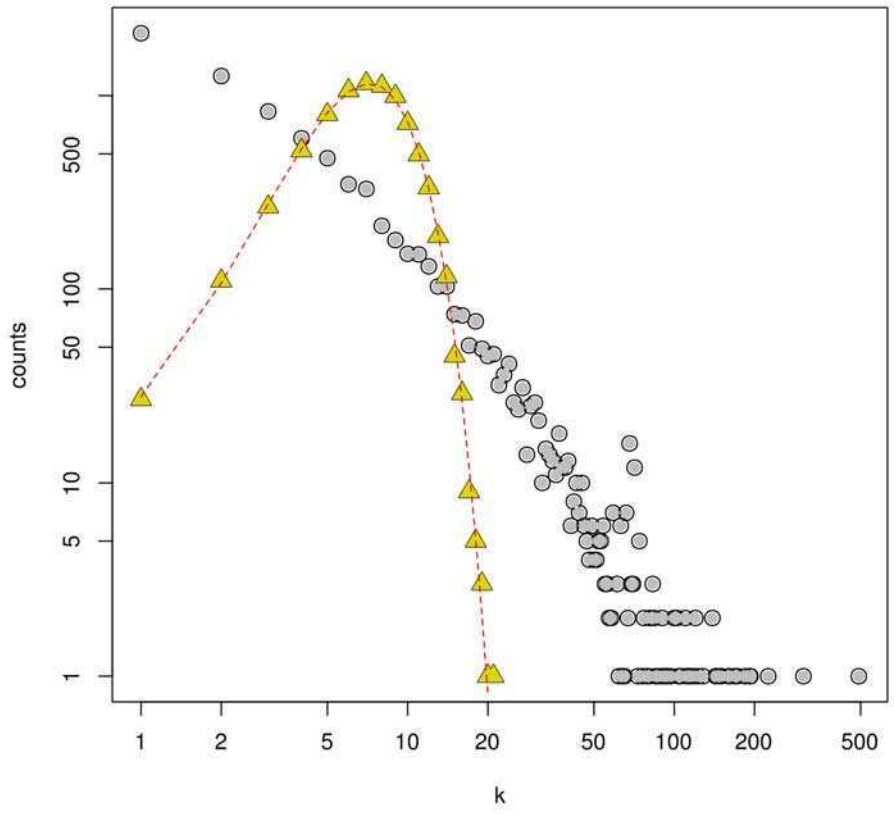


FIGURA 3.2: Distribución de grado para la red HIPPIE G_{HC} y un modelo nulo de Erdos-Renyi G_{ER} denotados con círculos grises y triángulos amarillos respectivamente. El eje de ordenadas consigna la cantidad de nodos de la red con un dado grado k. La línea roja de trazos describe una distribución de Poison de parámetro $\lambda = \overline{k} = 7.7$. En contraste con el modelo nulo G_{ER} que tiene una escala bien definida, la red G_{HC} una cola pesada en su distribución de grado, indicando la presencia de heterogeneidades no triviales en la distribución de conectividades de la red.

como tendencia general una correlación negativa entre estas variables para las tres redes. Sin embargo, es evidente también la tendencia general de la red G_{HC} a presentar nodos que, para un mismo nivel conectividad k, tienen mayores valores en sus coeficientes de agrupamiento comparados con los modelos nulos considerados. Esto sugiere la existencia de estructuras densamente conectadas compatibles con la presencia de organización modular en la red.

3.3.2. Correlaciones de grado de segundo orden

Para analizar correlaciones de mayor orden en las conectividades de la red, graficamos en la figura 3.5 el grado medio de primeros vecinos de cada nodo de la red como función del grado del mismo $knn(k)$ (ver sección 2.2.4.1). Para un dado nodo, el valor de knn es el promedio del

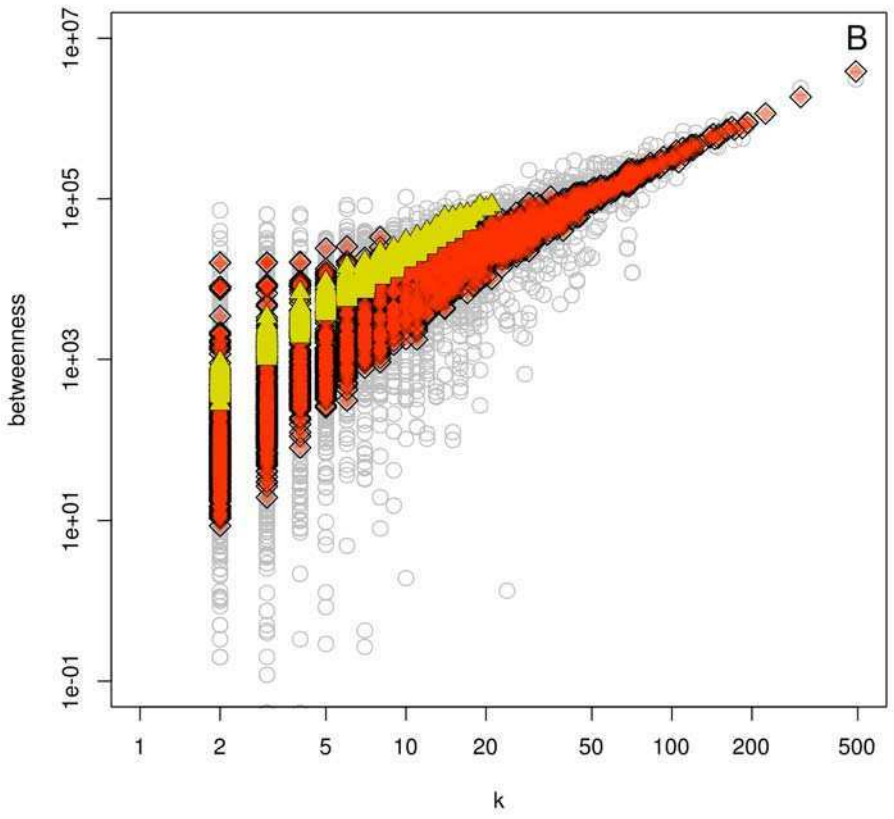


FIGURA 3.3: Betweenness en función del grado de conectividad k para la red HIPPIE G_{HC} el modelo nulo de Erdos-Renyi G_{ER} y un modelo configuracional G_{CF}, denotados con círculos grises y triángulos amarillos y rombos rojos respectivamente. Puede apreciarse en la red G_{HC} la presencia de múltiples nodos de baja conectividad ($k < \bar{k}$) pero alto nivel de Betweenness (por arriba del cuantil 90 %)

grado de sus primeros vecinos. Luego, tomando todos los nodos de grado k y promediando sus respectivos valores de *knn* podemos obtener un gráfico como el que se presenta en la figura 3.5, donde círculos grises, triángulos verdes y rombos rojos denotan los valores de $knn(k)$ para la red G_{HC}. Las notables fluctuaciones que se observan en la figura impiden establecer un comportamiento monótono creciente o decreciente en esta relación, dificultando la clasificación de la red en un comportamiento plentamente asortativo o disortativo. A modo de control se calcularon valores de $knn(k)$ sobre un ensamble de 1000 modelos configuracionales que respetan exactamente la distribución de grado de la red G_{HC}. Dentro de los límites de la región sombreada se encuentra el 95 % de los valores $knn(k)$ obtenidos en este ensamble de redes. Valores de $knn(k)$ del grafo oringinal por encima de esta región (triángulos verdes) o por debajo de ella (rombos rojos) resultan estadísticamente significativos a un nivel de confianza del 5 % (en referencia al

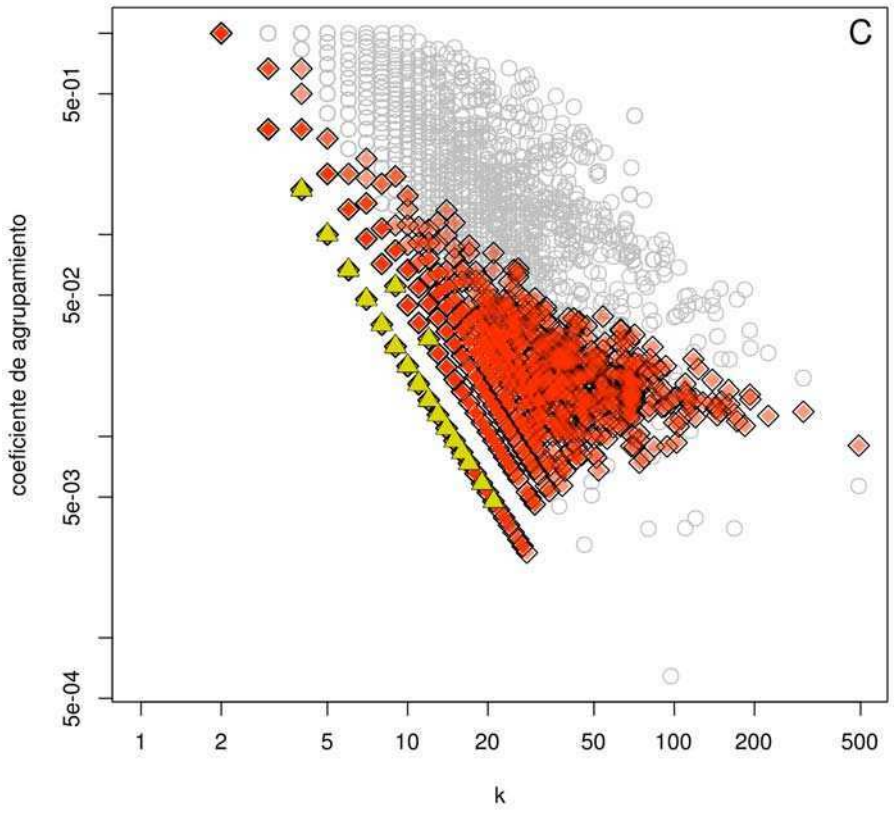


FIGURA 3.4: Coeficiente de agrupamiento en función del grado de conectividad k para la red HIPPIE G_{HC}, el modelo nulo de Erdos-Renyi G_{ER} y un modelo configuracional G_{CF}, denotados con círculos grises y triángulos amarillos y rombos rojos respectivamente. Puede observarse una tendencia general de los nodos de G_{HC} a presentar mayor coeficiente de agrupamiento que en ambos modelos nulos considerados.

modelo configuracional).

Para nodos de grado $k \leq 2$ (que representan el 42 % de los nodos de la red) el valor observado de $knn(k)$ esta por encima del control, indicando que estos nodos de bajo grado en la red G_{HC} están típicamente conectados a nodos de grado mayor al que cabería esperar por azar (compatible con un comportamiento disortativo). En concordancia se observan numerosos casos de nodos de alto grado conectados a nodos de menor conectividad que lo que cabría esperar por azar (rombos rojos), compatible también con un comportamiento disortativo. Por otro lado también es relevante destacar la existencia de nodos cercanos a la recta de pendiente unitaria (línea de trazos) cuyo valores de $knn(k)$ difieren significativamente del modelo nulo de referencia, tanto por exceso como por defecto. Estos son nodos conectados a otros de similar conectividad y de manera no trivial, o al menos estadísticamente significativa. Por lo tanto, esta región del gráfico

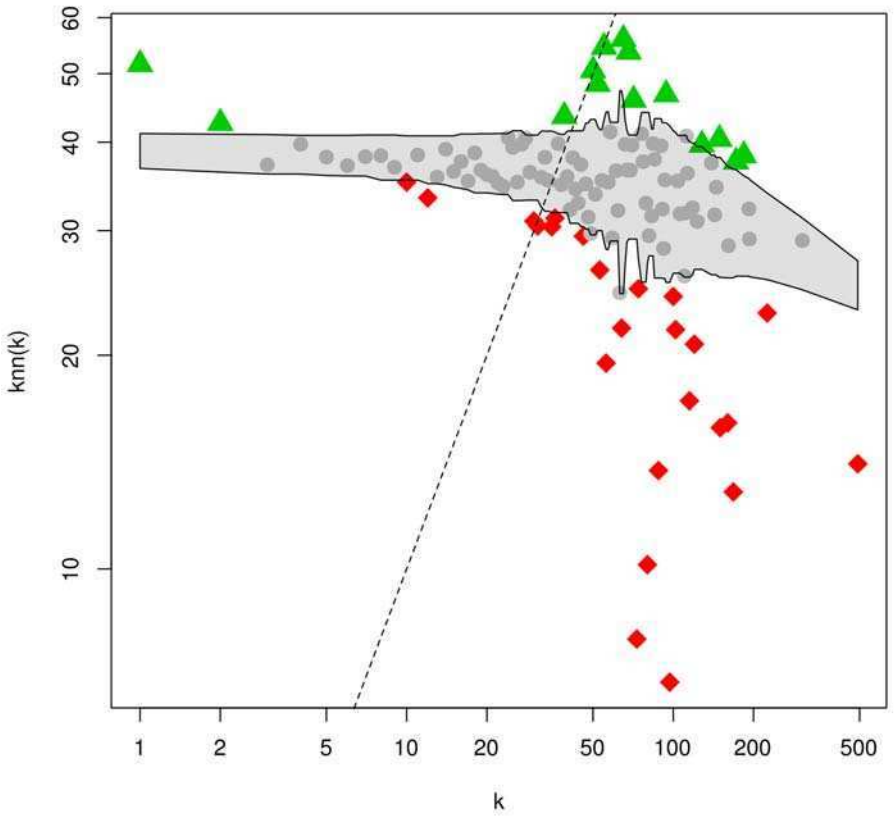


FIGURA 3.5: Grado medio de primeros vecinos en función del grado ($knn(k)$) para la red G_{HC} (puntos grises, triángulos verdes y rombos rojos). El área sombreada representa un intervalo de confianza del 95 % calculado sobre un ensamble de 1000 modelos configuracionales que respetan exactamente la distribución de grado de la red G_{HC}. Valores de $knn(k)$ por encima de esta zona (triángulos verdes) involucran nodos conectados a vecinos de grado más alto que lo que cabe esperar por azar en un modelo configuracional (significativo al 5 %). Análogamente, valores por debajo de esta zona involucran nodos conectados a otros de menor grado respecto al modelo configuracional. Los valores de correlación de grado de segundo orden para nodos dentro de la zona sombreada (círculos grises) no difieren del modelo nulo empleado. La línea de trazos representa una recta de pendiente unitaria que se muestra a modo de referencia. La figura es compatible en general con un comportamiento disortativo (nodos de bajo grado conectados a otros de alto grado), aunque también puede observarse un rango de comportamiento asortativo (ver texto).

presenta características compatibles con un comportamiento asortativo. En suma, la red presenta correlaciones de segundo orden no triviales que difieren del comportamiento esperado en redes provenientes de modelos configuracionales, mostrando tanto regiones de comportamiento disortativo como otras de comportamiento asortativo.

3.3.3. Los tripletes de la red se conforman típicamente con aristas soportadas por múltiples ensayos experimentales.

En secciones previas se ha mostrado que la red considerada presenta interacciones de alta fiabilidad, en el sentido que las conexiones están soportadas en su extensa mayoría por al menos un ensayo experimental de mediana o alta calidad. Más aún, se mostró que en promedio hay más de 3 ensayos experimentales que soportan la presencia de cada arista de la red. No obstante es relevante preguntarse si existe algún tipo de sesgo en la forma que las distintas evidencias experimentales están distribuidas, tanto en calidad como en cantidad de ensayos realizados. Con esta idea en mente, se analizará en que medida se relacionan los patrones de conectividad de la red con los niveles de confianza de cada arista reportados por HIPPIE (ver ec. 3.2). En particular, se estudiará la composición de los triángulos del grafo (una de las evidencias ya analizadas de conectividad no trivial que presenta la red) en relación al nivel de calidad y cantidad de experimentos involucrados en sus aristas constituyentes.

Recordemos que, el coeficiente de agrupamiento total de la red, una cantidad directamente vinculada al número de triángulos observados en la red, puede calcularse al menos de dos maneras. La primera es promediando el coeficiente de agrupamiento de cada nodo, calculado bajo la suposición que todas las aristas tienen la misma importancia relativa, $C = \sum_i^N c_i$ (ver ecuación 2.7). La segnda forma es mediante la asignación de un peso relativo a todas las aristas de la red, $C^w = \sum_i^N c_i^w$ (ver ecuación 2.8). Tal como fue explicado el capítulo 2.2.6, de la comparación de C^w y C puede concluirse si los triángulos están típicamente conformados por aristas de alto o bajo peso. En la figura 3.6 se muestran los valores de coeficiente de agrupamiento de la red G_{HC} calculados mediante la consideración de distintos pesos en sus aristas. La línea horizontal roja de trazos en esta figura, denota como referencia el valor de C (sin asignación de pesos relativos en las aristas). Se presentan además tres alternativas distintas para C^w. En la primer condición se calcula $C^{w=S}$ es decir, utilizando el vector de pesos relativos para las aristas S provisto por HIPPIE (ver ecuación 3.2) y se lo consigna en la figura con un rombo verde en la condición *HIPPIE Score*.

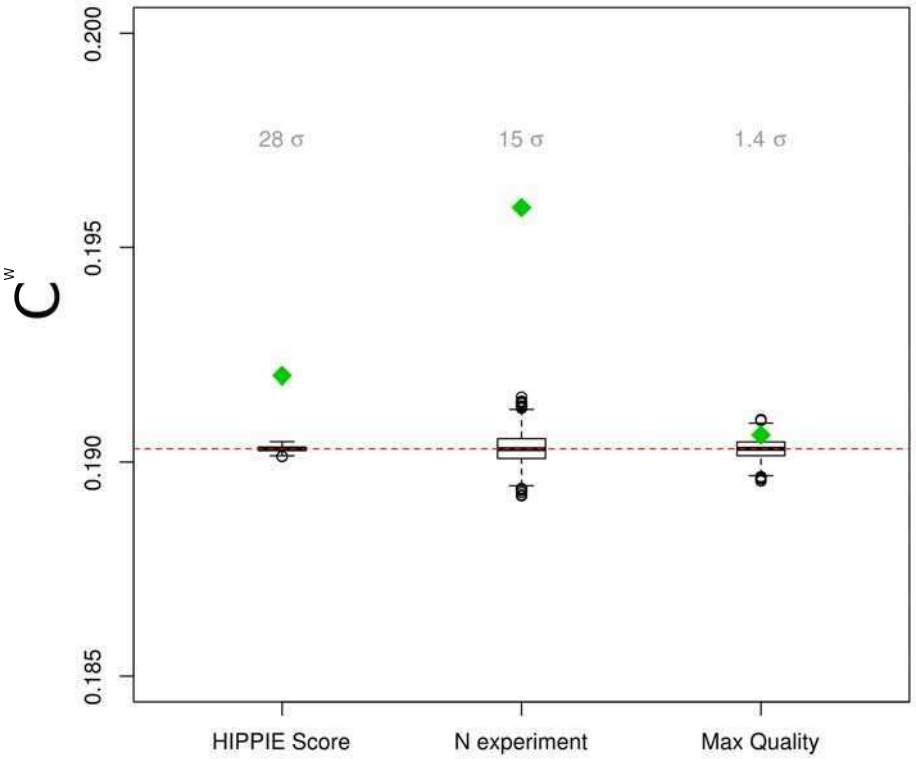


FIGURA 3.6: Composición de triángulos en la red G_{HC}. El eje de ordenadas consigna los valores de coeficiente de agrupamiento C^w calculados con cuatro tipo de pesos diferentes para sus aristas. Primero considerando todas las aristas equivalentes (es decir, $C^{w=1} = C$), valor que se denota mediante la línea colorada de trazos horizontal. Los restantes casos representados con rombos verdes consideran: el peso S calculado según 3.2 (*HIPPIE Score*), la cantidad de experimentos como observable para asignar los pesos (*N experiment*) y el máximo valor de calidad q observado en los experimentos de cada arista (*Max Quality*). Diferencias significativas entre los valores de C^w y C se observan para el uso de $w = S$ y $w = Nexperiment$ (a 28σ y 15σ respectivamente), dando soporte a la idea que los triángulos de la red están típicamente poblados por aristas que cuentan con numerosa evidencia experimental, pero no necesariamente con experimentos de máxima fiabilidad.

Observamos que $C^{w=S} > C$, lo que indicaría en principio que los triángulos del grafo puedan estar conformados por aristas de alto peso S. Para validar esta afirmación se debe garantizar que la diferencia observada entre $C^{w=S}$ y C no puede atribuirse a fluctuaciones aleatorias. Para ello, se calculó el coeficiente de agrupamiento $C^{w=S}$ sobre un ensamble de 1000 grafos de idéntica topología a G_{HC} pero con los pesos S redistribuidos aleatoriamente. Este control se muestra con el gráfico de cajas en la misma columna *Hippie Score* de la figura 3.6. Como consigna la leyenda superior de la figura, las diferencias observadas entre $C^{w=S}$ y C es de 28 desvíos estándar (notar que el valor de C coincide con la mediana de la distribución aleatoria) por lo que

podemos afirmar que la diferencia entre $C^{w=S}$ y C no puede ser atribuidas a fluctuaciones.

Ahora bien, el peso aquí utilizado $w = S$ es función del número de experimentos que soporta la evidencia experimental de cada arista, la calidad de esos experimentos y el número de experimentos complementarios en otros organismos. Por tanto cabe preguntarse si los triángulos de la red que presentan altos niveles de S tienen típicamente aristas soportadas por experimentos de alta calidad y/o con elevado número de ensayos. Calculamos entonces los valores de C^w primero considerando el número de ensayos que soportan cada interacción ($C^{w=NExp}$) y en segundo lugar, la máxima calidad q observada entre todos los experimentos que soportan a cada arista ($C^{w=max\{q\}}$). Ambos resultados se denotan con rombos verdes en la fig 3.6, y nuevamente, para ambos casos se consigna el cálculo de C^w, sobre ensambles de 1000 grafos idénticos con los mismos pesos distribuidos aleatoriamente. Resulta interesante notar que $C^{w=Nexp} > C$ (de hecho la diferencia es de 15 desvíos estándar), y que lo mismo no sucede en el caso donde sólo la máxima calidad q entre los experimentos de cada arista es considerada. Esto último se deduce en vistas que las diferencias observadas entre $C^{w=max\{q\}}$ y C no difieren significativamente de las fluctuaciones aleatorias. En efecto, estos resultados muestran que la red HIPPIE considerada tiene más triángulos de los que cabería esperar por azar (ver figura 3.4) y los mismos tienen un sesgo a constituirse por aristas con alto número de evidencias experimentales. Asimismo, los ensayos de máxima calidad en la red no tienen un sesgo claro a situarse en estos tripletes, lo que refuerza la confianza a priori que pueda tenerse en las conexiones de pares de nodos que no formen parte de tripletes en la red.

3.4. Conclusiones

En este capítulo se presentó y caracterizó la red de interacción de proteínas de *alta calidad* extensamente utilizada en los capítulos siguientes. La misma se extrajo de la base de datos integrativa HIPPIE y presenta interacciones soportadas casi en su totalidad por experimentos de mediana o alta calidad, tales como Inmunoprecipitación, Complejos reconstituidos, cristalografía de rayos X, o experimentos doble híbrido por ejemplo. En efecto, el 99.6 % de las aristas de la red G_{HC} están soportadas por al menos una evidencia de calidad media o alta, y presentan en promedio 3.4 ensayos independientes por cada arista del grafo. Por otro lado, desde el punto de vista topológico, los resultados presentados en este capítulo sugieren la existencia de heterogeneidades topológicas no triviales compatibles con una organización modular subyacente en la red biomolecular bajo estudio G_{HC}.

Por otra parte, las diferencias estructurales observadas con respecto a las redes aleatorias consideradas resaltan la presencia de componentes no triviales en la red G_{HC}, como proteínas de baja conectividad y alto betweenness, que usualmente actúan como importantes vínculos entre estructuras modulares de la red. Existe también una notable cantidad de proteínas que presentan altos valores de coeficiente de agrupamiento en relación a lo que cabría esperar en los modelos nulos considerados. Esto deja en evidencia la existencia de numerosos triángulos en la red que no se distribuyen de manera aleatoria. Más aún, se ha mostrado una tendencia general de estos triángulos a constituirse por aristas que están soportadas por un alto número de ensayos experimentales. En contraste, se observó que los ensayos con mayor nivel de calidad no tienen una sesgo particular a conformar estos triángulos, hecho que incrementa la fiabilidad en las conexiones de la red que no forman tripletes interconectados.

Por otro lado, la red G_{HC} presenta correlaciones de conectividad de segundo orden no triviales, dado que se observa una cantidad inusualmente alta de nodos de bajo grado conectados a otros de alto grado, comportamiento típico de una red disortativa. Al mismo tiempo se observan nodos conectados a otros de similar grado y cuyos niveles de correlación de segundo orden difieren de los modelos nulos empleados, evidenciando la presencia de zonas donde la red tiene un comportamiento asortativo.

Capítulo 4

Estructura modular en redes de interacción de proteínas a diferentes niveles de resolución

4.1. Resumen

El análisis de estructura modular en redes biológicas ha recibido mucha atención en los últimos años, impulsada por la creciente disponibilidad de grandes cantidades de datos generados por tecnologías ómicas (genómica, proteómica, transcriptómica, metabolómica, etc.). Por otro lado, la falta de una definición matemática unívoca para el concepto de módulo en una red, plantea el interrogante de cuan adecuados resultan los diferentes métodos de reconocimiento de comunidades en términos de su capacidad para descubrir nuevo conocimiento biológico. Una de las contribuciones principales de éste capítulo es el estudio de cómo la consideración de distintos algoritmos puede incidir en los análisis biológicos subsecuentes basados en la estructura modular de la red considerada.

Para abordar esta problemática, se ha considerado la red de interacción de proteínas humana (PIN)[69] presentada en el capítulo precedente, y dos algoritmos de reconocimiento de comunidades ampliamente utilizados, uno de ellos basado en teoría de información, *Infomap* [13], y otro basado optimización de *modularidad*, Clauset-Newman-Moorem [70] (*CNM*).

En primer lugar, se caracterizó cada algoritmo en términos de la granularidad de las particiones inferidas, la homogeneidad biológica de las estructuras detectadas y en base a la descripción cartográfica de la red que cada algoritmo es capaz de revelar. Se analizaron dos conjuntos

de proteínas de particular interés biológico. Uno de ellos comprende proteínas asociadas a diferentes vías de señalización, y el otro proteínas vinculadas a procesos de envejecimiento celular. Se encontró que el nivel de resolución provisto por la descripción de *Infomap*, resulta más apropiado para detectar esquemas modulares de conectividad inter/intra PIN específicos de estos conjuntos de proteínas. En particular, se encontró que proteínas asociadas a procesos de envejecimiento como SIRT1 y CDKN2A , presentan altos niveles de conexión intermodular cuando la partición de *Infomap* es considerada, lo que sugiere que podrían servir para coordinar flujos de información entre módulos de funcionalidad biológica específica. Además, en contraste con el algoritmo *CNM*, el procedimiento *Infomap* resalta principios de organización global en proteínas de señalización: mientras que proteínas receptoras están asociadas a nodos periféricos de la red, proteínas de andamiaje, como PICK1 , PLCG1 y GRB2, se encuentran enriquecidas en nodos de alto grado (conectando diferentes módulos *Infomap*). Así mismo, ligandos y factores de transcripción se ven asociados a nodos de la red escasamente conectados.

4.2. Introducción

Uno de los principales desafíos de la biología de sistemas es comprender las bases celulares y moleculares de funciones biológicas de alto nivel y de fenotipos complejos. Un enfoque prometedor para tratar estos problemas se basa en la caracterización de la funcionalidad celular en términos de una descripción global del intrincado conjunto de reacciones bioquímicas que tienen lugar dentro de la célula. Este enfoque sistémico ha recibido mucha atención en los últimos años, impulsado por la creciente disponibilidad de grandes cantidades de datos generados en escalas ómicas y el desarrollo de centros de datos públicos, así como los esfuerzos de curación encaminados a organizar y agregar valor a la acumulada evidencia experimental de interacciones moleculares.

En este contexto, y partiendo de la inmensa cantidad de datos disponibles, la metáfora de redes ha aparecido como un marco atractivo para organizar y revelar patrones globales de relevancia biológica. El enfoque de redes proporciona un lenguaje de descripción sistemático basado en relaciones de pares (es decir, aristas o arcos de la red) entre las entidades de interés (es decir, los nodos o vértices de la red). Este enfoque resulta particularmente útil para revelar patrones de conectividad e interacción en el contexto de diversas funciones biológicas [71, 72]. Ha mostrado también utilidad para asignar nuevas funciones a productos génicos no anotados [73], para proponer biomarcadores en patologías diversas [74], obtener información sobre relaciones genotipo-fenotipo [75–77], y establecer asociaciones significativas entre fenotipos patológicos

y perturbaciones disruptivas que involucran regiones particulares de las redes de interacción de proteínas subyacentes [2, 75, 78, 79].

El fundamento del enfoque basado en redes es que sus características topológicas pueden revelar biología relevante. En este contexto, una estrategia recurrente consiste en la identificación de nodos relevantes, de acuerdo a distintos índices de centralidad de la red, con la esperanza de poder reconocer entidades biológicas significativas. Siguiendo esta línea de investigación, varios estudios han sugerido, por ejemplo, que proteínas centrales en redes de interacción física en levaduras (*S. cerevisiae*) son más propensas a ser esenciales que otras proteínas, dando lugar a la llamada regla de centralidad-letalidad [80–84]. Las descripciones modulares de redes biológicas también han recibido mucha atención en los últimos años [5]. En este sentido, se ha puesto mucho esfuerzo en establecer y aprovechar correlaciones significativas entre grupos topológicos de la red (formados a partir de nodos más densamente conectados entre sí que respecto a otros nodos de su vecindad), y la idea original de Hartwell sobre *módulos de funcionalidad biológica*, definidos como un conjunto de componentes moleculares y sus interacciones que llevan a cabo una función biológica específica [6].

El análisis de estructura modular en redes biológicas moleculares ofrece una descripción amplia y global de patrones de interacción que contribuye a la comprender los complejos mecanismos de organización del sistema biológico en estudio. En particular, una interesante descripción modular en roles cartográficos de redes moleculares fue introducida por Guimera y Amaral [7, 8, 85](ver capítulo 2.5). Aprovechando la estructura modular de la red, y una vez establecida una partición en comunidades disjuntas, propusieron clasificar a los nodos de la red en función de sus patrones de conectividad dentro y fuera del módulo al que pertenecen diferenciando siete categorías cartográficas distitnas [7]. Con ese fin se introdujeron dos observables: la conectividad intramodular (Z), y el coeficiente de *participación* de un nodo (P). Mientras que el primer parámetro describe el grado de un nodo relativo al grado de los restantes nodos en su misma comunidad, la segunda cuantifica en qué medida un nodo se conecta con nodos en otras comunas (para más detalles, ver capítulo 2.5).

Utilizando esta metodología los autores fueron capaces de caracterizar distintas representaciones cartográficas informativas en diversas redes metabólicas. Además, demostraron que los nodos de alta *participación* y bajo *grado intramodular* detectados en la red metabólica de E. coli tienden a mostrar tasas de evolución inusualmente bajas, lo que sugiere que la biología relevante puede ser recapitulada con la metodología propuesta [8]. Un factor atractivo del análisis de Guimera es que por ser una metodología que hace uso explícito de la estructura modular, no se basa en características de la red estrictamente locales (es decir, considerando sólo las propiedades de

un nodo y sus vecinos directos), ni estrictamente globales, sino más bien que considera patrones de conectividad a nivel de mesoescala. De hecho, la misma noción de comunidad se utiliza con el fin de establecer una escala característica sobre la cual se realiza posteriormente el análisis de conectividad.

Sin embargo cabe destacar que la identificación de módulos en redes complejas es un problema matemáticamente mal definido, en el sentido de que no existe una definición objetiva y libre de hipótesis para "modulo"en una red. Esto da lugar a la coexistencia de diferentes procedimientos para reconocer comunas en grafos (ver [26] para una revisión extensa), los cuales pueden producir diferentes particiones de red. Por otra parte, cabe preguntarse, cuál es la capacidad de cada una de estas metodologías para dar a conocer patrones biológicamente relevantes.

En el presente capítulo se aborda esta temática, estudiando las implicancias que las eventuales discrepancias de distintas descripciones modulares pueden producir en el análisis biológico subsecuente. El trabajo está centrado en dos aspectos importantes del problema. Por un lado, se exploran posibles asociaciones entre grupos funcionales y los módulos topológicos identificados, analizando la homogeneidad biológica de las estructuras halladas. Por otro lado, teniendo en cuenta la descripción cartográfica de Guimera, hemos explorado características de conectividad de la red a nivel de mesoescala inducidas por las diferentes estructuras modulares consideradas.

En relación a este punto se estudió la capacidad de cada metodología de reconocimiento de comunas, para revelar patrones de conectividad características en distintos conjuntos de proteínas asociados a fenotipos o funcionalidades biológicas de interés, identificando sesgos y tendencias topológicas sensibles a los fenotipos bajo estudio. En particular, se presentan análisis para dos conjuntos de interés: el primero de proteínas relacionadas con procesos de envejecimiento celular y el segundo de proteínas involucradas en vías de señalización. De este modo, se analizó si estos fenotipos complejos y funcionalidades biológicas de alto nivel pueden relacionarse con patrones de conectividad intermodulares o intramodulares específicos.

Respecto a las metodologías de agrupamiento consideradas , el análisis se focaliza en dos procedimientos paradigmáticos de detección de comunidades en redes: el algoritmo de Clauset-Newman-Moore (*CNM*) [12], y el algoritmo *Infomap* [13]. Estas conocidas metodologías utilizan criterios de optimización cualitativamente muy diferentes. La primera es parte de una amplia familia de procedimientos de detección de comunidades basados en la optimización de una figura de mérito conocida como *modularidad* de la red. A pesar de su amplia utilización, cabe destacar que Fortunato y Bartolomé demostraron en el 2007 la existencia de un límite teórico de resolución para este tipo de algoritmos. Este límite de resolución, conduce a la fusión sistemática de pequeños grupos en módulos más grandes, incluso cuando los grupos estén bien definidos

y mínimamente conectados entre sí [86]. Desde entonces, muchas contribuciones desarrolladas principalmente desde la comunidad física, exploraron aún más este efecto, proponiendo metodologías alternativas y estableciendo estudios comparativos mediante modelos ad-hoc de redes de referencia [87–91].

Por otro lado, el algoritmo *Infomap* se basa en un criterio de optimización muy diferente. Los módulos son definidos mediante la minimización de la longitud con la que se describe un proceso de paseo al azar que tiene lugar sobre la red. Se sabe además, que este procedimiento no se ve afectado por el efecto de límite de resolución, al menos en el contexto de las redes de referencia propuestas por Lancichinetti [90].

Pese a estos avances, los algoritmos basados en maximización de modularidad son todavía una de las alternativas más populares para la detección de estructuras en comunidades sobre grafos. En particular, en concordancia con el procedimiento de detección de comunidades empleado por Guimera en su serie original de trabajos, muchos análisis recientes de diversos problemas biológicos basados en detección de comunas en redes [9–11] se abordan considerando ligeras variaciones del mismo tipo de algoritmo, todos ellos basados en optimización de *modularidad*.

En este contexto, el análisis comparativo presentado en este capítulo sirve para ilustrar y llamar la atención acerca de cómo la idiosincrasia de un algoritmo considerado puede impactar en el análisis y conclusiones biológicas subsecuentes, al utilizar redes de interacción de proteínas.

Este capítulo se organiza de la siguiente manera. Primero se presenta la red de interacción de proteínas utilizada (PIN) y dos vías alternativas de reconocimiento de comunidades. Se consideran dos metodologías bien conocidas y ampliamente utilizadas en la comunidad de redes: el algoritmo de Clauset -Newman -Moore (*CNM*) [12], y el algoritmo *Infomap* [13] los cuales tienen sus raíces en criterios de optimización cualitativamente diferentes. Luego se lleva a cavo una caracterización de las particiones obtenidas en cada caso en términos del nivel de resolución y coherencia biológica de las estructuras detectadas. Se presenta también un análisis cartográfico de la red y como el mismo se ve afectado por el nivel de resolución subyacente. Además se llevó a cabo la búsqueda de asociaciones significativas entre roles cartográficos específicos y diferentes conjuntos fenotípicos de proteínas, uno asociado a procesos de envejecimiento en homo sapiens, y otro vinculado a diversas vías de señalización de proteínas. Finalmente, los resultados obtenidos son analizados y discutidos en términos de sus posibles implicancias biológicas.

4.3. Red de interacción de Proteínas considerada

En lo siguiente, consideraremos la red de interacción de proteínas descripta y caracterizada en el capítulo precedente, HIPPIE (Human Integrated Protein-Protein Interaction rEference). Esta red es una base de datos integrada que recapitula información sobre interacciones físicas entre proteínas, puntuadas según la calidad del tipo de evidencia experimental que da origen a las mismas [69]. La versión de mayor calidad de la red (v1.5, descargada en abril 2012) incluye 32321 interacciones entre 8277 proteínas. El análisis realizado de aquí en adelante, se centra en la componente gigante de esta red, que comprende 8.000 nodos y 30835 interacciones, a la cual llamaremos PIN para referencia futura. Un análisis preliminar del tipo de experimentos que dan origen a las interacciones en esta red, la calidad y cobertura de los mismos, así como un análisis preliminar de las características topológicas básicas, fue presentado en el capítulo 3.

4.4. *CNM* e *Infomap* exploran estructuras modulares a diferentes niveles de resolución

Para explorar la organización modular de la PIN se consideran dos metodologías de reconocimiento de comunas ampliamente utilizadas: el algoritmo de Clauset -Newman -Moore (*CNM*) el cuál se basa en optimización de *modularidad* [12], y el algoritmo *Infomap* [13] basado en teoría de información. Una breve descripción de estos criterios de optimización utilizados por cada algoritmo se presentó en el capítulo 2.3.1. Un análisis más exhaustivo sobre comparaciones entre estos algoritmos puede encontrarse en [87, 89, 90].

En primer lugar, exploramos la estructura modular de la PIN mediante el uso de los algoritmos de detección de comunas *Infomap* y *CNM*. Ambos algoritmos reportan particiones que muestran similares niveles de *modularidad* $Q_{Infomap} = 0.52$, $Q_{CNM} = 0.54$. Estos valores son muy superiores a los esperados para un modelo nulo configuracional, es decir un modelo de redes con idéntica distribución de grado que la PIN (ver capítulo 2.4). Si consideramos como conjunto de control un ensamble de 1.000 redes aleatorias G^{conf} generadas según el modelo configuracional, se obtienen valores de *modularidad* $Q^{conf}_{Infomap} = 0.2546 \pm 0.0007$, $Q^{conf}_{CNM} = 0.313 \pm 0.001$. Las diferencias observadas entre Q y Q^{conf} con ambos algoritmos, hacen hincapié en la importancia de las correlaciones de segundo orden (u órdenes superiores) presentes en la PIN en relación a la estructura modular observada.

Pese a que las particiones halladas por ambos algoritmos alcanzan similares valores de *modularidad* (difieren en menos del 4 %), se observan grandes diferencias en términos de la distribución de tamaños de los grupos correspondientes en cada caso. Por ejemplo, mientras que no hubo módulos *Infomap* de tamaño superior a 392 nodos, la partición *CNM* incluye cuatro módulos con más de 1000 nodos cada uno. Otra forma interesante de estimar el tamaño de un *módulo* es calculando el número de arcos o conexiones internas del mismo, l_{int}. Este observable resulta particularmente relevante para entender las características cualitativas de las particiones obtenidas. En la figura 4.1, se muestra la relación entre esta cantidad l_{int} de un módulo y la masa del mismo ($Size$), medida como el número de nodos que ese módulo contiene. Cada cuadrado negro representa un módulo *Infomap*, y cada triángulo rojo representa un módulo CNM. En esa misma figura se muestra a modo de referencia la relación esperada en estructuras modulares completamente conectadas (línea discontinua de trazos) y mínimamente conectadas (línea punteada). Si un módulo tiene n nodos, el primer caso muestra una estructura donde la cantidad de aristas es la máxima posible ($\frac{n(n-1)}{2}$) y el segundo caso representa un módulo conexo con la mínima cantidad de aristas ($n+1$). En esta figrua se observa que para casi todo el rango de masas ($Size \lesssim 492$) existe al menos un módulo *Infomap* con mayor cantidad de aristas que las que se observan en cualquier módulo *CNM* de igual masa.

Podemos también analizar la distribución de masa acumulada a través los diferentes módulos, en relación al tamaño de los mismos medido en términos de l_{int} (Fig.4.2.A). En esta figura se muestra la función de distribución acumulada (fracción acumulada de nodos, $F_{cluster}$), en función de l_{int} para ambas particiones. Los círculos rojos corresponden a CNM y los cuadrados negros a *Infomap*. En el caso de *CNM* se observa un cambio abrupto de $F_{cluster}$ que tiene lugar en un número de arcos internos del orden de $\sqrt{L}$, siendo L es el número total de conexiones de la red. Este cambio cualitativo puede entenderse considerando que el 90 % del número total de nodos de la red queda agrupado en sólo 8 módulos *CNM* (círculos rojos , Fig. 4.2.A) .

Por el contrario, los módulos *Infomap* (cuadrados negros, 4.2.A) presentan un incremento suave en $F_{cluster}$. Esto implica que en este caso, la red puede dividirse en comunas con un amplio espectro de tamaños, sin la posibilidad de identificar una escala característica. De esta forma, los resultados obtenidos para la PIN considerada van en dirección a las observaciones realizadas por Lancichinetti y colaboradores: a diferencia de *CNM* , el algoritmo *Infomap* proporciona una descripción modular capaz de resolver simultáneamente estructuras modulares de muy diferente tamaño [90].

Por otro lado, cabe destacar que casi el 90 % de los arcos internos *Infomap* son también enlaces internos en comunas *CNM* (el 86 % de los pares de nodos dentro de módulos *Infomap*

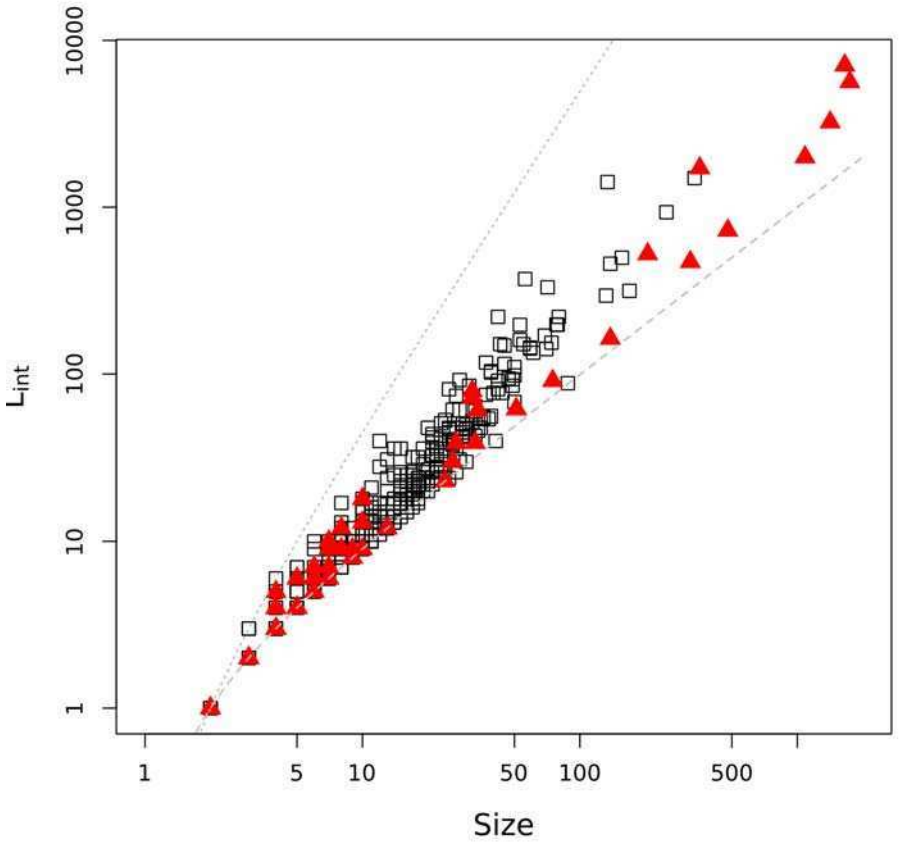


FIGURA 4.1: Análisis de las particiones obtenidas con los algoritmos *CNM* (triangulos rojos) e *Infomap* (cuadrados negros). Se consigna el número de conexiones internas a cada módulo (l_{int}) en función del número de nodos (*size*) del mismo. A modo de referencia se muestra la relación esperada para estructuras tipo *Clique* (máximo número de aritas posibles, línea punteada) y la relación esperada en estructuras lineales, es decir, con $n + 1$ aristas que nodos (línea de trazos discontinua).

se conserva en la partición alternativa CNM), mientras que sólo el 66 % de los enlaces internos *CNM* se conservan como enlaces internos *Infomap* (y sólo el 5 % de los pares de nodos *CNM* se conservan bajo la descripción *Infomap*). Este resultado sugiere que los módulos *Infomap* están virtualmente incluidos dentro de las estructuras de *CNM*.

La fig. 4.2.B consigna la fracción de arcos internos en estructuras *CNM* que unen nodos agrupados en distintos módulos *Infomap* (fracción de arcos recortados, fBL) en función de l_{int}. En esta figura puede observarse que casi la totalidad de fBL (99 %) tiene lugar en las 8 estructuras dominantes de la partición *CNM*, mientras que sólo el 1 % ocurre en módulos *CNM* con baja densidad de arcos internos.

A fin de ganar intuición respecto a cómo se relacionan ambas descripciones modulares, mostramos en la Figura 4.3 10 ejemplos de estructuras *CNM* de baja densidad de arcos, con un

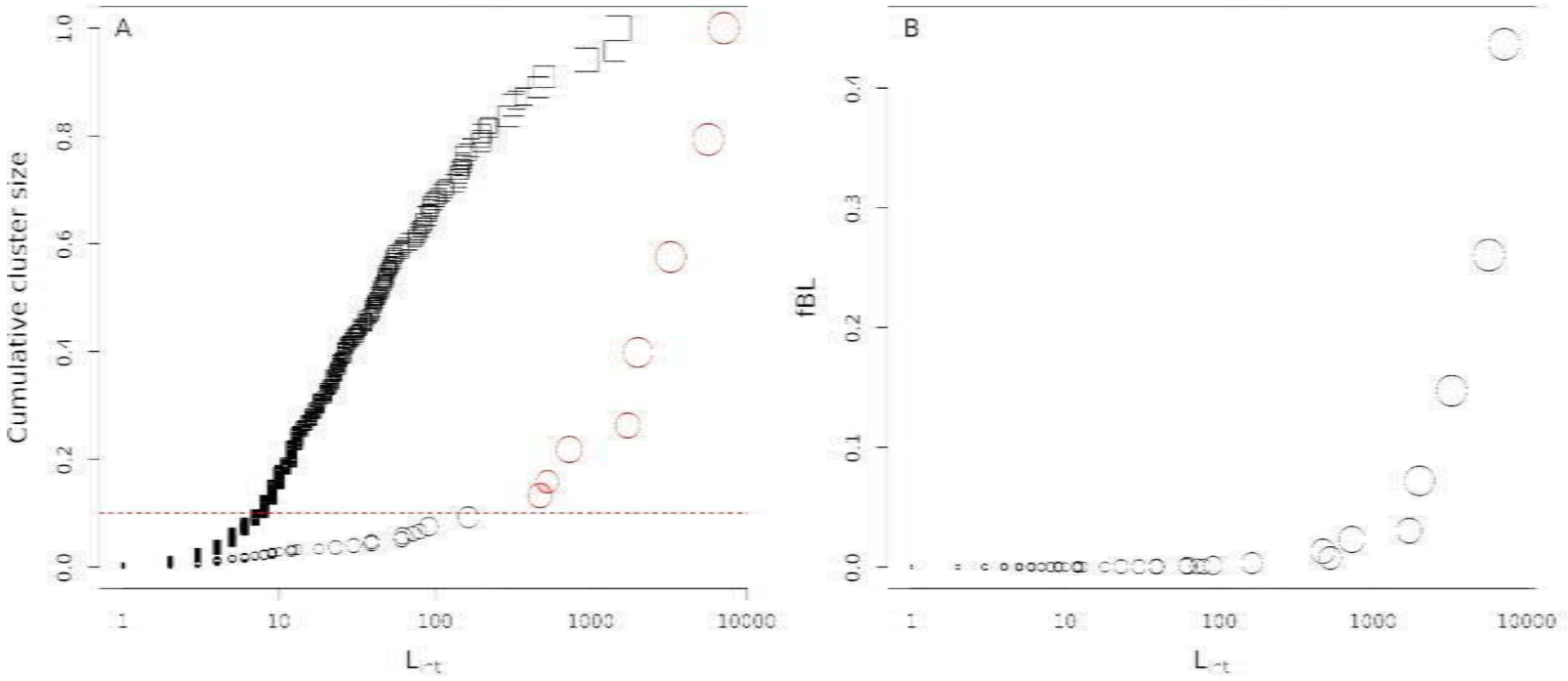


FIGURA 4.2: Análisis de las particiones obtenidas con los algoritmos *CNM* (circulos rojos) e *Infomap* (cuadrados negros). (A) Función acmulativa del tamaño de comunas (*Cumulativeclustersize*) para estructuras con número de links menor o igual a un dado l_{int}. El tamaño de cada símbolo es proporcional al logaritmo del tamaño del módulo que representa. Las líneas discontinuas verticales delimitan la región ($\sqrt{L/2}$, $\sqrt{2L}$), siendo L el número total de conexiones de la red. Esta región resulta una escala natural de operación en algoritmos basados en optimización de *modularidad* [86, 90]. (B). Número de conexiones internas a cada comuna *CNM* que no son internas a ninguna comuna *Infomap*, expresado como fracción del número total de conexiones de la red (fBL), en términos del l_{int} para cada comuna *CNM*.

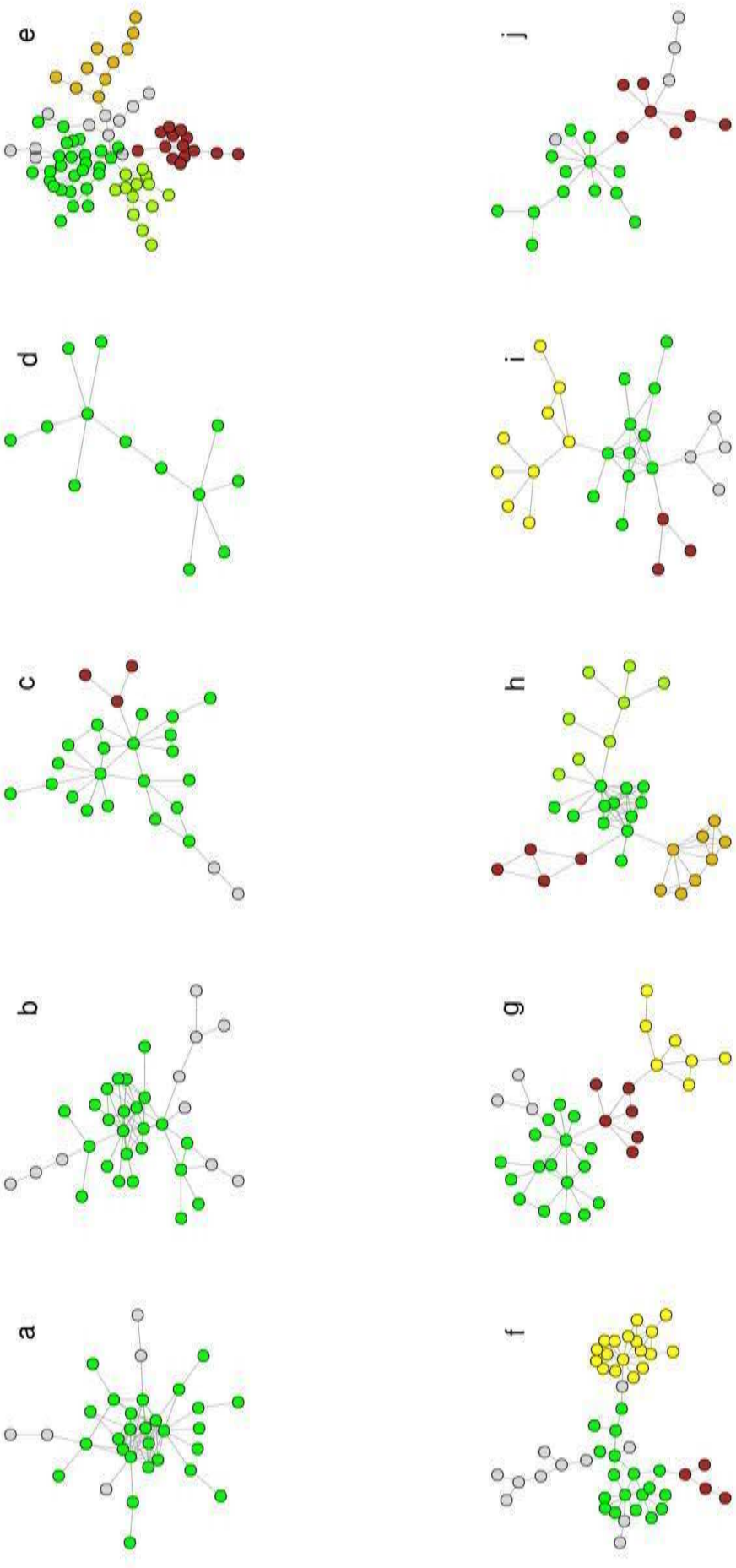


FIGURA 4.3: Módulos *CNM* con baja densidad de conexiones y un mínimo de 10 nodos. Las correspondientes comunas *Infomap* se representan con códigos de colores embevidos en el gráfico. Los nodos grises representan proteínas que en la descripción *Infomap* se unen a nodos de otros módulos.

tamaño superior a 10 nodos. En esta figura, los módulos *Infomap* fueron representados usando diferentes colores. Nodos de color gris corresponden a comunas *Infomap* que no están totalmente incluidas en la estructura *CNM* que se presenta. En esta figura, pueden reconocerse dos escenarios bien diferenciados: para los casos (a)-(d) se observa un buen acuerdo entre las descripciones modulares alternativas . Para los casos ilustrados en los paneles (e)-(j) la estructura interna existente en esos casos no logra resolverse mediante el uso de *CNM*, lo cuál se pone de manifiesto utilizando la descripción provista por *Infomap*.

En resumen, los resultados aquí presentados son consistentes con un escenario donde las estructuras *Infomap* presentan mayor resolución que las halladas por *CNM* y se fusionan en esta última descripción. En este sentido, ambas particiones de la PIN considerada son compatibles, pero la estructura de comunidades revelada por *Infomap* proporciona un nivel de granularidad más fino que el que puede obtenerse mediante el empleo de *CNM* .

4.5. Estructuras detectadas a alta resolución presentan mayor congruencia biológica

Una vez identificadas las diferentes descripciones modulares de *Infomap* y *CNM*, resulta esencial evaluar en qué medida estas particiones son consistentes con algún tipo de criterio biológico externo. Una metodología muy utilizada para medir el grado de congruencia biológica de un grupo de proteínas, es decir un *módulo* de la red, es comparar en que medida concuerdan sus anotaciones funcionales. Para ello, resulta especialmente útil disponer de una ontología que provea una forma sistematizada de organizar los distintos conceptos biológicos donde las proteínas pueden estar anotadas.

La ontología genética *Gene Ontology* [92] provee un vocabulario controlado utilizando tres grafos dirigidos y acíclicos (DAGs), es decir sin la presencia de *ciclos*, de manera que cada uno de ellos es capaz de organizar uno de tres tipos de conceptos biológicos distintos: *funciones moleculares* (MF), *componentes celulares*(CC), y *procesos biológicos* (BP). En particular en este capítulo haremos uso de la ontología correspondiente a procesos biológicos *BP*. Este DAG, es un grafo $G(C, \mathcal{E})$, donde cada nodo $c_i \in C$ representa un proceso biológico y cada arista $\overline{e}_{ij} \in \mathcal{E}$ representa una relación con dirección bien definida entre dos procesos biológicos c_i y c_j. El nodo c_i puede conectarse a otro nodo c_j, si el proceso biológico que representa c_i *es parte* del proceso c_j o bien *es un* tipo de proceso c_j. De esta manera, el consorcio de *Gene Ontology* (*GO*) mantiene un vocabulario controlado disponiendo de un identificador único para cada proceso biológico y sus relaciones con otros procesos claramente establecidas por las aristas del DAG.

Este tipo de ontología se utiliza en la práctica para mantener anotaciones funcionales de genes y proteínas. De esta forma, si dos proteínas se saben involucradas en un mismo proceso biológico encontraremos que las mismas se encuentran incluidas en el mismo nodo c_i del DAG provisto por *GO*. Es importante señalar que una misma proteína puede estar anotada en varios procesos biológicos c_i, $i = 1, 2...m$. Dada una proteína x denotaremos mediante $C_{(x)} = \{c_1, c_2, ...c_m\}$ al conjunto de m categorías (procesos biológicos) donde la proteína x se encuentra anotada.

Otro punto a considerar es que existen distintas estrategias para generar anotaciones de proteínas en los nodos del DAG y cada una de ellas tiene un nivel de fiabilidad diferente. El consorcio *GO* mantiene un detallado reporte de la evidencia experimental (*código de evidencia*) que da origen a cada anotación. Proveer un detalle exhaustivo de los distintos tipos de anotaciones en GO está fuera del objetivo de esta sección.

No obstante, cabe destacar que hay un único tipo de anotación funcional en *GO* que carece de curación manual y son aquellas denominadas *IEA* (*anotaciones inferidas electrónicamente*). Este tipo de anotaciones funcionales involucran por ejemplo comparaciones por similitud de secuencia, o anotaciones transferidas de bases de datos, las cuales carecen del una curación manual. Las anotaciones de tipo *IEA* se caracterizan tanto por su baja calidad, como por su gran cobertura (de hecho más del 40 % de las anotaciones en GO provienen de este tipo de evidencia).

En el contexto del análisis de congruencia biológica, hemos utilizado las anotaciones funcionales en GO para evaluar el grado de coherencia de las distintas *comunas* halladas pero sin considerar anotaciones funcionales con código de evidencia *IEA*. En particular, se ha considerado el índice de homogeneidad biológica BHI de una *partición*, introducido por Datta [93]. El observable BHI cuantifica el grado en que una partición dada de la PIN presenta grupos biológicamente homogéneos. Éste índice reporta, para cada comuna la máxima proporción de pares de genes agrupados que comparten una misma clase funcional de Gene Ontology [92].

Consideramos dos genes x, e y que pertenecen a un mismo módulo D en una partición de la PIN, que contiene un total de k módulos. Consideremos también todas las clases funcionales $C_{(x)}$ que contienen anotado al gen x y todas aquellas clases $C_{(y)}$, que contienen anotado al gen y. De esta forma, el índice de homogeneidad biológica de la partición queda definido según:

$$BHI = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j(n_j - 1)} \sum_{x \neq y \in D_j} I(C_{(x)} = C_{(y)}) \tag{4.1}$$

donde n_j es el tamaño o cardinalidad del cluster D_j, y la función indicadora $I(C_{(x)} = C_{(y)})$

toma el valor 1 si al menos una de las clases $C_{(x)}$ y $C_{(y)}$ coinciden. Como clases funcionales se consideraron todos los conceptos de Gene Ontology incluidos en la ontología de Procesos Biológicos, *BP*. La implementación de esta medida de homogeneidad biológica se llevó a cabo mediante el uso del paquete clValid R [94].

Los valores de BHI calculados para cada una de las 8 estructuras de mayor masa en la partición de *CNM* se representan como puntos rojos en la Figura 4.4, mientras que los triángulos verdes muestran valores de BHI para las respectivas particiones *Infomap* de cada uno de los 8 módulos *CNM* presentados. Notar que los grupos *CNM* fueron ordenados en el eje de abscisas según su masa en forma decreciente. En esta figura se aprecia que los niveles de BHI en particiones *Infomap* son sistemáticamente superiores a los observados en los correspondientes módulos *CNM*, lo cual sugiere que el mayor nivel de granularidad que proporciona *Infomap* resulta en un aumento significativo de la consistencia biológica global de la estructuras detectadas.

Además, para cada cada una de las 8 particiones *Infomap* presentadas, se consignan gráficos de cajas representando distribuciones de BHI estimadas en 1000 reasignaciones aleatorias de sus etiquetas. Este análisis permite asegurar que la ganancia en homogeneidad observada a través del incremento en BHI no proviene de efectos de tamaño exclusivamente, ya que en todos los casos, más del 95 % de las reasignaciones aleatorias de etiquetas (gráficos de cajas), presentan niveles BHI inferiores al valor indicado por la partición original de *Infomap*.

Estos resultados apoyan la idea de que los módulos *Infomap* identifican satisfactoriamente subestructuras de la PIN, con mayores niveles de congruencia biológica que las observadas con la metodología *CNM*.

4.6. Cartografía funcional a diferentes niveles de resolución

Los dos algoritmos de reconocimiento de comunidades analizados en este capítulo proporcionan descripciones modulares alternativas de la misma PIN, caracterizadas por diferentes patrones de conectividad intermodular e intramodular. Con el fin de obtener una mejor caracterización topológica de los nodos de la PIN, se consideran dos observables introducidos por Guimera y Amaral: la conectividad intramodular, Z , y el coeficiente de *participación*, P (ver sección 2.5). Estos observables toman en cuenta explícitamente la estructura modular de la red analizada. En la fig. 4.5.*a* se consigna la distribución de nodos en PIN sobre el plano Z-P, considerando la partición *CNM* . Las líneas discontinuas en la figura separan las regiones correspondientes a los 7 roles cartográficos universales propuestos por Guimera[7]. En la fig. 4.5.*b* se consigna también la distribución de roles cartográficos pero considerando la estructura modular provista

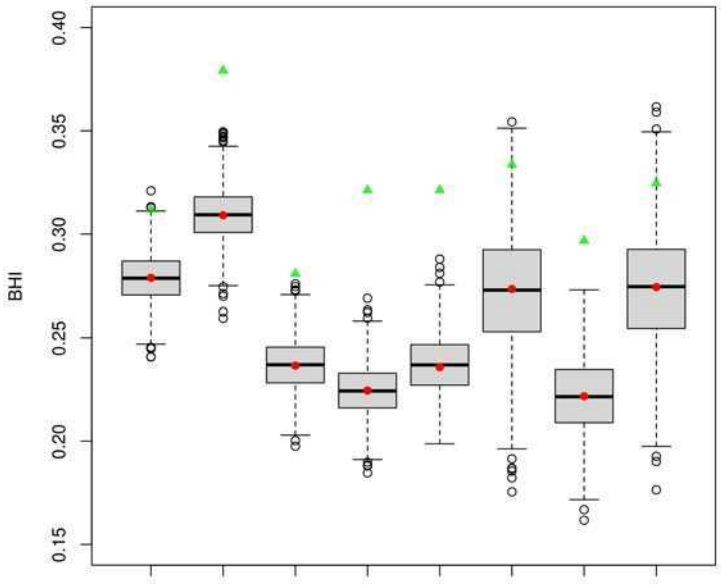


FIGURA 4.4: Indice de Homogeneidad Biológica, BHI, estimado para cada una de las 8 comunas *CNM* de mayor tamaño (puntos rojos), ordenados según el número de nodos de cada comuna. Los triángulos verdes denotan el correspondiente valor de BHI para las comunidades *Infomap* incluidas en el correspondiente módulo *CNM*. Para cada una de estas comunas *CNM*, los gráficos de cajas representan los valores de BHI calculados sobre un ensamble de 1000 particiones al azar, que respetan la distribución de tamaños provista por *Infomap*. Cabe destacar que, los valores medios de BHI en el ensamble aleatorio no difieren del nivel de homogeneidad biológica provisto por *CNM* en cada caso, y en contraste el respectivo valor BHI de *Infomap* está por encima del cuantil 0.95 de estos ensambles en todos los casos.

por *Infomap*. El esquema de colores consigna en cada caso una estimación de la distribución de densidad de puntos en este plano.

Comparando ambos paneles, se observa que los puntos no se distribuyen homogéneamente en el plano, sino que se dispersan en tres zonas locales de alta densidad: una zona de nodos ultra-periféricos (Z ~ -0.5 , P ~ 0), otra zona de nodos periféricos (Z ~ -0,5 , P ~ 0.5) y una zona de nodos conectores (Z ~ -0.5 , P ~ 0,65). Por otra parte, el menor nivel de resolución alcanzado por *CNM* da lugar a una tendencia general de asignar valores de *participación* inferiores a los asignados mediante el uso de *Infomap*. La tabla de contingencia 4.1 consigna la reasignación de roles entre una y otra descripción modular. En particular, el efecto de disminuir los valores de *participación* puede observarse en la aparición de numerosos genes Kinless (R4) según la descripción *Infomap*, (genes de bajo *grado intramodular*, Z<2.5 y alta *participación*, P>0.8) . De esta manera el 68 % de los nodos conectores *Infomap* fueron reasignados en roles

de *participación* más bajos (59 % periféricos , 9 % ultra-periféricos) considerando la partición *CNM* .

Infomap \ *CNM*	R1	R2	R3	R4	R5	R6	R7	Total
R1	3009	152	1	0	0	0	0	3162
R2	716	1497	84	0	18	32	0	2347
R3	146	936	500	0	7	21	0	1610
R4	9	304	288	11	4	19	0	635
R5	3	6	0	0	3	0	0	12
R6	1	44	5	0	18	34	0	102
R7	0	25	31	2	8	64	2	132
Total	3884	2964	909	13	58	170	2	8000

CUADRO 4.1: Distribución de las asignaciones en roles cartográficos según las descripciones *Infomap* y *CNM*. Abreviatura de roles cartográficos: ultra-peripheral (R1), peripheral (R2), connector (R3), kinless (R4), provincial Hubs (R5), connector Hubs (R6), Kinless Hubs (R7). El nivel de resolución provisto por el algoritmo *CNM* da lugar a una tendencia general de reducir el nivel de *participación* de los nodos de la red. Por ejemplo el 68 % de los vértices en la categoría *connector Infomap* reducen el nivel de *participación* reasignándose en la descripción *CNM* como nodos en categorías *peripheral* y *ultra-peripheral* (59 % y 9 % respectivamente). Más aún, la mayor parte de nodos *kinless Infomap* (94 % de los 635 nodos) fueron reclasificados como: *CNM-connector* (45 %), *CNM-peripheral* (48 %), y los nodos *CNM-ultra-peripheral* (1 %). Por último, también se puede observar que los nodos originalmente asignados a roles de alto *grado intramodular* (Z≥ 2.5) en la descripción *Infomap*, no sólo fueron afectados por una disminución en su nivel de *participación*, sino que además casi el 50 % de ellos fue reasignado a categorías de bajo *grado intramodular* (Z¡2.5) en la prescripción *CNM*

Más aún el 94 % de los 635 nodos kinless *Infomap* fueron se reasignados según *CNM* como: conectores (45 %) , periféricos (48 %), e incluso nodos ultra-periféricos (1 %). Por último, también se observa que los nodos inicialmente asignados a roles de alto *grado intramodular* usando *Infomap*, (Z>2.5, Provincial Hubs, Connector Hubs, y Kinless Hubs) no sólo sufren una notable disminución en la *participación* , sino además, casi el 50 % de ellos también sufre una fuerte baja en el nivel de *grado intramodular* al considerar *CNM*.

Todos estos resultados son consistentes con el hecho de que *CNM* da lugar a estructuras modulares de mayor masa, presentando menos superficies intramodulares que en el caso de *Infomap*. En otras palabras, dentro de los módulos *CNM* de mayor masa pueden aparecer "superficies internas" si se considera la partición de *Infomap*, lo cual implica que arcos originalmente internos a estructuras *CNM* se transformen en arcos intermodulares. Las consecuencias de estas discrepancias, tiene origen en la distinta naturaleza de la función objetivo que cada uno de los algoritmos de reconocimiento de comunas utiliza. Sin embargo estas discrepancias no son generalmente discutidas en la literatura dedicada a este tipo de análisis cartográfico. De hecho, varios estudios recientes en diferentes contextos biológicos, utilizan metodologías basadas en

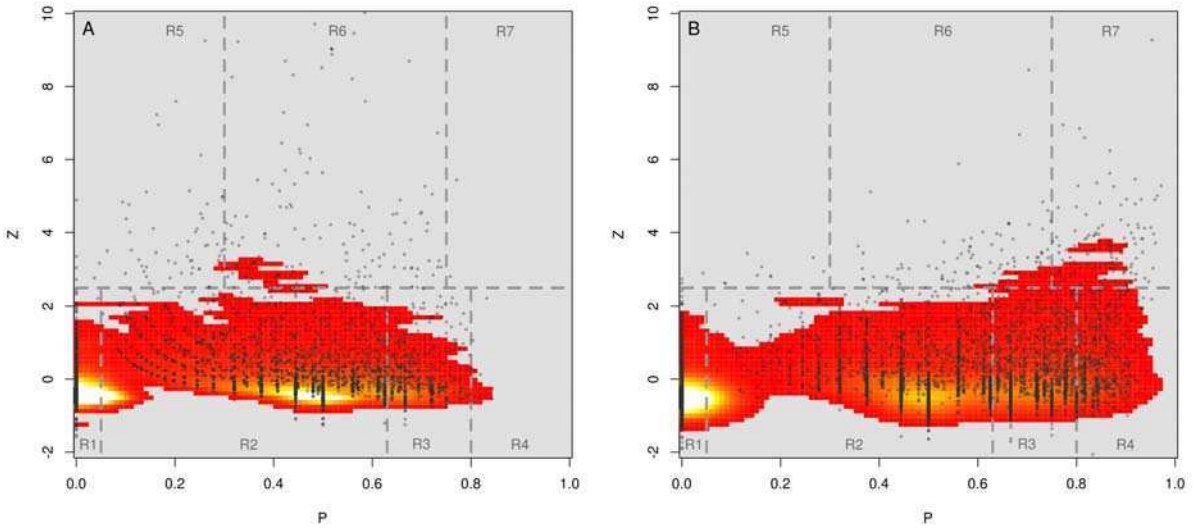


FIGURA 4.5: Distribución de nodos de la red HIPPIE en un plano Z-P, según la descripción *CNM* (A) e *Infomap* (B). Zonas blancas indican alta densidad de nodos mientras que rojo y gris prescriben zonas de baja densidad. Las líneas discontinuas delimitan regiones correspondientes a los 7 roles cartográficos descriptos en [7]

procedimientos de optimización de *modularidad* para caracterizar nodos PIN en términos de roles cartográficos [9–11]. En estos trabajos se reportaron por lo general un bajo número de nodos con alta *participación*. Sin embargo, los resultados expuestos en este capítulo hacen hincapié en que, la ausencia de nodos de alta *participación* en PINs, no es una característica intrínseca de este tipo de redes sino más bien una consecuencia de la metodología de reconocimiento de comunas empleada en todos estos casos . Si el algoritmo *Infomap* hubiera sido utilizado para la caracterización de esas redes, se habría observado un notable aumento en el número de nodos de alta *participación*, tal como explícitamente se consigna en la fig. 4.6

4.7. Análisis en conjuntos específicos de proteínas

En esta sección se pretende ilustrar algunos ejemplos motivados biológicamente, de cómo el índice de *participación* puede ser utilizado para revelar patrones de interacción característicos de proteínas asociadas adiferentes fenotipos complejos y funciones biológicas particulares. Se prestará especial atención a la utilización de diferentes metodologías de detección de comunidades con el fin de estudiar posibles sesgos en descripciones basadas en el uso de este índice. En particular, se focalizará en el uso de este tipo de caracterización cartográfica, para revelar patrones de conectividad en dos conjuntos fenotípicos de proteínas: uno relacionado a vías de

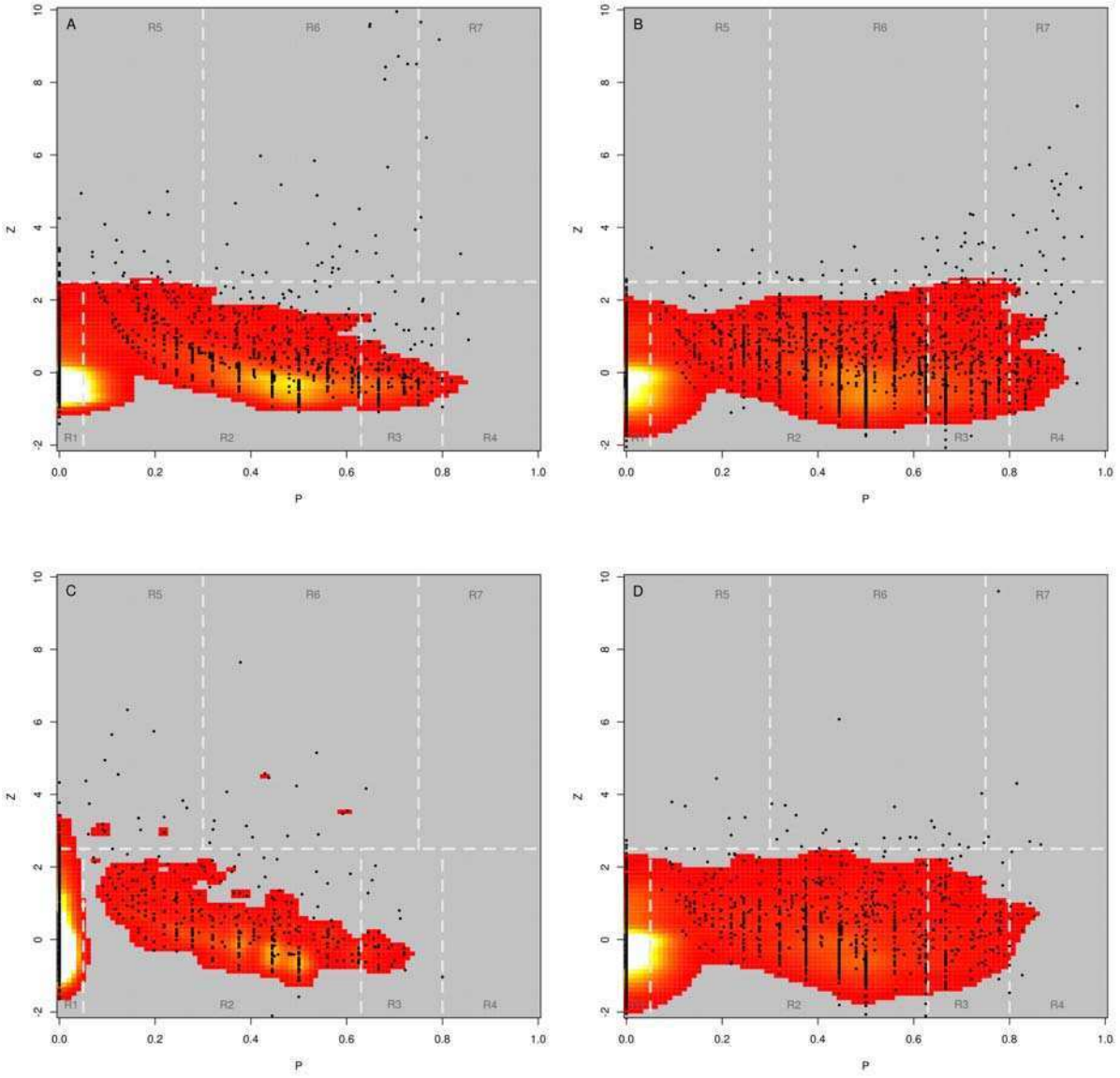


FIGURA 4.6: Distribución de densidad de nodos para dos redes PPI de Levaduras en un plano Z-P. Los paneles A y B corresponden a la PIN descripta en [82] para la descripción en comunas provista por *CNM* e *Infomap* respectivamente. Análogamente, los paneles C y D corresponden a la PIN descripta por [95] para *CNM* e *Infomap* respectivamente. Se observa en ambos casos un incremento general en los nivels de *participación* de los nodos de ambas redes cuando la partición de InfoMap es considerada.

señalización, y otro involucrado en procesos de envejecimiento humano. A continuación se describe con mayor detalle estos conjuntos de proteínas de interés.

4.7.1. Envejecimiento Celular

El proceso de envejecimiento celular (EC) está estrechamente asociado a enfermedades complejas, por lo que se ve afectado por diversos factores ambientales y genéticos [96]. Se ha dedicado un gran esfuerzo en la caracterización de la base genética del envejecimiento y como resultado se han identificado genes que: son capaces de modular el proceso de envejecimiento (por ejemplo, mutantes de genes que aumentan la esperanza de vida máxima en organismos modelo o vinculados a la longevidad humana) [97, 98] ,mostrar cambios transcripcionales que se correlacionan con la edad [99, 100], o mostrar patrones de metilación específicos del ADN [101].

Distintas metodologías basadas en la integración de redes moleculares han sido empleadas para proporcionar una comprensión a nivel sistémico del EC .[97, 102–104] . En particular , Xue y colaboradores consideraron un modelo de red modular de proteínas para estudiar procesos relacionados al EC [102] . En el mismo espíritu del presente trabajo, los autores analizaron la estructura modular de la red de interacción de proteínas relacionada a los procesos de EC considerados, encontrando que aquellas proteínas asociadas a procesos de envejecimiento celular (ARG) se encuentran irregularmente distribuidas en la red modular analizada. En ese trabajo, los autores reportaron que las interfaces entre módulos (laxamente definidos como vértices que presentan sus primeros vecinos ubicados en diferentes módulos) presentan una tasa de enriquecimiento entre dos y tres veces mayor en ARG que en las proteínas centrales de los distintos módulos.

En el contexto de este último hallazgo, el coeficiente de *participación* analizado en este trabajo resulta particularmente adecuado para proporcionar una mejor descripción topológica cuantitativa de las proteínas relacionadas a procesos de EC.

Para llevar a cabo este análisis, se considera un conjunto de genes asociados a envejecimiento celular sobre el organismo homo sapiens, extraído de la base de datos GenAge. Ésta, es una base de datos curada, que contiene genes asociados al fenotipo de envejecimiento humano [105]. Los datos, descargados en octubre 2013, constan de 298 genes de los cuales 261 codifican para alguna proteína en la red de interacciones aquí considerada (PIN).

Tal como se ha reportado en secciones precedentes, el uso de *Infomap* da lugar a un notable aumento general en los valores de *participación* de los vértices de la red con respecto a la caracterización basada en *CNM* (ver fig. 4.5 y Tabla 4.1) . No obstante, cabe destacar que el análisis de las curvas ROC a base de *participación*, calculadas para los genes ARG revela que esta tendencia está particularmente sesgada para genes ARG. Esto se desprende de la fig. 4.7 dado que el área bajo la curva ROC *Infomap* (AUC-InfoMap = 0,76) es mayor que la encerrada bajo la curva ROC de *CNM* (AUC-*CNM* = 0,65) y la diferencia resulta estadísticamente significativa (PV = 2,2 10-16, prueba de DeLong [106]). Este resultado sugiere que dado el mayor nivel de granularidad proporcionado por las comunidades *Infomap*, los genes ARG incrementaron selectivamente su grado de Participación respecto a otros genes de la red. Por lo tanto, en este último caso la componente de *participación* estimada usando la definición en comunas *Infomap* se presenta como un interesante observable para cuantificar y formalizar la tendencia informada en [102], de que ARG genes se ubican en las interfaces de las comunidades de la red.

Con el fin de extraer más información de la topología de la red de interacciones considerada, se ha explorado si los ARG genes presentan sesgos particulares en alguno de los distintos roles cartográficos de la red. Los resultados de enriquecimiento (prueba exacta de Fisher) presentados en la Tabla 4.2 (columna ARG - set) muestran que los roles provincial-hub y conector-hub presentan un fuerte enriquecimiento en ARG cuando se adopta el nivel de resolución proporcionado por la metodología de agrupamiento *CNM* . Por otro lado, las categorías de alta *participación*, Kinless y Kinless-hub, resultan altamente enriquecidas cuando se considera el nivel de resolución proporcionado por *Infomap*. En virtud de este último análisis alternativo, es interesante notar que bajo la descripción de *Infomap*, hubo un 64 % más de genes ARG, involucrados en categorías cartográficas enriquecidas, y además, los niveles de señal de significancia estadística alcanzados son más extremos, sugiriendo que el nivel de resolución proporcionado por *Infomap* permite realzar mejor el sesgo hacia roles de alta *participación* que presentan los genes ARG.

Aún cuando el coeficiente de *participación* proporciona información valiosa sobre las características topológicas de las redes de interacción, vale la pena analizar si tendencias similares podrían establecerse por sólo estudiar otras características topológicas, en particular, el grado de los nodos de la red. Este problema es especialmente relevante aquí, ya que el conjunto de genes ARG muestra elevados niveles de grado en la PIN (ver fig. 4.9), por lo que el grado puede resultar en un posible factor de confusión para nuestro análisis.

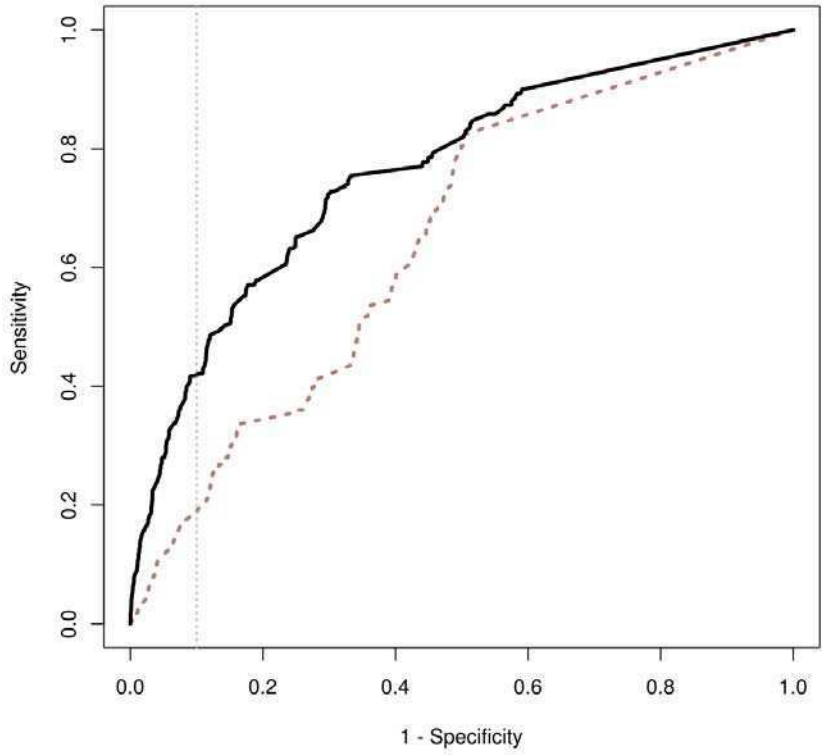


FIGURA 4.7: Participación como descriptor topológico en genes involucrados en envejecimiento (ARG) calculada para la prescripción en comunas *Infomap* (línea continua negra) y *CNM* (curva discontinua marron). La línea punteada vertical indica el nivel de 90 % de especificidad. El área bajo estas curvas (AUC) pueden utilizarse como medida de calidad de los descriptores topológicos utilizados, para el conjunto ARG estudiado. Las diferencias reportadas (AUC-*Infomap*=0.76, AUC-*CNM*=0.65) resultan estadísticamente significativas (DeLong Test, $pv \leq 10^{-16}$), sugiriendo que el nivel de resolución provisto por *Infomap* mejora la detección del sesgo topológico que presentan los genes ARG hacia altos niveles de *participación*.

Con el fin de deconvolucionar la señal medida de *participación*, se utilizó un análisis de muestreo estadístico realizando la misma prueba de Fisher para 1000 conjuntos aleatorios de proteínas tomadas al azar (RG) que respetan la misma distribución de grado del conjunto original de proteínas estudiado (ARG). Sobre cada uno de estos 1000 conjuntos RG se realizó la misma prueba exacta de Fisher que con el ARG original y se estimó un valor de significancia estadística (p-valor) en términos de la fracción de realizaciones aleatorias que presentan los mismos o mayores efectos (sobrerepresentación o subrepresentación) que los observados en los datos originales ARG.

Cabe señalar, que para construir los distintos conjuntos RG, se definieron previamente distintos “pools”de genes categorizados según su grado. A fin de garantizar que el muestreo este libre de sesgo, cada pool de genes fue generado garantizando un tamaño mínimo de muestra

Cartografía				ARG dataset		ARG' dataset		
.		Role	N	ARG	pv	ARG'	pv	pv'
CNM	Non Hubs	R1	3884	47	1	47	1	1
		R2	2964	119	0.01	119	4.60E-005	0.49
		R3	909	47	0	46	0	0.12
		R4	13	1	1	1	0.45	0.11
	Hubs	R5	58	15	**1.72E-009**	8	0	0.42
		R6	170	32	**2.19E-015**	12	0.01	1
		R7	2	0	1	0	1	1
Infomap	Non Hubs	R1	3162	26	1	26	1	1
		R2	2347	51	1	51	1.00E+000	1
		R3	1610	65	0.13	65	0.06	0.29
		R4	635	67	**4.95E-018**	67	**6.67E-021**	$< 10^{-4}$
	Hubs	R5	12	1	9.86E-001	1	0.42	0.62
		R6	102	11	2.32E-003	7	5.07E-002	0.98
		R7	132	40	**1.00E-027**	16	**4.37E-006**	0.35

Cuadro 4.2: Resumen de resultados obtenidos para evaluar la asociación estadística entre el conjunto ARG y los distintos roles cartográficos. Se consideró la descripción provista por *CNM* (primeras 7 filas) y la correspondiente a *Infomap* (últimas 7 filas). Las respectivas descripciones modulares se muestran en las 4 primeras columnas (Cartografía), donde N representa el número de proteínas anotadas a cada rol cartográfico.. Las columnas de *ARG-dataset* consignan el número de proteínas de ARG en cada rol (columna ARG) y el resultado de enriquecimiento mediante una prueba de Fisher (columna *pv*). Las columnas de *ARG'-dataset* consignan los resultados de enriquecimiento realizados para el subconunto ARG' donde se han removido el 10 % de proteínas de mayor grado. La columna ARG' consigna el número de proteínas por categoría cartográfica, la columna *pv* consigna el resultado de la correspondiente prueba de fisher. La columna *pv'* consigna el resultado de una prueba de remuestreo con un ensamble de 1000 conjuntos control que respetan la misma distribución de grado que el conjunto ARG'. Todos los p-valores realizados mediante la prueba de Fisher fueron corregidos por pruebas múltiples mediante False Discovery Rate (FDR adjustment). Notablemente, la única categoría que permite establecer asociaciones estadísticas significativas independientemente de la distribución de grado del conjunto ARG' es la Kinless *Infomap* (R4, pv'$< 10^{-4}$). Abrevituras de roles Cartográficos: Ultra-Peripheral (R1), Peripheral (R2), Connector (R3), Kinless (R4), Provincial Hubs (R5), conector Hubs (R6), Kinless Hubs (R7).

de 100 nodos por pool. Finalmente, cada realización aleatoria se construye extrayendo genes al azar del pool de genes correspondiente, de manera de respetar el grado que presenta el conjunto de genes originales. Una vez generados los distintos conjuntos RG, se realizó un análisis a posteriori para asegurar que las distribuciones de grado de los conjuntos control posean distribución de grado con características estadísticas similares que los observados en el grupo ARG.

Para asegurar que la distribución de grado de genes ARG esté debidamente muestreada evaluamos a cada cuantil de la misma, si el valor de grado correspondiente es comparable con el que se obtiene al mismo cuantil en las realizaciones aleatorias. Este análisis se muestra en la figura 4.8, que consigna el casode ARG (círculos azules), y los correspondientes conjuntos control (gráficos de caja). El eje de abscisas representa el cuantil de grado considerado en cualquiera de las distribuciones observadas (tanto en ARG, como en los RG), mientras que el eje

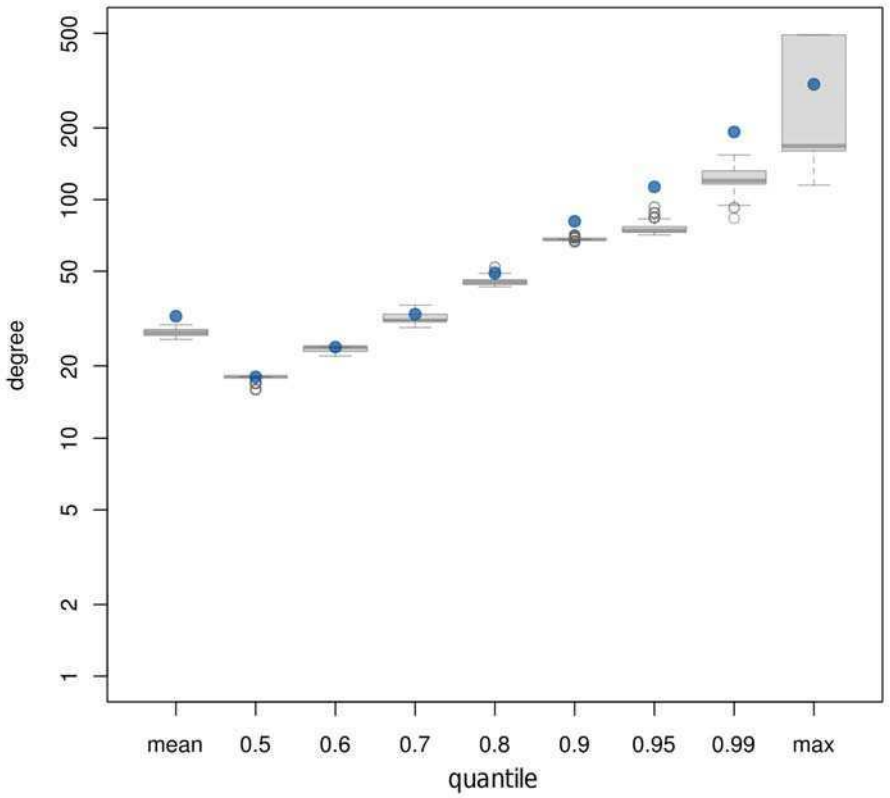


FIGURA 4.8: Control del muestreo estadístico basado en grado para genes asociados a envejecimiento. Cada grafico de cajas representa la distribución de grado para el cuantil seleccionado (eje de abscisas) en las 1000 muestras aleatorias del ensamble considerado. Los puntos azules denotan el grado del cuantil correspondiente para el conjunto original de genes ARG. Se puede observar que el 10 % de genes ARG de mayor nivel de grado, queda fuera de los niveles inter-cuartil de las muestras control correspondientes, indicando que los mismos no fueron debidamente representados.

de ordenadas consigna el valor de grado correspondiente a dicho cuantil. Se observa que, para niveles de grado por encima del cuantil 0.9 de la distribución, los niveles de grado del conjunto de ARG no puenen ser correctamente reproducidos por el procedimiento de muestreo aleatorio. En ese caso, se realizó el análisis de remuestreo considerando entonces un conjunto reducido ARG′ , que descarta el 10 % de genes de mayor grado que no pudo ser debidamente muestreado.

En este conjunto reducido ARG‘, encontramos que sólo la categoría Kinless bajo la prescripción *Infomap*, resulta enriquecida significativamente bajo el análisis de remuestreo (tabla 4.2, columna ARG‘ -genes). Estos resultados muestran que el nivel de resolución provisto por *Infomap* pone de manifiesto un enriquecimiento no trivial en genes ARG‘ en una única categoría cartográfica, el cual no puede ser explicado por efectos de la distribución de grado del

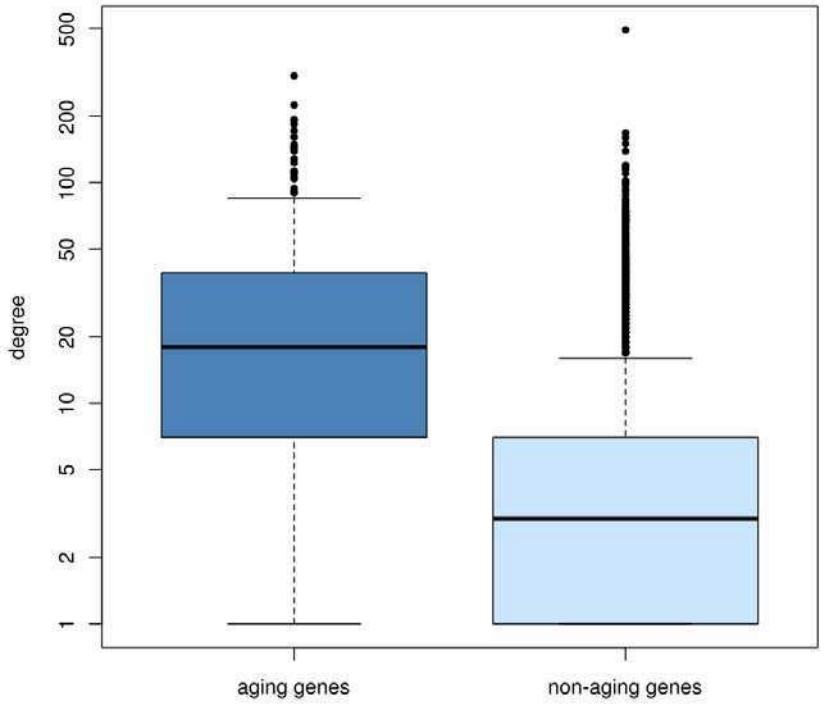


FIGURA 4.9: Distribución de grado para genes involucrados en envejecimiento celular (*aginggenes*, izquierda), y los restantes genes de la PIN (*nonaginggenes*, derecha). Se observan diferencias significativas en las correspondientes distribuciones de grados (Wilcoxon test, $p - val < 10^{-16}$).

correspondiente conjunto de genes.

Una vez establecida la relevancia de considerar la *participación* bajo la prescripción *Infomap* para el análisis de este conjunto de proteínas, se examinará si existen otras características topológicas de la red que puedan aportar evidencias no triviales para caracterizar el conjunto de genes ARG. En particular, se analizó la capacidad de diversos indicadores topológicos para realzar proteínas ARG de media o baja conectividad, es decir, aquellas proteínas de baja importancia desde la perspectiva del número de conexiones en la PIN. Con este fin, se consideró un subconjunto de la PIN que excluye el 10 % de los genes de mayor grado (esto implica remover del análisis todo gen de grado k > 18), y se compararon distintos observables topológicos como:

- la *participación* según la prescripción *Infomap*
- la *participación* según la prescripción *CNM*
- dos medidas alternativas de flujo de información: Betweenness y Bridge-centrality

- el grado de cada nodo.

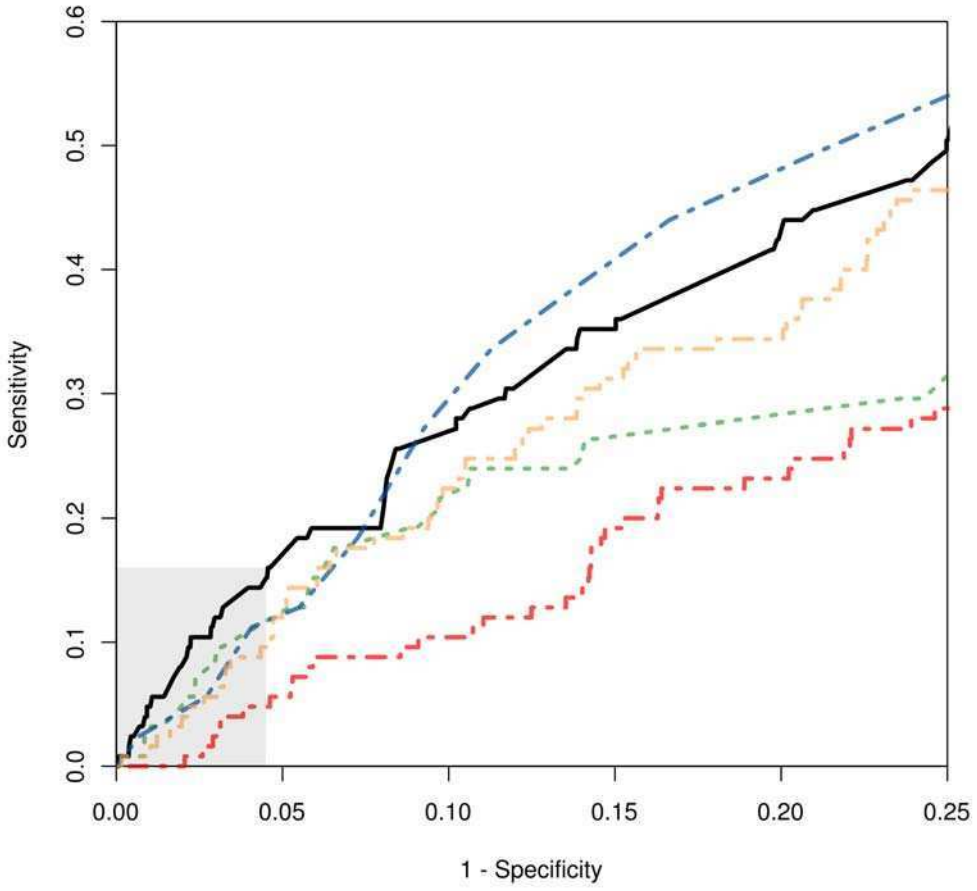


FIGURA 4.10: Descriptores topológicos para genes de bajo grado involucrados en envejecimiento(*ARG′*) para: Participación *Infomap* (línea negra continua), Participación *CNM* (línea naranja discontínua),grado (línea azul discontinua), bridge centrálity (línea roja discontinua) y betweenness centrality (línea verde discontinua). Sólo nodos con nivel de conectividad bajo o medio fueron considerados (i.e. nodos incluidos en el cuantil 90 % de toda la distribución de grado de la red, $k \leq 18$). El área sombreada denota el máximo nivel de especificidad y sensibilidad alcanzado por genes *Infomap* kinless en la detección de genes ARG.

La figura 4.10 muestra las curvas ROC obtenidas para el subgrafo PIN considerado y el correspondiente conjunto de genes ARG, limitado en grado también. El rango de 1-especificidad abarcado por *Infomap* Kinless está comprendido en el intervalo [0, 0.045] con niveles de sensibilidad (ordenadas) el valor de *participación Infomap* presenta la mayor sensibilidad entre los descriptores considerados. Independientemente de su rendimiento absoluto como predictor topológico, el comportamiento observado para la *participación* prescripta por *Infomap*, mejora tanto el alcanzado por la *participación* prescripta por *CNM* como cualquiera de las restantes

variables topológicas consideradas. Este resultado sugiere que la Participación *Infomap* proporciona la alternativa topológica más eficaz, entre las aquí consideradas, para caracterizar este subconjunto de genes considerados en el contexto de la red de interacción de proteínas utilizada. En particular, resulta más eficaz que otras cantidades relacionadas a la transmisión de flujo de información en la red como ser el Betweenness y el Bridge Centrality.

4.7.2. Proteínas involucradas en vías de señalización: base de datos SignaLink

Las vías de señalización pueden ser conceptualizadas como subredes de interacción de componentes moleculares capaces de realizar tareas de procesamiento de información complejas y altamente reguladas dentro de la célula. En la presente sección se estudiará si los patrones de interacción embebidos en PINs como HIPPIE, pueden servir para revelar principios de organización global relevantes en vías de señalización. Con el fin de probar esta hipótesis, se empleará un conjunto de componentes moleculares de vías de señalización proporcionado por la base de datos SignaLink (http://signalink.org/) [107]. La estructura multicapa de la base de datos SignaLink consta de componentes centrales de vías de señalización, sus reguladores (e.j. proteínas de andamiaje y endocitosis), sus enzimas modificadoras (e.j, fosfatasas o ligasas de ubiquitina), así como reguladores transcripcionales y postranscripcionales de todos estos elementos. En este capítulo se ha trabajado con la base de datos SignaLink 2.0 completa (descargada al 12 de octubre de 2013), que contiene información sobre 7 vías de desarrollo:

- RKT (receptor tirosina kinasa) cuya vía se caracteriza por pertenecer a la familia de receptores con actividad encimática intrínseca o asociada, teniendo como ligandos la insulina y diversos factores de crecimiento. Esta vía incluye además factores de crecimiento insulínicos (IGF), epidérmicos (EGF), y vías MAPK las cuales comunican la señal desde receptores como el RKT en la superficie celular hacia el ADN en el nucleo celular.

- TGF – ß, involucrado en procesos celulares tanto en organismos adultos como en el desarrollo embrional, incluyendo crecimiento celular, diferenciación celular, apoptosis homeostasis celular y otras funciones.

- Wingless/ Wnt, un grupo de vías de trasducción de señales formadas por proteínas que transfieren las señales del exterior de una célula a través de la superficie receptora de la misma hasta su interior.

- Hedgehog, una vía que transmite información hacia células embriónicas requeridas para un apropiado desarrollo.

- JAK / STAT, una vía que transmite información de señales químicas en el exterior de la célula a través de la membrana celular.
- NOTCH, una proteína transmembranal que sirve como receptor de señales extracelulares y que participa en varias rutas de señalización durante el desarrollo.
- NHR, receptores de hormona nuclear, los cuales forman ligandos que al ligarse a secuencias específicas de ADN actúan como interruptores para la transcripción dentro del nucleo celular. Estos interruptores controlan procesos de desarrollo y diferenciación de tejidos entre otros.

Luego de descargar la base de datos SignaLink y mapear los respectivos identificadores ENTREZ [108], se obtuvo información sobre los 2.814 productos génicos relacionados a estas vías de señalización. Finalmente, de este conjunto, 1.739 proteínas están presentes en PIN.

Desde una perspectiva sistémica, las vías de señalización celular pueden describirse como la composición de un módulo sensor de señales, otro módulo transductor, una vía de transmisión molecular (a lo largo de la cual pueden tener lugar complejos procesamientos de información), y una sección final de salida, responsable de la activación de moléculas efectoras. Además, para que estas vías de señalización alcancen un nivel de funcionalidad específico, distintos procesos de retroalimentación y regulación externa pueden desempeñar un papel relevante.

Para llevar a cabo el análisis se consideraron tres criterios de clasificación mutuamente no excluyentes, proporcionados por SignaLink. Estos criterios de clasificación están basados en: el rol que una proteína tiene en su respectiva vía de señalización (sección de la vía molecular), su importancia o esencialidad para el proceso de transmisión de la señal (proteínas centrales, o no centrales) , y la *participación* en una o más vías moleculares (proteínas de una o múltiples vías) [107]. Dado este esquema de clasificación de proteínas basado en vías de señalización, se analizará a continuación si puede o no establecerse conexiones estadísticamente significativas, entre estas categorías biológicamente motivadas y distintas variables topológicas de la PIN utilizada.

En particular, se evaluará si puede o no establecerse asociaciones significativas entre cualquiera de estas clases con roles cartográficos específicos. Los resultados de estos análisis de alto y bajo enriquecimiento entre categorías cartográficas y biológicas mediante el uso de una prueba exacta de Fisher, se resumen en la fig. 4.11. En esta figura la escala de colores logarítmica representa los p-valores obtenidos en cada prueba de Fisher, ajustados por la realización de pruebas múltiples vía False Discovery Rate (FDR)[109]. Resultados para las descripciones modulares

de *CNM* e *Infomap* se presentan en los paneles izquierdo y derecho respectivamente.
Además , resulta relevante analizar si las asociaciones significativas son apoyadas principalmente por su patrón de conexiones intramodular e intermodular, más que por sesgos en otras características topológicas propias del conjunto de proteínas analizado (distribución de grado de los nodos analizados por ejemplo). Por lo tanto, para cada condición de prueba se llevó a cabo un control estadístico como el realizado en la sección anterior con genes ARG (ver capítulo 4.7.1, y figura 4.8).

Así, en lo subsiguiente se considerará como asociaciones no triviales, aquellas que resulten significativas en virtud de este procedimiento de bootstrapping. En la figura 4.11, los símbolos "asteriscos", indican las categorías que presentan asociaciones no triviales (p-valor ajustado $FDR < 0.05$). En otras palabras, estos símbolos indican asociaciones estadísticamente significativas entre roles cartográficos y vías de señalización específicas, que no pueden ser explicadas por la distribución de grado del conjunto de proteínas analizado.
En los parrafos sucesivos se analizan las pruebas de enriquecimiento mostrdas por esta figura.

Proteínas que participan en múltiples vías de señalización

Según la clasificación SignaLink , proteínas "multipathway" o "cross- talk", son aquellas proteínas que participan en más de una vía de señalización. En la PIN empleada en este capítulo hay un total de 60 proteínas que pertenecen a este grupo, y 586 proteínas anotadas en una sola vía ("monopathway"). Los resultados de subrepresentación y sobrerepresentación, de estos grupos en los respectivos roles cartográficos, se presentan en las dos primeras filas de la fig. 4.11. Conforme a la descripción modular *CNM*, se observa que proteínas "multipathway" están asociadas con el rol de hub connector, que normalmente comprende nodos de alta *participación* y de alto grado. Asimismo se observa la escasa presencia de estas proteínas en el rol ultra-periférico (diametralmente opuesto, en el sentido que, está compuesto principalmente por nodos de baja *participación* y bajo grado).

Si este mismo análisis se realiza bajo la perspectiva de la descripción proporcionada por *Infomap*, se observa también una disminución general de este tipo de proteínas en roles con bajo nivel de *participación* (ultra-periféricos y periféricos). Por otra parte, en este caso, los dos roles asociados a niveles de alta *participación* (kinless y kinless-hub) se encuentran fuertemente enriquecidos en esta clase de proteínas "cross-talk". Más aún, la categoría Kinless en la descripción *Infomap*, es la única que presenta una sobrerrepresentación no trivial con proteínas "cross talk", es decir, una asociación que permanece significativa en la prueba de bootstrapping y por tanto

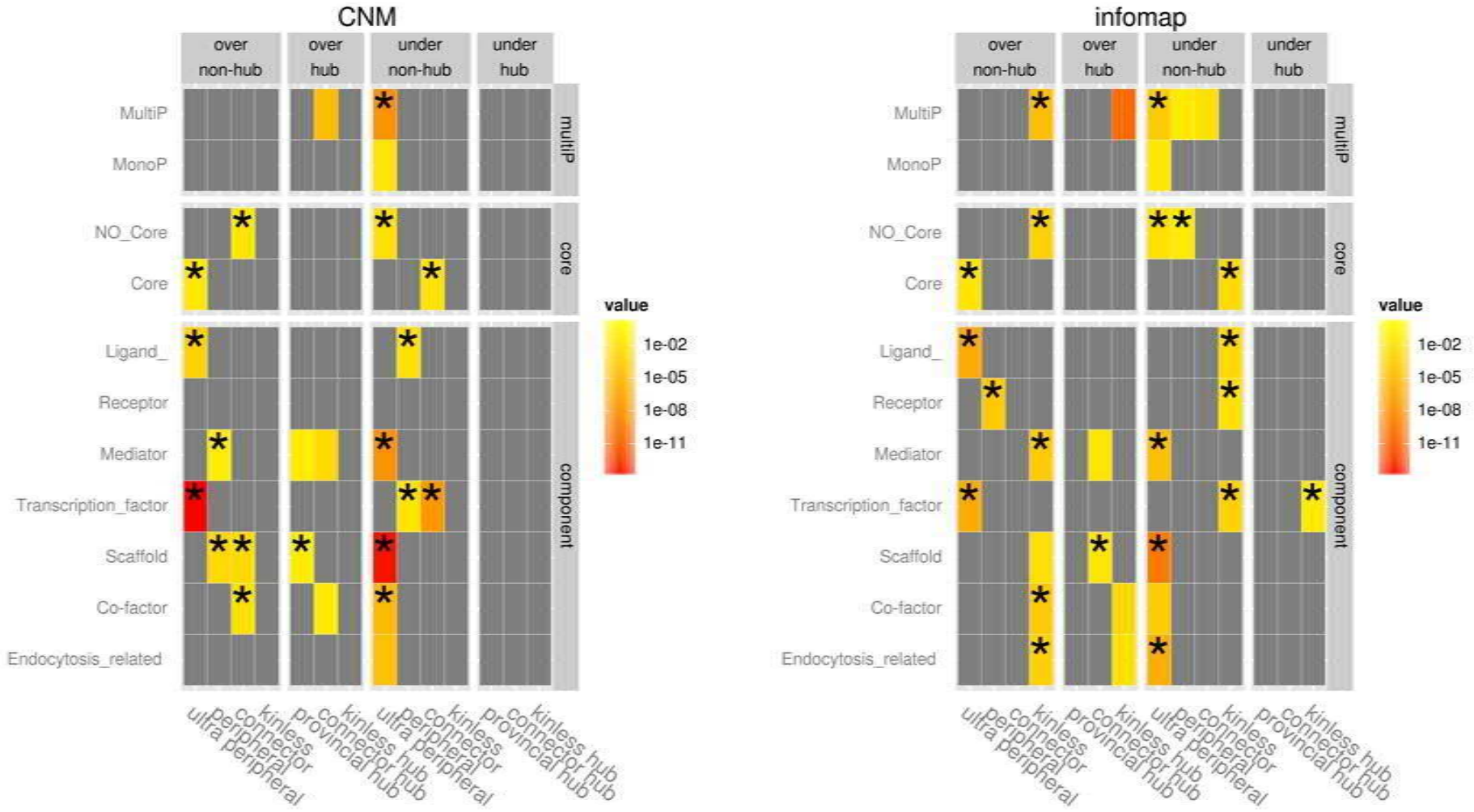


FIGURA 4.11: Sobrerrepresentación y subrepresentación de roles cartográficos definidos acorde a *CNM* (panel izquierdo), e *Infomap* (panel derecho) para categorías SignaLink. El código de colores representa el resultado de una prueba de Fisher exacta para éstas categorías dicotómicas, indicando el correspondiente nivel de significancia estadística que posee (pv ¡0.05). Celdas en color gris no resultan en asociaciones estadísitcamente significativas a un nivel de confianza de 0.05. Símbolos asteriscos representan asociaciones no triviales, es decir aquellas que no dependen de la distribución de grado del conjunto específico de genes estudiado en cada caso.

no puede ser explicada simplemente por la distribución de grado de estas proteínas. En contraste, el enriquecimiento de proteínas "cross talk" observado en el rol de hub-connector (tanto en *CNM* como en *Infomap*) presenta una dependencia con la distribución de grado de los nodos involucrados que no puede ser descartada.

Proteínas centrales

Según la definición de SignaLink, se dice que una proteína es central, si su presencia es esencial para transmitir la señal en al menos una de las vías de señalización a las que pertenece, y tiene por lo menos una de las características bioquímicas de la dicha vía (por ejemplo tirosina actividad kinasa). En contraste, una proteína se denomina "no central" cuando modula a las proteínas centrales [107].

La PIN aquí empleada consta de 359 proteínas centrales y 287 no centrales según la definición de SignaLink. Se puede apreciar a partir de la tercera y cuarta filas de la figura 4.11 que para esta clasificación de proteínas, se encuentran asociaciones estadísticas cualitativamente similares (respecto a los roles cartográficos involucrados), independientemente del nivel de resolución empleado para revelar la estructura modular de la red. En ambos casos, las proteínas centrales se ven fuertemente asociadas con el rol ultra-periférico, mientras que las proteínas no centrales se vinculan a roles de alta *participación* (*Infomap*-Kinless y *CNM*- conectores). Cabe destacar que todas estas asociaciones resultan no triviales, en el sentido que todas ellas presentan independencia estadística respecto a su distribución de grado.

Secciones en vías de señalización (SVS)

Como última clasificación motivada biológicamente para evaluar posibles sesgos en la distribución de roles cartográficos se consideraron distintas secciones de vías de señalización (SVS), acorde a la definición proporcionada en la base de datos SignaLink. Se evaluaron los siguientes grupos:

- Ligandos: proteínas que inician la señal de una vía.
- Receptores: proteínas receptoras inmediatas de la señal.
- Mediadores: proteínas que median la señal de los receptores hacia los factores de transcripción.
- Co-factores: proteínas que modulan la función de cualquier otra proteína de señalización.

- Factores de transcripción: proteínas que se unen a una región promotora específica en el ADN, o que transmiten la señal recibida a hacia otros factores de transcripción o forman complejos con estos.

También se consideraron dos categorías complementarias: proteínas involucradas en endocitosis y proteínas de andamiaje. Vale la pena mencionar, que de acuerdo con la política interna de anotación de SignaLink, las proteínas pueden estar asociadas a más de una y hasta un máximo de dos secciones [107] . Sin embargo en la última actualización descargada en Octubre 2013, se detectan 41 proteínas se están asociadas a más de dos secciones. En vistas a que la clasificación de proteínas en diferentes secciones de vías de señalización provista por SignaLink fue concebida en un escenario biológico de complejas relaciones entrelazadas que tienen lugar dentro de la célula, vale la pena examinar y relevar algunas de sus características específicas antes de proceder con los análisis de enriquecimiento de categorías.

		SVS Alternativa						
		Ligand	Receptor	Mediator	T.factor	Scaffold	Co-factor	Endocytosis
SVS	Ligand	78	0	0	1	3	6	4
	Receptor	0	81	1	6	13	3	10
	Mediator	0	1	92	17	47	27	11
	T. factor	1	6	17	820	20	15	2
	Scaffold	3	13	47	20	250	78	47
	Co-factor	6	3	27	15	78	130	34
	Endocytosis	4	10	11	2	47	34	33

CUADRO 4.3: Distribución de frecuencias conjunta en las distintas secciones de vías de señalización (*SVS*). Por ejemplo, de las 90 proteínas anotadas en la sección de Ligandos, 78 son exclusivas de esta sección, mientras que las 12 restantes muestran vías alternativas. Pese que la base de datos *SignaLink* reporta un máximo de 2 secciones por proteína se observa que 2 proteínas de la categoría Ligando están en realidad asignadas a tres secciones simultáneamente. En general se han identificado un total de 41 proteínas asociadas a más de dos vías.

En primer lugar puede apreciarse que, la mayor parte de ligandos , receptores y factores de transcripción pertenecen sólo a una sección (tablas 4.3 y 4.4) y son proteínas centrales (tabla 4.5). Por otro lado, las proteínas anotadas como mediadores así como aquellas que pertenecen a otras categorías no directamente relacionadas a vías de señalización sino a procesos regulatorios (tales como co-factores, proteínas de andamiaje y proteínas relacionadas a endocitosis) pertenecen generalmente a más de una sección y tienen simultáneamente un rol central en alguna de sus vías así como uno no central en otra u otras de sus vías (ver tablas 4.4 y 4.5). En general, encontramos que mientras pertenecer o no al núcleo de una vía depende en sí mismo de la vía en cuestión, la sección resulta una propiedad en principio independiente de la vía, para ligandos,

receptores y factores de transcripción cuando menos. En cuanto a los resultados expuestos en la figura 4.11, pueden observarse diferentes patrones de asociación, soportados estadísticamente entre funciones cartográficas y secciones de vías. Por ejemplo, se encuentra que las proteínas pertenecientes a ligandos y factores de transcripción (quinta y octava filas de la figura respectivamente) están fuertemente asociadas al rol ultra-periférico bajo cualquiera de los niveles de resolución empleados y sobrerepresentados en roles de alta *participación* (conectores *CNM*, y Kinless para el caso *Infomap*). Todas estas asociaciones son no triviales, es decir, permanecen significativas cuando fueron controladas por el efecto de distribución de grado correspondiente.

SVS	Total	SPS	PSS
Transcription factor	870	820	0.94
Ligand	90	78	0.87
Receptor	110	81	0.76
Scaffold	420	250	0.6
Mediator	180	92	0.52
Co-factor	260	130	0.5
Endocytosis related	110	33	0.29

CUADRO 4.4: Número total de proteínas asociadas a las distitnas secciones de vías de señalización consideradas (*SVS*). Número de proteínas anotadas en una úncia vía (SPS) y especificidad de cada vía (PSS) definida como el cociente entre SPS y el número total de proteínas reportadas en cada vía.

SVS	no-core	core	%core
Ligand	7	83	0.92
Receptor	8	95	0.92
Mediator	82	85	0.51
Transcription factor	27	92	0.77
Co-factor	207	32	0.13
Scaffold	81	57	0.41
Endocytosis related	31	24	0.44

CUADRO 4.5: Número de proteínas centrales y no centrales anotadas en cada sección de vía de señalización (*SVS*), reportadas en la primera y segunda columna respectivamente. La tercer columna representa la fracción de proteínas centrales en correspondiente a cada vía. Es importante recalcar que varias vías de señalización cuentan con proteínas sin especificar el carácter de central o no central.

Las proteínas que pertenecen a la SVS “receptor”, no se observan asociados a ninguna función cartográfica bajo la prescripción *CNM* . Por el contrario, el uso de grupos *Infomap* para representar la estructura de la red de interacción de proteínas , hace posible establecer una asociación no trivial entre esta SVS y nodos periféricos, así como notar que esta SVS se encuentra subrepresentada en roles de alta *participación*, tal como se observa para nodos Kinless.

En este punto, cabe señalar que el extenso grado de solapamiento que presentan las restantes SVS (ver tabla 4.3) puede dificultar el análisis de los respectivos patrones de enriquecimiento. Aproximadamente la mitad de mediadores, proteínas de andamiaje, y co-factores se encuentran involucradas también en una SVS alternativa, mientras que en proteínas de endocitosis esta cantidad llega hasta un nivel del 70 % (véase tabla 4.4). Aunque un análisis marginal exhaustivo está fuera del alcance de la presente tesis, algunos sesgos generales pueden ser resaltados para estas cuatro categorías del análisis de la Figura 4.11.

Por ejemplo, se observa que la SVS "mediadores" está asociada al rol de proteínas periféricas cuando se considera grupos *CNM*, y al rol Kinless cuando *Infomap* es utilizado. Independientemente del criterio de agrupamiento en comunas empleado, las proteínas mediadoras están subrepresentadas en la categoría ultra-periférica. Resulta interesante, que estos últimos resultados se mantienen incluso cuando el mismo análisis se realiza utilizando sólo las 92 proteínas que pertenecen exclusivamente a la SVS "mediadores" (datos no mostrados).

En relación a la SVS de proteínas de andamiaje, se observa una tendencia similar de subrepresentación en la categoría de baja *participación* ultra-periférica, y una asociación no trivial de sobrerepresentación en categorías de alta *participación* y alto grado, en cualquiera de los niveles de resolución adoptados para el agrupamiento en comunidades. Más aún, esta es la única SVS asociada a roles de alto *grado intramodular*: hubs provinciales para *CNM* y hubs conectores para *Infomap*.

Por último notamos que las SVS de proteínas co-factores y endocitosis, presentan una tendencia general a la subrepresentación en la categoría ultra-periférica (baja *participación*), y de sobrerepresentación en roles de mayor *participación*.

4.8. Discusión

En el presente análisis de estructura modular en una red de interacción de proteínas humana (PIN), encontramos que ambos algoritmos de agrupamiento considerados y ampliamente utilizados en la comunidad de redes, *Infomap* y *CNM*, producen particiones de la red con alta calidad en términos de los valores de *modularidad* alcanzados. Sin embargo, surgen importantes diferencias en cuanto a la granularidad de cada descripción. En particular, las diferencias observadas podrían tener origen en el efecto de límite de resolución [86] que afecta el rendimiento del algoritmo de detección de comunidades *CNM*. Más aún, se verifica que las estructuras dominantes (de mayor tamaño) detectadas por esta metodología se subdivididen en pequeños grupos

de acuerdo a la descripción modular provista por *Infomap*. Es importante destacar que este comportamiento descripto no es característico de la red PIN utilizada ni del organismo estudiado, dado que se observa el mismo comportamiento al analizar redes de interacción alternativas en Levaduras, oportunamente publicadas [82, 95], (fig.. 4.6). Este resultado, sin duda relativiza la afirmación original de Guimera, de la ausencia de nodos kinless en redes reales [7]. En lugar de ello, hemos mostrado que esto puede eventualmente surgir sólo como consecuencia de la metodología de detección de la Comunidad empleada, más que por una característica intrínseca de las redes analizadas.

Las discrepancias observadas en el nivel de resolución de comunas, tienen incidencia no sólo en la coherencia biológica de los grupos de proteínas hallados (las estructuras *Infomap* presentan mayores niveles de coherencia biológica que las *CNM*), sino también, en los patrones de conectividad que cada algoritmo es capaz de revelar para el análisis de proteínas consideradas. A modo de ilustrar este último punto, se presentan análisis de varios conjuntos de proteínas relacionados a vías de señalización y a procesos de envejecimiento celular.

En el caso de proteínas asociadas a envejecimiento celular (ARG) se observaron relaciones estadísticamente significativamente entre este grupo de proteínas y su pertenencia a roles cartográficos de alto *grado intramodular* tanto para un amplio rango en los niveles de *participación* (roles Provincial Hubs y Connector Hubs) cuando la partición *CNM* es considerada (Tabla 4.2). Sin embargo, si se considera la partición provista por *Infomap*, se encuentra que el mismo conjunto de proteínas está enriquecido significativamente en roles de alta *participación* (categorías Kinless y Kinless – hub).

Cabe en este punto mencionar, que en ninguna categoría Hub (tanto *CNM* como *Infomap*) se puede descartar que el enriquecimiento observado sea consecuencia de la distribución de grado particular que exhiben los correspondientes nodos ARG, ya que, utilizando muestras de proteínas tomadas al azar y que respetan la distribución de grado del conjunto ARG considerado, muestran asociaciones de similares niveles de significancia estadística que en el caso de ARG. Sólo la categoría Kinless *Infomap* (de bajo *grado intramodular* y alta *participación*), resulta no trivialmente asociado al conjunto ARG. Esto último implica que, la asociación establecida entre este grupo y la correspondiente categoría cartográfica Kinless, está principalmente apoyada en el patrón de conectividades intermodular e intramodular de los nodos de la red cuando los mismos no son definidos con el grado de resolución provisto por *Infomap*.

Todos estos resultados nos permiten hipotetizar, que los nodos ubicados principalmente en las interfaces de las comunidades detectadas al nivel de resolución provisto por *Infomap*

pueden servir para la coordinación y/o transmisión de flujo de información entre los módulos detectados con funcionalidades biológicas específicas.

Un ejemplo paradigmático de una proteína kinless *Infomap* relacionada con el EC es Sirtuin1(SIRT1). Esta proteína, y otros miembros de la familia sirtuin (SIRT3 y SIRT6) contribuyen a un envejecimiento saludable en mamíferos [110]. En particular, la asociación de SIRT1 con EC se ha propuesto sobre la base de su rol en varios procesos, tales como procesos de estabilidad genómica, de eficiencia metabólica, de biogénesis mitocondrial, proteostasis y respuestas inflamatorias relacionadas con el envejecimiento [110]. Las proteínas codificadas por el gen CDKN2A son otro ejemplo interesante de productos génicos pertenecientes a la categoría Kinless *Infomap*, relacionados a EC. Este gen da lugar a varias isoformas conocidas para funcionar como inhibidor de la CDK4 kinasa, tales como p16 y p19. Los niveles tanto de p16 y p19 se correlacionan con la edad cronológica de tejidos en seres humanos y modelos animales. Más interesante aún, el locus del gen CDKN2A se encontró asociado a varias de las enfermedades relacionadas con EC en un metaanálisis de GWAS [111]. En base a estas evidencias, el gen CDKN2A es considerado como el mejor gen documentado que controla el envejecimiento humano y está asociado a enfermedades relacionadas con el proceso de envejecimiento.

Respecto al estudio de proteínas pertenecientes a múltiples vías de señalización ("cross-talk"), se encuentra que hay varias categorías de alto *grado intramodular* enriquecidas. Cabe aquí también señalar, que el rol Kinless *Infomap* el único asociado a éstas proteínas con un enriquecimiento no trivial, es decir independiente de la distribución de grado de éstas proteínas.

Desde una perspectiva biológica, el reconocimiento de la existencia de proteínas "cross-talk", empleadas en la descripción de procesos fisiológicos moleculares implica que, participando en más de una vía de señalización, estas entidades moleculares podrían servir para propósitos de interconexión y coordinación entre módulos funcionales que de otra forma se encontrarían separados. El hecho de que las comunas *Infomap* se muestren asociadas a funcionalidades biológicas bien definidas, concuerda entonces con la asignación cartográfica de *Infomap* que se esperaría para este tipo de proteínas. El gen SIRT1 ya mencionado, es un miembro bien estudiado de esta categoría cross-talk en SignaLink. El mismo codifica una proteína histona NAD dependiente (nicotinamida adenina dinucleótido) que coordina distintos procesos, tales como el ciclo celular, la respuesta al daño en el ADN, el metabolismo, la apoptosis y la autofagia.

Para las proteínas mapeadas a la SVS receptor, fue la descripción *Infomap* nuevamente, la única que proporcionó evidencia suficiente de asociaciones significativas y no triviales con funciones cartográficas específicas. Sin embargo, esta vez fue el rol de nodos periféricos (caracterizado por relativamente bajo nivel de *participación*) el que se encuentra asociado al conjunto

de genes correspondientes. Esta tendencia es consistente con la fuerte (y no trivial) subrepresentación observada en el rol Kinless (de alta *participación*). Dada la tendencia general observada en *Infomap*, la cual típicamente aumenta los niveles de *participación* de los vértices de la red, la presentación de asociaciones significativas que implican roles de baja *participación* al nivel de resolución *Infomap*, y no bajo la descripción *CNM*, resulta en sí mismo un resultado destacable. Además, esto sugiere que *Infomap* detecta estructuras biológicamente sensatas, en el sentido de que las superficies adicionales que aporta, no impiden detectar el enriquecimiento de categorías de baja *participación*.

Vale la pena también hacer mención, que en algunos casos se ha encontrado ambas particiones en comunidades (*Infomap* y *CNM*) consistentes en los resultados de asociaciones significativas y no triviales. Por ejemplo, el análisis de proteínas *centrales* y *no centrales* resulta en una sobrerepresentación en categorías de baja *participación* en el primer caso, y de alta *participación* en el segundo caso ya sea bajo la prescripción en comunas *CNM* o *Infomap*.

Para ligandos y factores de transcripción los patrones de sobrerepresentación y subrepresentación detectados por ambos algoritmos también concuerdan. En este caso se observó una importante sobrerepresentación (y no trivial) en el rol ultraperiférico, mientras que roles de alta *participación* están subrepresentados. El patrón reportado apoya la idea de que los puntos de partida en las vía de señalización (ligandos y receptores) así como los de finalización de las mismas (factores de transcripción) se encuentran establecidos en la periferia de la red de interacción de proteínas, en concordancia con las observaciones de Csermely y colaboradores hechas en el contexto de búsqueda in-sílico de fármacos [2].

El mismo acuerdo cualitativo ocurre en las proteínas de andamiaje. A su vez, esta es la única SVS SignaLink que puede ser significativa y no trivialmente asociada a un rol de alto *grado intramodular* (provincial-hub y conector-hub para *CNM* e *Infomap* respectivamente). Las proteínas de andamiaje unen y colocalizan tres o más miembros de una vía, mientras que las proteínas adaptadoras unen y colocalizan dos miembros funcionales que interactúan en una vía catalítica [112, 113]. Por lo tanto, proteínas de andamiaje y adaptadoras actúan como plataformas de organización, reclutando diferentes componentes de una vía dada y sus parejas aguas arriba y aguas abajo, hacia una ubicación específica en la célula para lograr alguna función particular.

Las proteínas de andamiaje ofrecen muchas posibilidades en la regulación de vías y también resultan puntos claves en la comunicación entre vías que tengan lugar. Por ejemplo, permiten a Kinasas desempeñar diferentes roles, de acuerdo con el complejo de señalización en la que se ensamblan [113]. En este sentido, es razonable que las proteínas de andamiaje como PICK1,

PLCG1 y GRB2 fueran enriquecidas en roles con alto grado y alta *participación* (por ejemplo hub conector). En particular, el gen GRB2 codifica la *proteína 2 de unión al receptor de factores de crecimiento*, la cual es una proteína de andamiaje y adaptadora que conecta los receptores de factores de crecimiento de la superficie celular con la vía de señalización Ras, actuando como un importante mediador en esta vía. Además, también está implicado en otras vías, como la de señalización del receptor de insulina, o en vías relacionadas a procesos inmunológicos y de desarrollo. El gen PICK1 codifica una proteína andamiaje que contiene un dominio PDZ (dominio estructural común en varias proteínas de señalización) que sirve para reclutamiento de proteínas kinasas en regiones subcelulares. Este último interactúa con receptores de glutamato, transportadores de membrana plasmática de monoamina, con canales de sodio no activables por voltaje, y pueden dirigir la proteína kinasa PRKCA hacia estas proteínas de membrana, regulando su función y distribución.

Otra proteína de andamiaje interesante perteneciente al rol "hub conector" es la fosfoinosítido fosfolipasa C gamma 1, codificada por el gen PLCG1. Como otros miembros de las enzimas fosfolipasa C, la PLCG1 es un componente clave en vías de señalización reguladas por diversas señales extra celulares. Además esta proteína está implicada en respuestas celulares anormales asociadas a enfermedades inmunológicas [114].

Por último, también podemos observar una interesante correlación entre los roles cartográficos y la función biológica para conjunto de proteínas relacionados con endocitosis. Aquí, nuevamente, es la descripción *Infomap* la que proporciona asociaciones no triviales con roles de alta *participación*, particularmente, la categoría "Kinless". Un importante aspecto de la comunicación entre las vías es la localización de las correspondientes proteínas de señalización y de las proteínas de entocitosis que participan en endosomas. Estas proteínas fueron recientemente propuestas como agentes claves en la intercomunicación entre vías metabólicas endosomáticas [115].

4.9. Conclusiones

En este capítulo, se ha estudiado de manera sistemática cómo dos descripciones modulares alternativas de una red biológica (HIPPIE), obtenidas por diferentes algoritmos de detección de comunidades, pueden condicionar los resultados subsecuentes del análisis biológico en redes de interacción de proteínas. En particular, se analizaron dos algoritmos de reconocimiento de comunidades ampliamente conocidos y paradigmáticos, como *CNM* e *Infomap*, caracterizando

en profundidad el desempeño de los mismos en términos de la granularidad de las correspondientes particiones inferidas y de la homogeneidad biológica de las mismas. Hemos observado que la partición *Infomap* resulta en una descripción más apropiada de la estructura modular de la red de la prescripción *CNM*, y argumentamos que el límite de resolución inherente a los algoritmos basados en optimización de *modularidad* como *CNM*, puede ser una de los orígenes en las diferencias cualitativas de comportamiento. En nuestro conocimiento, el efecto de límite de resolución no ha sido evaluado con anterioridad en el contexto de una red biológica específica. Más aún, encontramos que las comunas detectadas por *Infomap*, no sólo resultan en estructuras congruentes desde el punto de vista topológico, sino también que muestran niveles más altos de homogeneidad biológica. Por otro lado, las discrepancias en el nivel de resolución que presenta cada algoritmo también incide en el tipo particular de patrones de conectividad de mesoescala, que cada metodología es capaz de revelar.

En este sentido, presentamos un detallado análisis de las diferencias que surgen en las asociaciones estadísticamente significativas que pueden ser establecidas entre los patrones de conectividad intermodular e intramodular con diversos conjuntos de proteínas específicas relacionadas a distintos fenotipos complejos y funcionalidades biológicas, tales como envejecimiento celular y diversas vías de señalización.

En líneas generales, los resultados aquí expuestos proporcionan una llamada de atención respecto a las herramientas técnicas empleadas en el análisis biológico de redes de interacción de proteínas. En particular encontramos que el algoritmo *Infomap* supera a la prescripción proporcionada por *CNM* en términos de su capacidad para detectar asociaciones estadísticamente significativas y biológicamente sensibles entre diversos conjuntos de proteínas considerados y los respectivos roles cartográficos.

Capítulo 5

Reposicionamiento de blancos de drogas en organismos patógenos causantes de enfermedades tropicales desatendidas: un enfoque de redes multicapa.

5.1. Resumen

Las enfermedades tropicales desatendidas (NTD) son enfermedades infecciosas humanas que se producen en regiones tropicales o subtropicales y se asocian a menudo con condiciones extremas de pobreza. Dado el notable desinterés de la industria farmacéutica, las contribuciones académicas resultan de fundamental importancia para la investigación y desarrollo de dianas terapéuticas en el tratamiento de estas enfermedades. Un objetivo usual en este contexto es la identificación de potenciales blancos de compuestos químicos que permitan fomentar el desarrollo de fármacos para atacar estas enfermedades.

En este capítulo se propone una forma original de abordar esta problemática desde la perspectiva de redes complejas. En particular, se propone un modelo de red multicapa para integrar una amplia cantidad y variedad de datos quimiogenómicos, permitendo guiar la búsqueda de potenciales blancos de drogas. Los datos integrados incluyen información de similitud química

entre compuestos, información de bioactividades entre moléculas y proteínas de múltiples organismos, así como también relaciones estructurales, metabólicas y funcionales entre proteínas. Haciendo uso del modelo de red multicapa propuesto se abordaron dos temáticas de amplio interés en el área. En primer lugar, la priorización de potenciales blancos de droga en una especie patógena de interés. En segundo lugar, la búsqueda de blancos proteicos para drogas con actividad probada, pero cuyo mecanismo de acción y blanco específico permanece desconocido.

En particular, se emplearon dos técnicas de priorización en redes, una de ellas basada en propagación de información a primeros vecinos y la otra basada en una técnica de simulación de flujo en redes complejas. Los resultados obtenidos fueron evaluados con dos criterios alternativos. En primer lugar se realizó una validación computacional mediante técnicas de validación cruzada para estimar el poder predictivo de las técnicas de trabajo utilizadas. En segundo lugar, se reportaron nuevos blancos putativos cuya pertinencia fue analizada con esfuerzos de curación manual de literatura omitida en los datos originalmente empleados para la construcción de la red multicapa utilizada.

En suma, los resultados presentados en este capítulo muestran cómo el enfoque de redes multicapa utilizado puede ayudar a guiar ejercicios de priorización de blancos de drogas en organismos completos así como también para la búsqueda de blancos para un fármaco de interés específico, aún en ausencia de cualquier tipo de conocimiento de bioactividades del mismo.

5.2. Introducción

Las enfermedades tropicales desatendidas (NTD) han devastado a la fecha, las vidas de cientos de millones de personas, con otros miles de millones en riesgo [116]. Estas enfermedades afectan principalmente a personas en condiciones de pobreza en África, Asia y las Américas. Los tratamientos actuales para estas enfermedades presentan varios problemas y limitaciones como el costo, las dificultades en la administración, las condiciones de seguridad deficientes, la falta de eficacia y el aumento de resistencia a los medicamentos, entre otros [117]. Por otra parte, ha habido un limitado interés comercial en el desarrollo y mejoras terapéuticas, en particular, debido a la naturaleza costosa y arriesgada del proceso de desarrollo de fármacos [118, 119] y el bajo retorno de la inversión que se espera cuando se trata de pacientes en condiciones de pobreza [14]. Como consecuencia, apenas del orden del 1 % de los nuevos medicamentos que llegaron al mercado en los últimos años fueron para enfermedades desatendidas [117, 120].

La situación de las enfermedades humanas que afectan al mundo desarrollado es radicalmente diferente. En este caso, se hacen muchas contribuciones importantes al descubrimiento de

fármacos cada año tanto desde laboratorios académicos como gubernamentales, lo cual lleva a la aprobación de aproximadamente 20 nuevos medicamentos por año en promedio. Como parte de este proceso de desarrollo de fármacos, se acumula mucha información acerca de compuestos bioactivos (sus actividades, blancos y mecanismos de acción), que pueden ser utilizados en estrategias de reposicionamiento.

El reposicionamiento de drogas, reutilización o reperfilado, consiste en el proceso de búsqueda de nuevas indicaciones para medicamentos ya existentes [121]. Los beneficios de este enfoque son muchos, siendo el principal, el bajo costo de desarrollo [118, 121–123]. Varias historias de éxito apoyan el uso de este tipo de enfoques. Dos de los ejemplos más conocidos son el *sildenafil* (*Viagra*) originalmente desarrollado como una droga para hipertensión, y luego reutilizada para una terapia de disfunción eréctil [123], y la *talidomida*, reutilizado para el tratamiento de mieloma múltiple y lepra [124].

Debido a los enormes ahorros en los costos que representa el reposicionamiento de un fármaco aprobado, esta estrategia es particularmente atractiva en el caso de las NTDs. En estos casos, también hay ejemplos concretos de éxito del enfoque de reposicionamiento: la *eflornitina*, que fue desarrollada como un compuesto contra el cáncer y se está utilizando para tratar la tripanosomiasis africana (enfermedad del sueño), la *pentamidina*, *anfotericina B* (originalmente un medicamento antimicótico) y la *miltefosina* fueron todos reutilizados en indicaciones de quimioterapia para el tratamiento de la leishmaniasis, una enfermedad tropical transmitida por hembras de los insectos *flebotomos* (para más ejemplos debatidos recientemente, ver [125, 126]).

La priorización de blancos de drogas, y el reposicionamiento de las mismas resultan particularmente atractivas para la utilización de técnicas computacionales de minería de datos, las cuales ofrecen un alto nivel de integración de la información disponible [127]. Estas estrategias hacen uso de herramientas quimioinformáticas y bioinformáticas, para explotar al máximo el conocimiento existente sobre blancos, fármacos, biomarcadores de enfermedades y vías de señalización, pudiendo así acelerar los correspondientes estudios clínicos. La exploración en esta forma de un gran espacio farmacológico ha dado lugar a nuevas ideas sobre los blancos y mecanismos de acción de las drogas existentes [128–131].

Desafortunadamente, éstas y otras estrategias de minería e integración de datos se han centrado en su vasta mayoría en atacar el problema desde la perspectiva de enfermedades que afectan al mundo desarrollado [15–17]. Afortunadamente, resulta relativamente sencillo el planteo de estrategias de inferencia para mapear asociaciones informativas desde organismos modelos a otras especies de interés. Recientemente, Kruger et al [132] mostraron que ligandos unidos

a más de 150 proteínas humanas se conservan mayoritariamente a través de los ortólogos de mamíferos, lo cual apoya este tipo de inferencias.

Vale la pena también mencionar que especialmente en el caso de NTDs el reposicionamiento de drogas no debe ser utilizado estrictamente para incluir medicamentos aprobados de uso clínico en humanos. Extendiendo el criterio de reposicionamiento, de manera tal de incluir medicamentos de uso veterinario, o aún más, cualquier compuesto bioactivo, se puede aumentar significativamente las posibilidades de éxito en la guía de esfuerzos tanto en el ámbito académico como de la industria farmacéutica.

En los últimos años Aguero, Crowther y colaboradores [133, 134] han desarrollado una extensa base de datos para guiar la priorización de blancos putativos en el desarrollo de fármacos en NTDs. Inicialmente, las priorizaciones de blancos se basaban sólo en características de proteínas con un uso limitado de la información disponible sobre compuestos bioactivos en la guía de estas priorizaciones. Recientemente los autores han integrado información a esta base de datos (TDR targets [135]) sobre un gran número de compuestos bioactivos, a partir de fuentes de dominio público y de una serie de ensayos de alto rendimiento, a una escala inusual para las NTDs [136–138]. Estos trabajos han llevado actualmente la integración de datos quimiogenómicos asociados a NTDs a una etapa en la cuál los ejercicios de minería de datos de gran escala son tan factibles como prometedores.

La consideración de relaciones de similitud entre pares de compuestos y proteínas pueden ser eficientemente descriptas usando conceptos de redes complejas. Bajo este paradigma pueden explorarse patrones de interconectividad no triviales para descubrir principios de organización subyacentes, identificar entidades relevantes, y novedosas asociaciones entre fármacos y blancos [139–143].

En este capítulo se presenta la construcción de una red multicapa formada por compuestos químicos, blancos de drogas (productos génicos), así como propiedades biológicas y estructurales de proteínas para guiar esfuerzos de descubrimiento y reposicionamiento de fármacos. Dado que nuestro interés puntual subyace en enfermedades tropicales NTDs, se ha hecho uso de la información contenida en la red multicapa, principalmente derivada de organismos modelos (organismos ampliamente estudiados, tales como humano, raton, etc) para orientar la selección de blancos y compuestos con miras a una posterior evaluación experimental en organismos patógenos.

En este contexto, se han abordado dos problemas bien diferenciados. Primero, se ha analizado la priorización de blancos de drogas en ausencia o escasa cantidad de información de bioactividades para un dado organismo de interés. Para este organismo, se ha considerado los datos

quimiogenómicos y de bioactividad disponibles en el resto de la red multicapa para conseguir una lista global de potenciales blancos en el organismo bajo estudio. En segundo lugar, se ha utilizado la información embebida en la red para sugerir blancos de drogas huérfanas, es decir, compuestos químicos que han mostrado actividad en una célula u organismo completo pero cuyos blancos de acción permanece aún desconocido. En este caso el objetivo es obtener una lista reducida de potenciales blancos para el compuesto huérfano de interés.

5.3. Datos quimigenómicos

En esta sección se presentará una amplia gama de datos quimiogenómicos utilizados para la construcción de nuestra red multicapa. Los mismos incluyen varios génomas completos que comprenden del orden de 1.7 x 10^5 proteínas en 221 especies (organismos patógenos y modelos), sus relaciones de homología, dominios estructurales y funcionales (Pfam) y anotaciones en diversas vías metabólicas. Así mismo, se incluye una vasta cantidad de datos estructurales para aproximadamente 1.48 x 10^6 compuestos químicos, su peso molecular, relaciones de similitud y subestructura así como sus bioactividades sobre los distintos genomas considerados. Todos estos datos fueron obtenidos de la base integratíva de dominio público TDR targets [133, 135].
En total se consideraron 12 organismos patógenos completos, todos ellos causantes de enfermedades tropicales desatendidas(NTDs): Plasmodium falciparum, Trypanosoma brucei, Trypanosoma cruzi, Leishmania major, Mycobacterium tuberculosis, Brugia malayi, Schistosoma mansoni, Toxoplasma gondii, Plasmodium vivax, Leishmania braziliensis, Leishmania infantum, y Leishmania mexicana.

Además, de estas especies patógenas se integra información de organismos modelos como: vertebrados (humano y ratón entre otros), plantas (Arabidopsis thaliana, Oryza sativa), invertebrados (Drosophila melanogaster), y nemátodos (Caenorhabditis elegans).
A continuación se describirán en detalle los distintos tipos de datos quimigenómicos considerados.

5.3.1. Grupos de homólogos

Un gen puede pensarse como una secuencia de nucleótidos en la molécula de ADN que contiene la información necesaria para la síntesis de una macromolécula con función específica, típicamente proteínas aunque también pueden ser moléculas de ARN mensajeros, ribosomales o de transferencia. Dos genes se dicen homólogos si comparten un origen evolutivo común.

Las secuencias homólogas pueden ser de dos tipos, ortólogos o parálogos. Los ortólogos son secuencias que se han separado por especiación, es decir, cuando una especie diverge en otras dos especies. Los parálogos en cambio, son secuencias separadas por un evento de duplicación, es decir, el caso donde un gen se duplica y ocupa dos posiciones distintas en el mismo genoma [144]. En general, las secuencias de genes homólogas presentan alta similitud entre sí y a menudo comparten una función común. Por este motivo suele usarse métodos basados en búsqueda de similitud de secuencia para inferir homología y conjeturar que ambos comparten la misma función, aunque tal asunción tienda a sobrestimar el número real de homólogos existentes. No obstante, la técnica suele ser suficientemente robusta para extraer patrones evolutivos y funcionales [145] y será el tipo enfoque que se utilizará en el presente capítulo.

La hipótesis subyacente asume que genes homólogos puedan ser blancos de un mismo compuesto químico. En el caso de ortólogos, este enfoque es especialmente útil para sacar provecho de los estudios realizados en especies de organismos modelo y dirigirlos a la priorización de genes como blancos de drogas en organismos patógenos causantes de NTDs.

La identificación de secuencias homólogas de las 221 especies estudiadas se extrajo de bases de datos internas del consorcio TDR targets [135]. Las mismas fueron calculadas con el algoritmo de agrupamiento OrtoMCL[146],basado en conjuntos de la base de datos OrthoMCL-DB. La idea básica es, primero producir un alineamiento de secuencias de a pares vía BLAST [147] buscando secuencias con alta similitud recíproca. Luego esta medida es utilizada para generar una matriz de distancias, que permitirá emplear el algoritmo de agrupamiento OrtoMCL[146] y encontrar grupos de alta similitud de secuencia que serán inferidos como homólogos [148]. Se encontraron 50779 grupos de homología (sólo grupos con un mínimo de dos genes anotados fueron considerados), cubriendo un total de 305872 genes a lo largo de 197 especies.

5.3.2. Dominios Funcionales de Proteínas: Pfam

Las proteínas son moléculas formadas por cadenas de aminoácidos (cadena polipeptídica). Las mismas se ensamblan a partir de la información contenida en los genes que las codifican. Cada proteína tiene su propia secuencia de aminoácidos, la que se conoce como estructura primaria. Luego estas cadenas se pliegan en pequeños plegamientos locales y regulares (estructura secundaria) que pueden tomar formas de hélice alfa, hojas plegadas beta o giros beta que a menudo conectan estructuras alfa y beta. La distribución tridimensional final que adopta la cadena polipeptídica se conoce como estructura terciaria. Esta estructura es esencial para determinar las propiedades biológicas y funcionales de la proteína, ya que condiciona su capacidad

de interacción con otros grupos funcionales [149].

Cabe mencionar también que algunas proteínas se conforman por más de una cadena polipeptídica, y en tales casos la forma en que se conjugan las mismas da lugar a lo que se conoce como estructura cuaternaria [150]. La estructura tridimensional de las proteínas contiene generalmente una o más regiones funcionales, comúnmente llamadas dominios. Los mismos pueden ser funcionales si llevan a cabo una función bioquímica determinada, o estructurales si refieren a un componente estable de la estructura [151]. Además, las diferentes combinaciones de dominios son esencialmente lo que da lugar a la diversa gama de proteínas que se encuentran en la naturaleza.

La identificación de estos dominios suele usarse también para ganar intuición sobre la función que puede cumplir una proteína. En particular, en el contexto de búsqueda de blancos de drogas resulta razonable pensar que la presencia de alguno de estos dominios funcionales o estructurales de una proteína pueda resultar esencial para que ésta constituya un blanco de acción para un determinado fármaco. Con esta idea en mente, se ha considerado en este capítulo la hipótesis de que proteínas que comparten un dominio común puedan ser blanco de la misma droga.

En la práctica, un recurso usual para buscar dominios comunes entre proteínas es utilizar búsquedas de similitud de secuencias, es decir mediante el análisis de las estructuras primarias correspondientes.

Una de las bases de datos más completas que hace uso de este enfoque para hallar dominios de proteínas es la base de datos Pfam [152]. La misma reúne información de familias de proteínas que comparten algún dominio común, y son inferidos por comparación de estructuras primarias. La base de datos Pfam contiene dos grandes clases de familias: Pfam-A que son familias de alta calidad, las cuales involucran usualmente una etapa de curación manual, y aquellas familias generadas automáticamente, Pfam-B las cuales son en general de menor fiabilidad por carecer del proceso de curación manual pero muy útiles para inferir funciones de proteínas en los casos que se carezca de dominios de tipo Pfam-A. Una misma proteína puede tener varios dominios funcionales, y por tanto pertenecer a más de una familia Pfam.

La identificación de dominios Pfam en proteínas de las 221 especies estudiadas se extrajo de bases de datos internas del consorcio TDR targets [135]. Las mismas fueron calculadas mediante el identificador de dominios de proteínas InterProScan [153], algoritmo que utiliza modelos ocultos de Markov del paquete HMMER [154]. En suma, se han considerado anotaciones de 237306 proteínas en 217 especies anotadas en 7156 familias Pfam. Cabe mencionar que sólo fueron consideradas familias con al menos 2 proteínas anotadas.

5.3.3. Vías metabólicas

Las vías metabólicas son cadenas de reacciones bioquímicas que conducen típicamente de un sustrato inicial a uno o más productos finales. Estas reacciones bioquímicas requieren típicamente para su ocurrencia temperaturas mucho más elevadas de las que existen dentro de la célula. Por lo tanto, cada reacción necesita de una disminución de la energía de activación para tener lugar. Este proceso de catálisis, es llevado a cabo por proteínas especializadas llamadas enzimas. Dada la alta selectividad de las enzimas a los sustratos que catalizan, este mecanismo le permite a la célula controlar los procesos y reacciones que tienen lugar dentro de ella.[1]

En muchas ocasiones el blanco de acción de un fármaco es una vía metabólica particularmente relevante para la patología de interés. Con esta idea en mente, se ha considerado la participación de proteínas en distintas vías metabólicas para guiar procesos de reposicionamiento de drogas. La hipótesis subyacente es que proteínas involucradas en una misma vía de metabólica puedan ser blanco del mismo fármaco.

La base de datos Kegg [155, 156] contiene información tanto de vías metabólicas, de señalización y de otros procesos celulares tales como trasducción y transcripción. No obstante, las anotaciones disponibles en Kegg no cubren completamente la totalidad de los organismos estudiados en este capítulo, especialmente en caso de organismos causantes de NTDs [157]. Por este motivo, hemos considerado anotaciones a vías metabólicas tomadas de bases de datos internas al consorcio TDR targets.

Las mismas fueron calculadas mapeando las anotaciones disponibles en KEGG a las especies de interés. Dado que las proteínas anotadas en las vías metabólicas son enzimas, el consorcio Kegg identifica cada proteína mediante un esquema de anotación exclusivo de enzimas basado en las reacciones que ellas catalizan (*Número EC*). Los genomas de las 221 especies consideradas, tienen una nomenclatura basada en la notación de UniProt, una base de datos de secuencias de proteínas ampliamente utilizada, tanto por su completitud como por ser de acceso libre. Para mapear los *Números EC* a identificadores UniProt en distintas especies, el consorcio TDR targets utiliza un alineamiento vía BLASTP, considerando mapeos positivos aquellos que cuya similitud sea significativa con un nivel de confianza $p - value < 10^{-5}$, que tengan además un porcentaje de identidad mínimo del 30 % y una cobertura mínima de un 80 % en la proteína estudiada. Como resultado, el consorcio TDR targets reporta mapeos para 18957 proteínas de 155 especies en 167 vías metabólicas distintas.

5.3.4. Datos de compuestos químicos

Se consideró información estructural y de bioactividades en 1487919 compuestos químicos extraídos principalmente de la base de datos ChEMBL [158] y complementada con curaciones manuales obtenidas de la base de datos TDR targets (especialmente para bioactividades contra organismos patógenos). Cabe destacar que estos compuestos incluyen información sobre fármacos aprobados por la FDA (U.S. Food and Drug Administration).
Los diferentes tipos de relaciones entre estos compuestos y proteínas consideradas, se describen a continuación.

5.3.4.1. Similitud estructural entre compuestos

Cada compuesto químico puede ser representado bajo la descripción de una cadena binaria, conocida como *fingerprint* o *bitmap*. Cada bit o posición de la cadena binaria de un *fingerprint* corresponde a alguna propiedad química específica, usualmente la presencia de algún elemento, enlace, anillo, etc, en la estructura química del compuesto.
Una vez adoptada la representación binaria de las moléculas, se puede medir similitud estructural mediante alguna medida de solapamiento entre las cadenas binarias. En este capítulo se ha adoptado la medida conocida como *similitud de Tanimoto*, que consiste en el cálculo del índice de Jaccard [159] sobre dos representaciones binarias de longitud fija. Dadas dos cadenas binarias A y B, tenemos la similitud de Tanimoto definida según

$$Tan(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{5.1}$$

lo cual resulta una medida de solapamiento entre ambas cadenas calculada a partir del cociente entre el número de patrones comunes a ambos bitmpas y la cantidad total de posiciones no nulas. Esta es una medida simétrica y puede ser empleada como criterio de similitud estructural entre dos compuestos químicos.

5.3.4.2. Cálculo de subestructuras

Valiéndonos de la representación binaria de cada fármaco, podemos calcular si una molécula es (o no) subestructura exacta de otra analizando si su *bitmap* está perfectamente contenido en el *bitmap* de otra molécula. Este tipo de relación de subestructura resulta en una relación

asimétrica entre moléculas y puede ser también utilizada como criterio adicional de similitud estructural entre compuestos.

No obstante, cabe mencionar que muchas de estas relaciones de subestructura exacta pueden ser poco informativas, en el sentido que moléculas muy pequeñas (cómo etanol u óxido nítrico por ejemplo) pueden estar trivialmente contenidas en estructuras más complejas. Por este motivo es necesario establecer algún criterio de filtrado para las relaciones de subestructura en base a otros indicadores que den idea de la relevancia que tiene la relación de subestructura hallada.

Dadas dos moléculas, A y B, tal que $A \subset B$, se consideran dos observables alternativos: el peso molecular de la molécula más pequeña MW y el número total de moléculas B_i en nuestra base de datos que verifiquen $A \subset B_i$, que denominaremos NS. Es decir,

$$NS = \sum_{B_i} \delta_i \qquad \delta_i = \begin{cases} 1 & si \quad A \subset B_i \\ 0 & si \quad A \not\subset B_i \end{cases} \tag{5.2}$$

Esperamos a priori, que moléculas con bajo peso molecualar y elevado NS resulten en casos de subestructura triviales y poco informativas. Para dar un ejemplo, podemos considerar la benzilamina, un compuesto orgánico que tiene un peso molecular 107.15 gr/mol y una estructura molecular trivialmente incluida en otros 381674 compuestos. Descartaremos entonces todas las relaciones de subestructura $A \subset B$ donde el peso molecular de A sea menor a un dado umbral Tmw (es decir, $MW < Tmw$) y esté involucrada en más de Tns casos de relaciones de subestructura (es decir, $NS > Tns$). La figura 5.1, muestra la cantidad de moléculas involucradas en relaciones de subestructura inespecíficas (*N.mol*) en función de los distintos valores de Tmw y Tns considerados. Cada curva se traza a un valor umbral Tns constante, mientras que en abscisas se consigna el umbral de peso molecualr Tmw y en ordenadas la cantidad de moléculas *N.mol* La definición de *subestructura inespecífica* dependerá entonces de los valores umbrales Tmw y Tns fijados.

En particular un límite razonable para el peso molecular podemos pensarlo en $Tmw = 150 gr/mol$. Situados en este umbral, la cantidad de moléculas a filtrar dependerá del límite de promiscuidad que toleremos, variando entre 165 moléculas y 347 moléculas dependiendo el límite de Tns considerado. Se fijó ad-hoc Tns=100, eliminando toda relación de subestructura donde la molécula de menor tamaño tenga peso molecular menor a 150 gr/mol y esté contenida en la estructura de más de 100 moléculas distintas. Las moléculas con estas características son $N.mol = 282$, pero hay que recalcar que en su conjunto están involucradas en un total de 4802228

relaciones de subestructuras poco informativas.

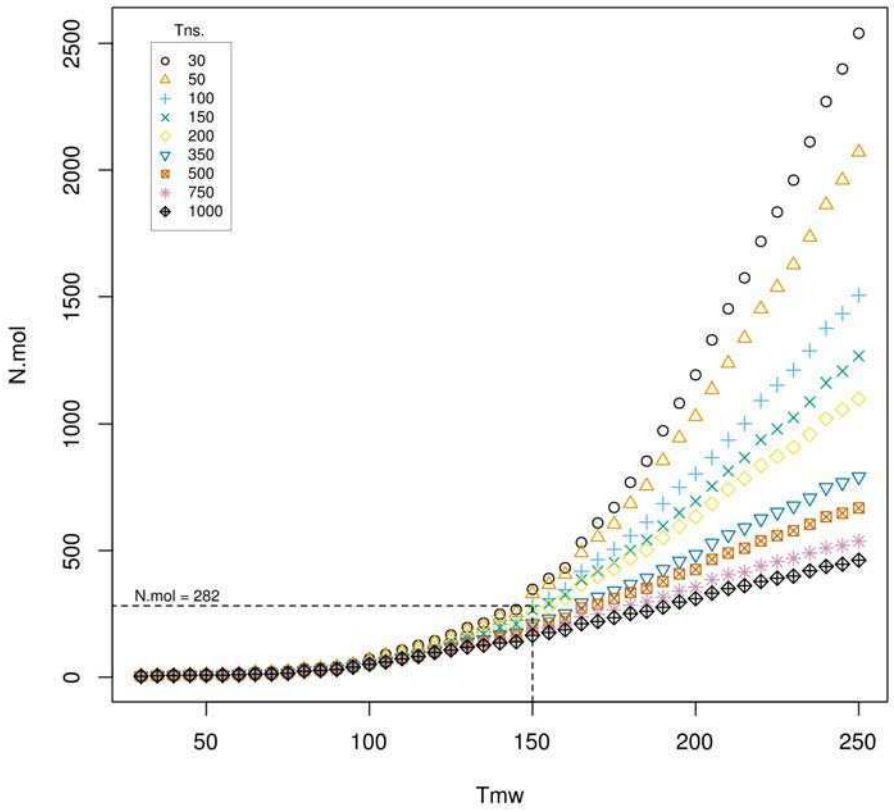


FIGURA 5.1: Número de moléculas (*N.mol*) involucradas en relaciones de subestructura inespecíficas, en función de los umbrales de peso molecualr Tmw y máxima promiscuidad tolerada Tns (ver texto para una descripción de estas cantidades). Cada curva representa un un umbral de promiscuidad Tns fijo. Las líneas de trazo indican la cantidad de moléculas que deben descartarse para para los umbrales $Tmw = 150$ y $Tns = 100$ seleccionados ad-hoc.

5.3.5. Relaciones entre compuestos y Proteínas: Bioactividades

Cuando una molécula muestra actividad específica sobre una proteína, por ejemplo inhibir las funciones que ésta lleva a cabo, diremos que existe una relación de bioactividad entre las mismas. Las bioactividades abarcan una amplia gama de experimentos con diferentes tipos de ensayos, como deleciones, inhibiciones, efectos de toxicicidad, etc. En este capítulo las relaciones de bioactividad entre fármacos y proteínas de las 221 especies estudiadas, se han extraído de bases de datos internas al consoricio TDR targets [133, 135]. Esto es en realidad una base de datos integrativa que recopila y unifica información de distintas bases de datos como ChEMBL

[158] y PubChem [160] incluyendo además curaciones manuales de literatura, especialmente en organismos patógenos.

Estas bioactividades incluyen más de 14 clases de experimentos, con ensayos de deleciones homocigotas y heterocigotas [161], mediciones de inhibición de targets bajo el efecto de un dado compuesto, mediciones de la eficacia de un compuesto en ensayos funcionales que involucran mediciones de propiedades farmacocinéticas del compuesto, su interacción con encimas metabólicas claves así como también efectos de toxicicidad en celulas y tejidos. Un detalle de los experimentos involucrados se detalla en la tabla 5.1.

Además se incluyen ensayos de EC50, es decir el reporte de valores de concentración de una droga necesario para observar el 50 % del máximo efecto sobre cualquier individuo de una población. Este valor suele usarse frecuentemente como una medida de la potencia de la droga. En contraste, se reportan también ensayos de ED50 que reportan el valor de concentración de una droga, necesario para que en el 50 % de la población se observe el efecto de la misma.
Se dispone asimismo de ensayos IC50, los cuales reportan la máxima concentración de un compuesto necesaria para inhibir una función biológica o bioquímica específica al 50 %, medida usualmente empleada como potencia antagonista de la droga. Además, se cuenta con ensayos de determinación de las constantes de disociación e inhibición en equilibrio de un compuesto, K_d y K_i respectivamente. La primera es una medida de la tendencia a disociación del conjunto receptor-ligando, mientras que la segunda es una medida de la potencia de inhibición de la droga.

Para cada tipo de experimento de bioactividad, es posible definir un umbral de corte que permita clasificar las bioactividades en dos clases: *Activas* o *Inactivas*. Un detalle de los criterios de corte establecidos en este trabajo (criterios extraídos del consorcio TDR targets) se reportan en el cuadro 5.1.

Cuadro 5.1: Clases de Bioactividades

Tipo de ensayo	Compuestos	Proteínas	Bioactiv	Activas	Umbral de corte	Fuente
Deleciones homocigotas	95	3542	889,407	65,148	$Pv < 0.01$	[161]
Deleciones Heterocigotas	247	5857	3,572,775	154,535	$Pv < 0.01$	[161]
I50	2,240	97	3,502	1,145	2 uM	[158]
IC50	152,722	2,238	297,136	184,86	2 uM	[158]
Inhibición	29,604	1,404	55,659	9,350	80 %	[158]
Kd	3,034	440	5,697	3,923	2 uM	[158]
Actividad	5,898	654	12,804	3,751	80 %	[158]
Ki	77,368	1,519	181,578	134,904	2 uM	[158]
EC50	16,221	528	30,089	20,961	2 uM	[158]
ED50	1,550	117	2,361	1,240	2 uM	[158]
Eficacia	2,748	102	5,346	1,900	80 %	[158]
Pf DHOD EC50	172	1	172	2	2 uM	[137]
Pf FP-2 EC50	172	1	172	0	2 uM	[137]
Variadas	142	24	397	148	2 uM/80 %	[135]

Tipos de experimentos de bioactividades consideradas. Las primeras tres columnas resumen el número de compuestos, proteínas y las correspondientes bioactividades entre ellos reportadas. La cuarta columna, referida como *Activas*, consigna el número de actividades que superan los valores umbrales reportados en la columna *Umbral de corte*, el cuál depende del tipo de ensayo.

5.4. Enfoque de redes multicapa

5.4.1. Integración de datos

El enfoque abordado en el presente capítulo, resume la información quimiogenómica descripta en las secciones previas en una red multicapa $G''(V = \{V_D, V_P, V_B\}, E = \{E_{DD}, E_{DP}, E_{PB}\})$, con tres tipos de nodos y distintas conexiones, distribuidos en diferentes capas (ver fig. 5.2a). La primer capa de la red consta de nodos que representan compuestos químicos o drogas (V_D). Las conexiones entre nodos embebidos en esta capa (E_{DD}) representan pares de moléculas con alto grado de similitud estructural (coeficiente de Tanimoto mayor o igual a 0.8 [162, 163]) o bien relaciones de subestructura exacta entre pares de éstos (evitando relaciones de subestructura inespecíficas debidamente reportados en la sección 5.3.4.2. En el primer caso, la conexión es bidireccional con un peso igual al valor del índice de Tanimoto (recordar que este índice es

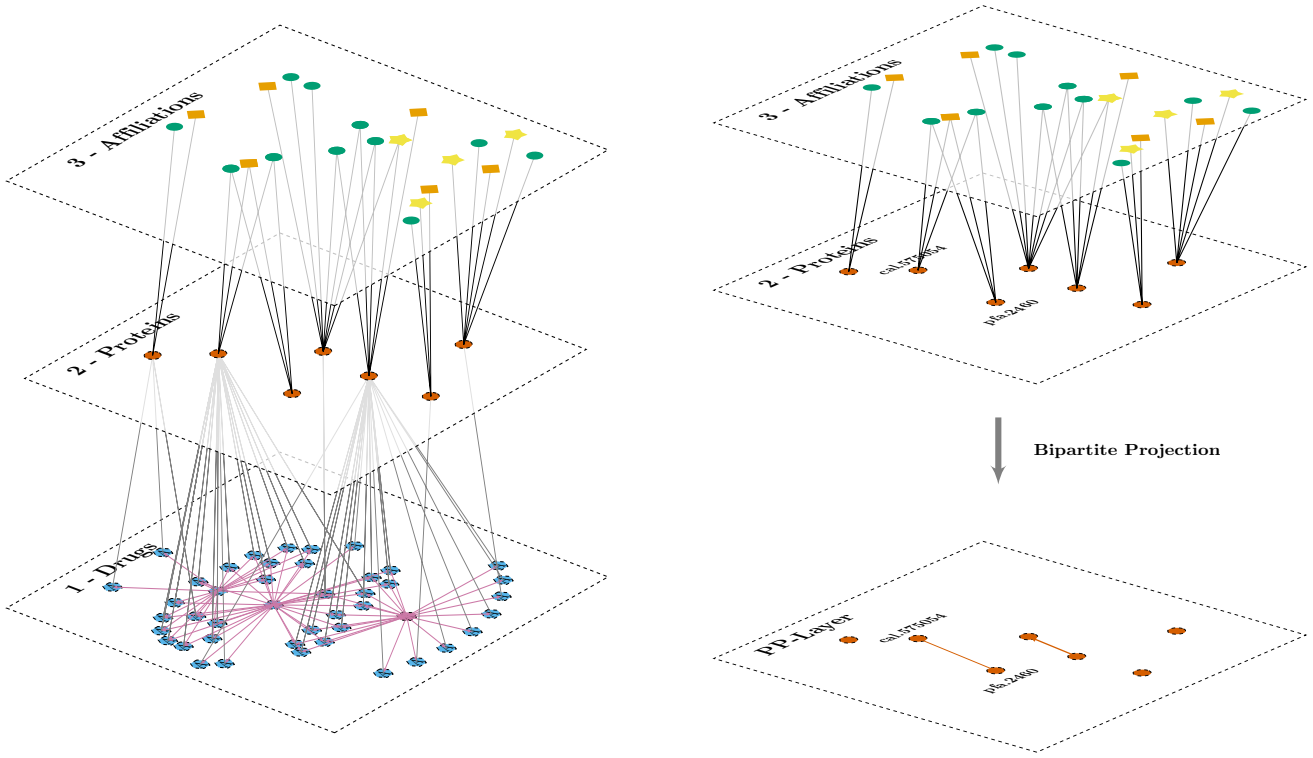
simétrico) y en el segundo caso la conexión es dirigida, con un peso uniforme w_s. Hemos considerado de mayor relevancia, la existencia de una relación recíproca de similitud estructural (índice de Tanimoto ≥ 0.8) que las relaciones de subestructura unidireccionales. Por este motivo el peso de las relaciones de subestructura se han fijado con una cota superior dada por la mínima similitud de Tanimoto considerada, y dado que el peso de las relaciones de subestructura es uniforme, se ha fijado $w_s = 0.8$.

La segunda capa consta de nodos que representan proteínas a través de las 221 especies consideradas (V_P). Notar que los nodos en esta capa no están directamente conectados entre sí, sino que lo hacen a través de nodos representados en la tercer capa de la red, es decir nodos de tipo afiliación (V_B). Se establece entonces una conexión entre una proteína en la segunda capa y un nodo de tipo de afiliación en la tercer capa (E_{PB}), en base a esquemas de análisis de secuencia y anotación que comprenden tres tipos posibles de asociaciones: relaciones de homología, dominios estructurales y/o funcionales (PFam) y por último, mapeos a vías metabólicas. De esta manera, la segunda y tercera capa en conjunto definen una red de afiliación o membresía $G_{memb}(V = \{V_P, V_B\}, E = E_{PB})$, la cual es un caso particular de red bipartita [164]. Por último, se han utilizado relaciones de bioactividad entre compuestos y sus blancos (extraídas de los experimentos referidos en la tabla 5.1) para establecer conexiones entre la primer y segunda capa, es decir, entre compuestos y proteínas(E_{DP}). La red resultante $G''(V = \{V_D, V_P, V_B\}, E = \{E_{DD}, E_{DP}, E_{PB}\})$ contiene tres capas descriptas cuantitativamente en la tabla 5.2 (columna,G')

5.4.2. Filtrado de la red e importancia relativa de nodos de afiliación

Desde el punto de vista de búsqueda y reposicionamiento de potenciales fármacos, es de suponer que la red tal cual fue descripta en la sección previa, contiene una gran cantidad de relaciones no necesariamente informativas. Por ejemplo, sólo una fracción de relaciones de afiliación (anotaciones PFam, homólogos y vías metabólicas) están conectados a proteínas que sean blanco de algún compuesto activo.

Además, existen casos bien diferenciados en cuanto a los patrones de conectividad entre compuestos bioactivos y sus blancos. Esto es, hay compuestos muy promiscuos, activos sobre múltiples blancos muy dispares desde el punto de vista de sus afiliaciones, en contraste con otros compuestos más específicos que sólo son activos frente a un conjunto de proteínas con anotaciones de afiliación bien definidas. Por lo tanto, será de utilidad filtrar en la red aquellos nodos



(A) Red Multicapa (B) Proyección Bipartita

FIGURA 5.2: **Esquema de la red multicapa propuesta**.**(A)** Red mulicapa compuesta por 3 capas distintas. En la primer capa, cada nodo representa un compuesto químico (V_D, exágonos azules en plano *1-Drugs*) y cada arista E_{DD} representa algún tipo de relación de similitud estructural. En la segunda capa, cada nodo representa una proteína de alguna de las 221 especies incluidas (V_P, círculos rojos en el plano *2-Proteins*), y la tercer capa se conforma por nodos de afiliación (V_B, incluidos en el plano *3- Affiliations*) que representan 3 tipos de criterios alternativos para evaluar similitud entre proteínas, (presencia de dominios *Pfam*, círculos azules), actividad en vías metabólicas (estrellas amarillas) y pertenencia a grupos de homología (rombos naranjas). Las aristas entre las capas de drogas y proteínas (E_{DP}), representan relaciones de bioactividad probadas según se describe en la tabla 5.1. Las conexiones ente nodos V_P y V_B (E_{PB}) establecen relaciones de pertenencia entre proteínas y nodos de afiliación, y en su conjunto con las dos capas que los contienen conforman una red bipartita. **(B)**. Proyección de la red bipartita conformada por las dos capas superiores en una única capa monopartita definida en la ecuación 5.7, donde cada nodo representa una proteína y las aristas resultantes de esta proyección E_{PP} indican que dos proteínas comparten algún nodo de afiliación. El peso o intensidad de estas conexiones viene dado por la ecuación 5.8 y depende de la importancia del nodo de afiliación compartido para el problema de búsqueda de blancos de droga.

y conexiones espúreas para el problema de búsqueda de potenciales fármacos, manteniendo exclusivamente los que sean pertinentes para el problema de predicción de blancos de drogas. Con este objetivo, se han conservado sólo nodos de afiliación V_B que contengan al menos una proteína V_P con una conexión directa de bioactividad E_{DP} bajo los umbrales de actividad definidos en el cuadro 5.1 (*blanco* para referencia futura). Es decir, se conservaron sólo nodos de afliación que contengan anotado al menos un blanco protéico.

Como resultado del filtrado expuesto, se obtiene una red multicapa $G'(V = \{V_D, V_P, V_B\}, E = \{E_{DD}, E_{DP}, E_{PB}\})$ comprende 5186 nodos de afiliación V_B informativos, de los cuales 2252 corresponden a dominios funcionales PFam, 2789 grupos de homólogos y 145 vías metabólicas. El cuadro 5.2, consigna el numero total de nodos y conexiones de la red en los diferentes planos, antes y después de filtrar nodos con los criterios mencionados (columnas G'' y G' respectivamente).

		G''	G'
V_D		1.488.034	1.487.919
V_P		385.711	167.815
V_B	Todos	58.102	5.186
	PFam	7.156	2.252
	Ort	50.779	2.789
	V.Metab.	167	145
E_{DD}	Todos	170.272.699	67.629.415
	Tanim	44.403.424	44.402.716
	Subestr	125.869.275	26.714.379
E_{DP}		427.338	325.843
E_{PB}	Todos	738.682	718.277
	PFam	333.188	331.928
	Ort	325.017	305.872
	V.Metab.	80.477	80.477

Cuadro 5.2: Número de nodos y aristas de la red multicapa construida. G'' corresponde a la red resultante de la integración de datos originales, sin ningún tipo de filtrado. G' corresponde a la red multicapa después de aplicar distintos criterios de filtrado en cada capa de la red (ver texto para el detalle de los mismos)

Finalmente resultará particularmente útil asignar un peso relativo a cada nodo de afiliación V_B en base a la proporción de *blancos* que tenga anotado. Con esta idea en mente, para cada nodo de afiliación V_B podemos construir una tabla de contingencia de 2 x 2 que condense

la información de cuantas proteínas V_P están (o no) anotadas a él, y cuantas de ellas son (o no) blancos de droga. La red contiene un total de N proteínas de las cuales $\#B$ son blancos de algún compuesto químico. Por otro lado, el nodo de afiliación V_B contiene en total $\#V_B$ proteínas anotadas de las cuales V'_p son blancos de alguna droga (ver tabla 5.3). Bajo esas condiciones, la probabilidad de que por azar el nodo V_B tenga exactamente ese número de blancos anotados sigue una distribución hipergeométrica y puede ser calculada de forma exacta mediante la ecuación 5.3

$$p_{V_B}(V'_P) = \frac{\binom{\#V_B}{V'_P}\binom{N-\#V_B}{\#B-V'_P}}{\binom{N}{\#B}} \tag{5.3}$$

	blanco	$\overline{blanco}$	Total
V_B	V'_P		$\#V_B$
$\overline{V_B}$			
Total	$\#B$		N

Cuadro 5.3: Tabla de contingencia construida para un nodo de afiliación V_B, con $\#V_B$ proteínas anotadas en él, de las cuales V'_P son blanco de algún compuesto bioactivo. El número total de proteínas en toda la red es N de las cuales $\#B$ son blancos. Esta tabla permite realizar una prueba de Fisher que determine el nivel de relevancia del nodo de afiliación V_B en el contexto de búsqueda de blancos de droga.

Notemos que, cuanto más relevante sea el nodo de afiliación V_B para el problema de búsqueda de blancos menor será la probabilidad p_{V_B} de hallar por azar una tabla de contingencia que contenga las mismas probabilidades marginales ($\frac{V_B}{N}$ y $\frac{\#B}{N}$) y una proporción mayor o igual a la observada de blancos protéicos $\frac{V'_P}{\#V_B}$. Por lo tanto, la probabilidad de haber observado por azar una tabla de mayor o igual proporción de blancos proteicos, es decir, la significancia estadística de la prueba de Fisher calcula mediante

$$pv(V'_P) = \sum_{x \geq V'_P} p_{V_B}(x) \tag{5.4}$$

Para transformar estos valores de probabilidad a un valor de peso, de manera que nodos de afiliación relevantes tengan mayor pesoque aquellos de menor significancia, se debe transformar estos valores con algún tipo de función monótona decreciente. Una búsqueda y selección de esta función de transformación se presenta en la sección 5.4.4.

5.4.3. Proyección de la red de afiliación

Consideramos la red de afiliación comprendida por la segunda y tercera capa de nuestra red completa $G_{memb}(V = \{V_P, V_B\}, E = E_{PB})$. Como se mencionó en la sección2.7.1.1, una propiedad importante de este tipo de redes de afiliación, es que nodos en la tercer capa (V_B) pueden utilizarse para inferir conexiones entre los nodos de la segunda capa (V_P) y viceversa. En este caso, se utilizarán los nodos de afiliación V_B para inferir conexiones entre proteínas E_{PP} mediante una proyección bipartita. Es posible condensar la información contenida en el subgrafo $G_{memb}(V = \{V_P, V_B\}, E = E_{PB})$ en una única capa $G_P(V = V_P, E = E_{PP})$, donde los nodos sean proteínas y las conexiones entre ellas estén dadas por el número nodos de afiliación compartidos y su relevancia estadística. Formalmente, se trata de la proyección de una red bipartita conformada por proteínas y nodos de afiliacíon (PFam, Ortólogos y vías metabólicas) sobre la capa de proteínas. Para llevarla a cabo, se implementó una versión modificada de la proyección bipartita propuesta por Zhou [46]. Dicha proyección, conecta dos proteínas sólo si éstas comparten al menos un nodo de afiliación, y el peso o relevancia de dicha conexión está dado por un balance entre el número de afiliaciones compartidas entre ambas proteínas, la cantidad de afiliaciones de éstas, y el número de proteínas anotadas a éstas afiliaciones (para más detalle ver capítulo 2.7.1.1). En la versión propuesta en la presente tesis, se tiene en cuenta adicionalmente la importancia relativa de cada afiliación, que fue cuantificada considerando su proporción de blancos de drogas (ver ecuación 5.4).

La red de afiliación $G_{memb}(V = \{V_P, V_B\}, E = E_{PB})$ compuesta por la capa de n proteínas V_P y m afiliaciones V_B, puede ser descripta mediante su matriz de adyacencia $M \in n\text{x}m$, donde el elemento m_{ij} de la misma puede tomar sólo dos valores posibles:

$$m_{ij} = \begin{cases} 1 & \text{si} \quad \exists \quad E_{PB} \in E \quad \backslash E_{PB} = E_{ij} \\ 0 & \text{en otro caso} \end{cases} \tag{5.5}$$

en otras palabras, $m_{ij} = 1$ sólo si la proteína V_{p_i} pertenece al nodo de afiliación V_{B_j}, ya sea PFam, Ortólogo o vías metabólicas. Bajo descripción matricial, el hecho de que los nodos de afiliación posean una importancia relativa dada por un vector de pesos w_j, puede ser representado con una matriz diagonal $S \in m\text{x}m$, con

$$S_{jk} = \begin{cases} 0 & \text{si } j \neq k \\ w_j & \text{si } j = k \end{cases} \tag{5.6}$$

En términos de esta notación, el subgrafo $G_{memb}(V = \{V_P, V_B\}, E = E_{PB})$ descripto por la matriz M puede proyectarse en una capa de proteínas $G_P(V = V_P, E = E_{PP})$ (*capa-PP*) descripta por su matriz de adyacencia $M_P \in nxn$, mediante la transformación:

$$M_P = \widehat{M}\, S\, \widehat{M^T} \quad , \quad \widehat{m_{ij}} = \frac{m_{ij}}{\sum_j m_{ij}} \tag{5.7}$$

donde, M^T indica la operación de trasposición de la matriz M, y $\widehat{M}$ señala la operación de normalización por columnas. Notar que la proyección bipartita desarrollada por Zhou y colaboradores [46] (ver capítulo 2.7.1.1), se reduce a un caso particular de esta proyección, tomando S como la matriz identidad, es decir, asignando igual importancia a los nodos de afiliación.
Como resultado de esta proyección, la red global original resulta ahora en una red multicapa de sólo dos niveles $G(V = \{V_D, V_P\}, E = \{E_{DD}, E_{DP}, E_{PP}\})$, tal como se representa en la figura 5.2b. Nos referiremos a la capa inferior como la capa de compuetos (*capa-D*) y la capa superior como la capa proyectada en proteínas (*capa-PP*). Ambas capas integran diferentes tipos de información de compuestos y proteínas y se utilizarán para propagar información en distintos problemas de priorización y reposición de fármacos y sus blancos, como se describe en las siguientes secciones.

5.4.4. Asignación de pesos para nodos de afiliación

Como se mencionó en el capítulo precedente, el nivel de importancia de los distintos nodos de afiliación fue ponderado en base a la probabilidad calculada en la ecuación 5.4. La idea ahora es utilizar explícitamente estos valores de probabilidad para construir la matriz diagonal S (ecuación 5.6). Construida de esta manera, S permitiría generar una proyección bipartita contemplando la importancia relativa entre nodos de afiliación para la búsqueda de blancos proteicos. La red multicapa resultante tendrá una capa $G_P(V = V_P, E = E_{PP})$ donde las proteínas que compartan nodos de afiliación con alta proporción de blancos proteicos serán conectadas por aristas de alto peso. De esta manera, la red G_P podrá ser utilizada para abordar distintos problemas de priorización de blancos y reposicionamiento de fármacos que competen al presente capítulo.

Para cada nodo de afiliación V_{B_j} tenemos asignado según ec. 5.4, un valor de probabilidad p_j. Es posible transformar estos valores en un vector de pesos w_j mediante la aplicación de una función no lineal monótona decreciente y utilizarlo así para definir la matriz diagonal S (ver ec. 5.6). En lo siguiente se propone como función de transformación una ley de potencias (ec. 5.8),

con un parámetro libre α que permite variar su forma funcional (ver figura 5.3). Los valores de p_j obtenidos en la ecuación 5.4 abarcan varios órdenes de magnitud ($p_j \in (10^{-251}, 1)$), por lo que no resultaría apropiado aplicarles directamente una ley de potencias. En cambio resulta particularmente útil el empleo de la función logarítmica y su posterior normalización al intervalo (0,1) para aplicar luego la ley de potencias propuesta. Adicionalmente se ha incluido un factor de corrección K_j que representa el grado del nodo de afiliación V_{B_j} y penaliza la promiscuidad del mismo. Por otro lado, es deseable que la presencia de algunos pocos nodos con valores extremos en la distribución ($p_j \sim 10^{-250}$) no resulten predominantes en la transformación propuesta. Con esta idea en mente, el 20 % de los nodos de afiliación más relevantes (el cuantil 0.2 de la distribución de p_j, $q_{0.2}$) no se ha afectado por la ley de potencias sino que sólo se los ha penalizado por su grado de promiscuidad.

$$w_j = \begin{cases} \frac{1}{K_j} & si \quad p_j < q_{0.2} \\ \frac{1}{K_j} \left[\frac{-log_{10}(p_j)}{\max_j\{-log_{10}(p_j)\}} \right]^{\alpha} & si \quad p_j \geq q_{0.2} \end{cases} \qquad (5.8)$$
$$1 < j < m\,, \quad \alpha \in [0, +\infty)$$

La figura 5.8 consigna distintas formas funcionales en el rango de $\alpha \in (0.1,10)$, para un valor de K_j fijo (tomamos por simplicidad $K_j = 1$, aunque recalcamos que todo nodo de afiliación V_B tiene grado mayor o igual a 2 para hacer la estructura del grafo G conexa). Valores de $\alpha \sim 0$ proporcionan buena resolución para diferenciar casos de poco nivel de significancia estadística ($-log_{10}(p) \lesssim 1$) y pierden resolución para diferenciar casos muy significativos (i.e, para $-log_{10}(p) \gg 1$, tenemos $w \sim \frac{1}{K_j}$.). El caso opuesto desde el punto de vista de resolución se tiene para $\alpha > 1$, pues aquí para $-log_{10}(p) \lesssim 1$, tenemos $w \sim cte \sim 0$., perdiendo resolución para diferenciar casos poco significativos, y penalizando el w correspondiente. Por otro lado, para $\alpha = 1$ se tiene una transformación lineal respecto al logaritmo de la probabilidad.

El parámetro libre α se fijó de manera de optimizar la predicción de blancos de droga en la red vía validación cruzada de 10 iteraciones. El procedimiento correspondiente se detalla en la sección 5.4.5. Notar que la proyección bipartita de la red depende de la matriz diagonal S (ecuación 5.7) y ésta a su vez de la asignación de pesos dada por la ecuación 5.8 cuyo único parámetro libre es α. En suma, la proyección bipartita planteada depende explícitamente del parámetro α y como veremos en la sección 5.4.5 el procedimiento de validación cruzada de 10 iteraciones provee un método robusto para fijar este parámetro.

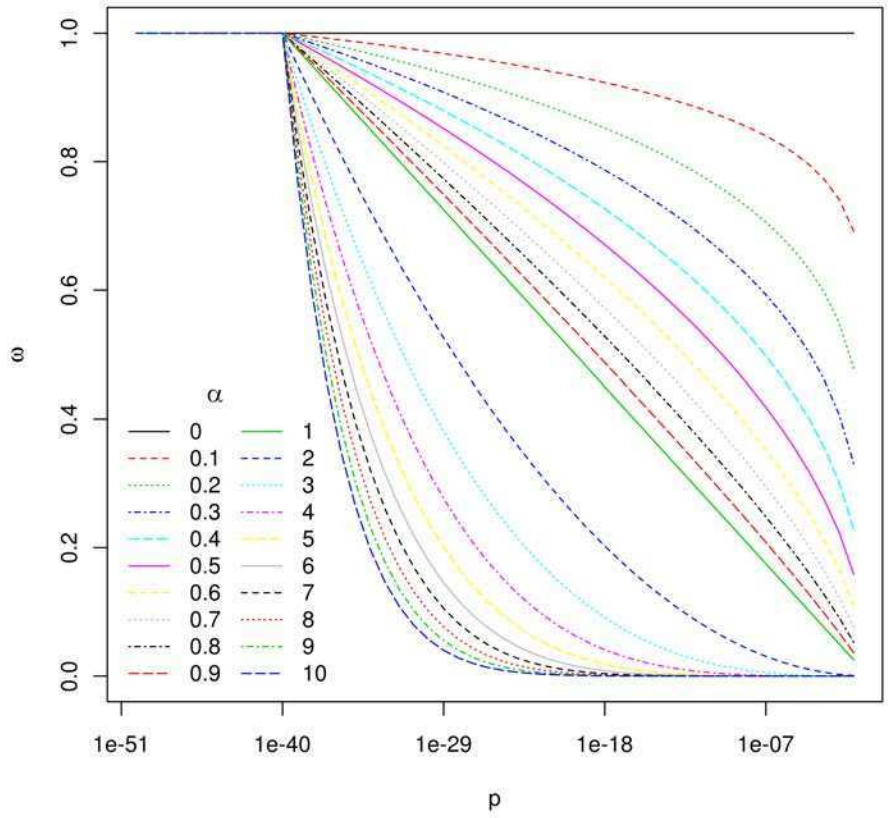


FIGURA 5.3: Ley de poteincias (ec. 5.8) empleada para transformar los valores de significancia estadística p_j de los nodos de afiliación (ec.5.4), en un vector de pesos relativos w_j. Estos pesos son utilizados explícitamente en la proyección bipartita de la red multicapa. Esta figura se corresponde a un valor fijo de $K_j = 1$. Notar que para valores de $\alpha > 1$ la transformación incrementa su resolución en los nodos más significativos, mientras que para $\alpha < 1$ la resolución incrementa en nodos de menor significancia estadística.

5.4.5. Selección de parámetros

Como se mencionó en la sección anterior, la proyección bipartita y por tanto la obtención del grafo $G_P(V = V_P, E = E_{PP})$ de la red multicapa depende del parámetro libre α. Es decir, cada α genera un grafo $G_P(V_P, E_{PP}, w(\alpha))$ distinto que abreviaremos G_P^α. En vista a que se pretende utilizar esta red para priorizar blancos de droga resulta razonable pensar en fijar este parámetro optimizando la capacidad de la red para priorizar blancos proteicos. Para ello, podemos considerar una fracción de blancos V_P (supongamos un 10 %) a los que nos referiremos como *conjunto de evaluación*. Luego, se eliminan temporalmente las bioactividades E_{DP} que apuntan a proteínas en este conjunto y se utiliza el restante (90 %) conjunto de proteínas, *conjunto de entrenamiento*, para propagar información en cada una de las redes G_P^α y tratar de

predecir como blancos a las proteínas del conjunto de evaluación. Para cada valor de α el resultado típico de un ensayo de priorización de este tipo es una lista ordenada de proteínas que idealmente tienen a los blancos de drogas en las primeras posiciones del ordenamiento. Si al recorrer la lista se observan a todos los blancos en las primeras posiciones sin la presencia de ningún falso positivo diremos que tenemos un predictor perfecto. Si en cambio el orden de la lista está completamente descorrelacionado con la cualidad de ser blancos de droga se dice que el predictor es aleatorio. Es posible cuantificar la calidad del método de predicción (y por tanto la calidad de G_P^α) mediante el uso de alguna métrica de desempeño como el área bajo una curva ROC (Receiver Operating Characteristic) limitada a una tasa de falsos positivos del 10 %, que se denota AUC-01 (ver sección 2.6.2). Con un valor de AUC-01 normalizado según la ecuación 2.25, un predictor aleatorio oscilará en un rango próximo a AUC-0.1$\sim \frac{1}{2}$, mientras que un predictor perfecto se corresponde con un valor AUC-01=1. En un procedimiento de validación cruzada de 10 iteraciones, se divide el conjunto de blancos en 10 partes iguales y se utiliza alternativamente cada una de éstas como conjunto de evaluación. Luego, a lo largo de las 10 iteraciones podremos obtener un valor de desempeño para cada grafo $G_P(V = V_P, E = E_{PP})$ promediando los 10 valores de AUC-01 obtenidos en cada caso. Además, dado que la división del conjunto de blancos en 10 partes iguales es aleatoria, es usual repetir el procedimiento m veces para considerar las fluctuaciones aleatorias correspondientes (aquí hemos considerado m=30).

La pregunta que abordaremos ahora es, si existe alguna red G_P^α que resulte óptima para la predicción de blancos de droga. Se realizó un procedimiento de validación cruzada contemplando todos los blancos de droga en la red exceptuando aquellas que pertenecen a especies de Tripanosoma Cruzi y M. Musculus. Esta exclusión se debe a que en secciones posteriores los blancos de estas especies serán utilizados para validar los métodos predictivos empleados, y es necesario excluirlos aquí, para evitar un posterior sobreajuste. En total se consideran 5748 blancos proteicos que fueron divididos en 10 conjuntos de igual tamaño utilizados alternativamente como conuntos de evaluación. Para propagar la información del restante 90 % de blancos a través de las conexiones de la red G_P^α correspondiente, se utilizó el algoritmo de propagación a primeros vecinos VS presentado en el capítulo 2.6.1. Dado que la proyección bipartita G_P^α empleada utiliza información del conjunto de entrenamiento, es importante recalcar que la misma tuvo que ser recalculada en cada una de estas 10 iteraciones para evitar un trivial sobreajuste.

Luego, en el algoritmo VS, los nodos V_P del conjunto de entrenamiento se utilizan como semillas que transmiten la información a sus vecinos directos con una intensidad proporcional al peso de las aristas que los unen a éstos. Cada nodo de la red obtiene una puntuación total igual

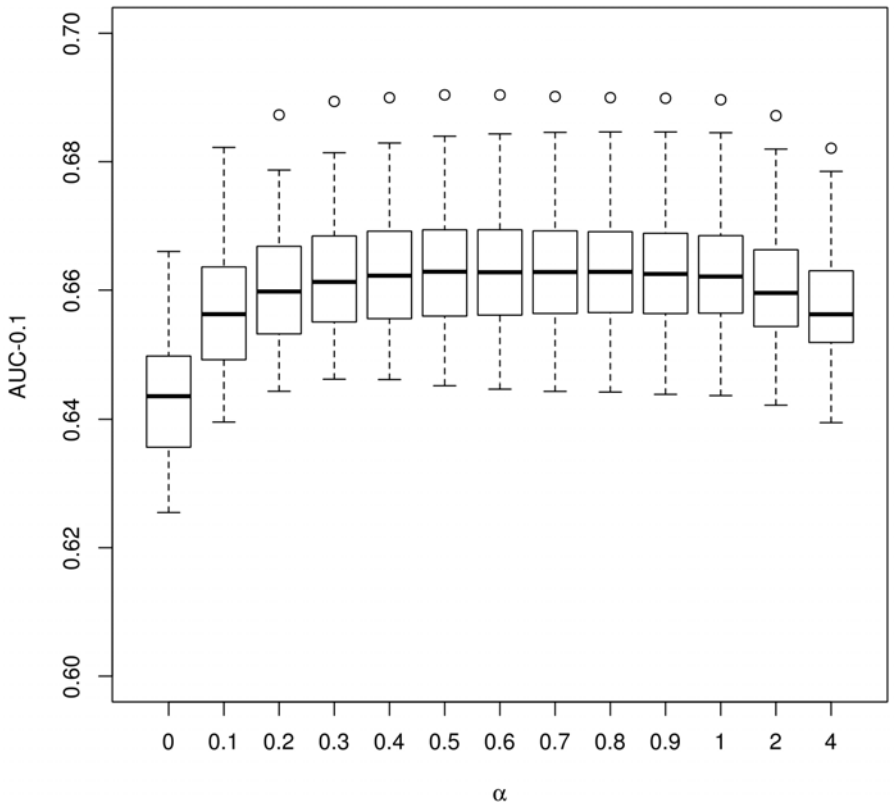


FIGURA 5.4: Validación cruzada de 10 iteraciones para ajustar el parámetro α correspondiente a la transformación presentada en la ec. 5.8. En ordenadas se consigna la distribución de AUC-01 obtenida a lo largo de 30 repeticiones de la validación, en función del parámetro α considerado. Resulta interesante notar que las distribuciones de AUC-01 para valores de $\alpha = 0$ y cualquier $\alpha \in [0.2, 1]$ difieren significativamente a un nivel de confianza $p - val \lesssim 10^{-24}$ (prueba de wilcoxon)

a la suma de las intensidades de todos los mensajes que recibe. De esta manera se constituye la lista ordenada (donde las semillas no son incluidas) y se calcula el AUC-01 como medida desempeño de para recuperar blancos en el conjunto de evaluación considerado. El valor de AUC-01 se promedia sobre las 10 iteraciones realizadas. Dado que el valor medio de AUC-01 obtenido puede depender de la forma particular en que el conjunto de blancos fue dividido, se repite el procedimiento de validación cruzada en un ensamble de 30 particiones diferentes de este conjunto. La figura 5.4, consigna la distribución de AUC-01 en este ensamble para cada red G_P^α considerada. Se observa que para valores de $\alpha \in [0.2, 1]$ el valor medio de AUC-01 no presenta diferencias apreciables, disminuyendo el desempeño para valores de α fuera de este rango. Dos conclusiones a priori pueden extraerse de esta figura. La primera es que se no se observa una red G_P^α óptima para predecir blancos de droga, sino que hay un conjunto de

ellas con desempeño equivalente en el rango de $\alpha \in [0.2, 1]$. En adelante consideraremos el punto medio de este intervalo $\alpha = 0.6$. La segunda observación, es que $\alpha = 0$ (que equivale a no considerar la transformación por ley de potencias) tiene un valor de desempeño inferior al que se obtiene con la transformación en cualquiera de los α mencionados. En particular, si comparamos las distribuciones de AUC-01 para $\alpha = 0$ y $\alpha = 0.6$ las distribuciones de AUC-01 presentan diferencias significativas a un nivel $p - valor \sim 10^{-28}$ (prueba de Wilcoxon). Esto sugiere que incluir una medida de relevancia como la considerada para los nodos de afiliación, mejora el desempeño de priorización de manera estadísticamente significativa.

5.5. Priorización de blancos en organismos completos

En esta sección se considerará el problema de priorización de blancos de drogas para una especie completa Q, la cual tiene escasa (o completamente nula) cantidad de datos de bioactividad. El objetivo es identificar dentro del proteoma de un organismo de interés potenciales candidatos que puedan resultar blancos de alguna droga conocida. Este problema es típico en organismos patógenos causantes de enfermedades tropicales desatendidas. Se pretende hacer uso de la información disponible en otros organismos patógenos y modelos embebida en la red multicapa construida en las secciones precedentes (fig 5.2b). Las estrategias de priorización llevadas a cabo en este capítulo hacen un uso exhaustivo de la red $G(V = \{V_D, V_P\}, E = \{E_{DD}, E_{DP}, E_{PP}\})$ y la capa *Capa-PP* definida en la sección 5.4.3, es decir del grafo proyectado $G_P^{\alpha=0.6}$. Este grafo puede representarse matricialmente mediante $M_P \in n\mathbf{x}n$ (ec. 5.7) la cual se conforma por proteínas de distintas especies conectadas según la cantidad y relevancia de los nodos de afiliación compartidos. La idea es identificar el conjunto de blancos V_P conocidos en la red M_P, y utilizarlos como semillas de dos diferentes estrategias de priorización sobre la capa M_P que permitan revelar nuevos blancos en la especie de interés Q. En la primer estrategia que denominaremos VS, las semillas transmiten información sólo a sus primeros vecinos de la red. En la segunda, que denominaremos *Functional Flow* o *Flujo Funcional* [41], se implementa un proceso iterativo donde vecinos de las semillas a mayor distancia geodésica pueden ser alcanzados en la priorización. Ambas estrategias se describieron en detalle en la sección 2.6.1

5.5.1. Validación in-sílico de estrategias de priorización

En esta sección se pretende realizar una evaluación exhaustiva de las estrategias de priorización propuestas. La idea detrás del enfoque planteado es que las asociaciones entre blancos y drogas en la red $G(V = \{V_D, V_P\}, E = \{E_{DD}, E_{DP}, E_{PP}\})$ puedan ser utilizados para propagar

asociaciones no triviales y significativas, resaltando o proveyendo putativos blancos V_P en la especie de interés Q. Para validar esta idea, el enfoque que se propone consiste en eliminar toda bioactividad E_{DP} asociada a proteínas de la especie Q elegida, y evaluar la capacidad de cada estrategia de priorización para priorizar los blancos conocidos entre todas las demás proteínas del genoma Q. Para evaluar cuantitativamente la capacidad de predicción de cada algoritmo, utilizaremos el área bajo curvas ROC restringidas a un 10 % de falsos positivos (AUC-01). Se realizó este análisis para dos especies de interés M. musculus y T. cruzi. El primer caso es un genoma de mamíferos usualmente utilizado como modelo para blancos de medicamentos humanos, por lo cual existe amplia evidencia acumulada en esta especie (284 nodos V_P que representan el 4.7 % de los blancos de la red, y 8429 $E_{DP} \in E$ que representan un 2.6 % de todas las conexiones E_{DP}). El segundo es un genoma protozoo, vector de la enfermedad de Chagas, que representa un caso típico de organismos poco estudiados. En este caso, se cuenta con evidencia para solamente 19 proteínas V_P que son blancos conocidos. Los mismos representan sólo el 0.1 % de todo los blancos la red multicapa e involucran 323 conexiones $E_{DP} \in E$ que representan un 0.3 % del total de las Bioactividades E_{DP}.

La tabla 5.5.1, resume los valores de AUC-0.1 obtenidos para los métodos VS y FF. Adicionalmente se reportan los valores de AUC-01 utilizando dos proyecciones G_P distintas para construir el grafo $G(V = \{V_D, V_P\}, E = \{E_{DD}, E'_{DP}, E_{PP}\})$. La primera genera una red multicapa con una capa proyectada G^k_P, es deicr, utiliza la corrección de grado en la función de transformación dada en la ec. 5.8 (columna **1/K**) y la segunda omite esta corrección (columna **s.c**) generando una red multicapa con una capa proyectada G^{sc}_P.

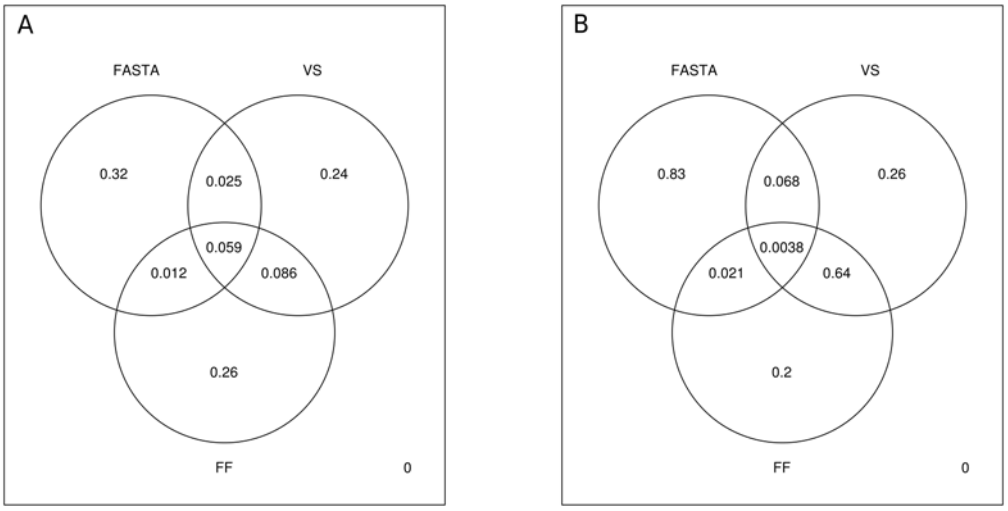
Método	VS	FF	FASTA	P-valor
	1/K | s.c	**1/K** | s.c		
Mus Musculus	**0.72** | 0.64	**0.66** | 0.65	0.64	$2.8\ 10^{-6}$
T.cruzi	**0.81** | 0.52	**0.72** | 0.52	0.72	$8.1\ 10^{-2}$

Cuadro 5.4: Comparación de desempeño para distintas estrategias de priorización utilizadas. Cada estrategia se evalúa calculando la correspondiente área AUC-01 normalizada (Mc.Clish Correction [42]). Para dos especies de interés (T.curzi, y M.musculus) se compara el desempeño de los algoritmos VS y FF ambos con y sin corrección de grado. Se consigna además los valores de AUC-01 para una priorización basada en alineamiento de secuencias realizada con el algoritmo FASTA (ver texto). La última columna consigna los resultados de p-valor de la prueba estadística realizada por remuestreo para evaluar si las diferencias entre valores de AUC-01 observados en las priorizaciones de FASTA y VS (con corrección 1/K) son estadísticamente significativas.

En ambas especies y para ambos algoritmos el valor de predicción mejora al emplear la corrección que penaliza la promiscuidad en nodos de afiliación, especialmente en el caso de VS sobre la especie TCR donde pasamos de un predictor random AUC-0.1 = 0.52 (s.c) a otro de AUC'01=0.81 utilizando la corrección. Además notamos que en ambas especies considerando las priorizaciones realizadas con corrección de grado el algoritmo VS supera el desempeño de FF mientras que sin la corrección ambas estrategias de priorización obtienen similares resultados.

Por otro lado, a modo de control, se consideró una tercer estrategia de priorización que no utiliza la estructura general de la red construida. Esta estrategia, FASTA en adelante, está basada en la estimación de similitud de secuencia (estructura primaria) entre los blancos de todas las especies (exceptuando M.Musculus o T.Cruzi según se el caso bajo análisis) y las proteínas de la especie de interés Q. Para este fin, se utilizó la herramienta de alineamiento de secuencias FASTA [165] que produce alineamientos de zonas continuas de la proteína en lugar de dividir en pequeñas regiones de alta similitud de secuencia como otros conocidos algoritmos hacen (ej. Blast, [166]). La hipótesis subyacente es que un compuesto se liga una proteína a través de un sector contiguo y no en pequeños fragmentos. La tabla 5.5.1, (columna FASTA) consigna los valores de AUC-0.1 normalizados para esta estrategia de priorización. La comparación muestra que VS mejora la estrategia basada en alineamiento FASTA para ambos organismos estudiados, aunque las diferencias resultan estadísticamente significativas sólo en el caso de M.Musculus (prueba de remuestreo con 2000 réplicas, tabla 5.5.1, columna "p-valor"). Esto sugiere que el aprovechamiento de la información embebida en la red puede mejorar predicciones basadas meramente en similitud de secuencia.

Por otro lado, existen formas alternativas al cálculo de AUC-01 para comparar las listas de distintos algoritmos de priorización. El cálculo de AUC-01 ofrece una medida cuantitativa, pero está restringida sólo a la capacidad de detectar blancos conocidos del conjunto de evaluación. No obstante, es posible comparar dos listas de priorización cuantificando el grado de solapamiento que tiene en la zona de proteínas mejor puntuadas. Más allá de que dos algoritmos presenten similares resultados desde el punto de vista de los valores de AUC-01 obtenidos, es importante cuantificar en que medida las listas de priorización que proveen se solapan o complementan entre sí. Con este fin, se ha considerado el 1 % de proteínas mejor puntuadas en las listas de los tres métodos: VS , FF, y FASTA. En el caso de T.Cruzi esta lista se restringe a las primeras 65 proteínas mientras que para el caso de M.Musculus la lista se restringe a las primeras 134 proteínas. La fig 5.5 consigna los diagramas de venn indicando en cada caso la fracción de proteínas que comparten las distintas listas. Por ejemplo para T.Cruzi (ver figura 5.5.B) vemos



ch

FIGURA 5.5: Intersección de listas para las distintas estrategias de priorización utilizadas. Tdoas las listas de priorización analizadas en estos diagramas se restringen al 1 % de las proteínas en cada especie que reciben mejor puntaje. El panel (A) representa el caso de M.Muculus que involucra 134 proteínas y el panel (B) el caso de T.Cruzi que involucra 65 proteínas. Como es de esperar, en ambas especies el acuerdo entre métodos basados en la red multicapa es mayor al que éstos presentan con aquél basado en alineamiento de secuencias (FASTA). Por otro lado, notamos que el acuerdo común entre los tres algoritmos aumenta un orden de magnitud en el caso de M.Musculus

que el acuerdo entre las listas de los algoritmos basados en la red multicapa (18.6 %) es considerablemente mayor al que cualquiera de éstos tiene con el algoritmo basado en alineamiento de secuencias, FASTA (3.1 y 1.9 % para FF y VS respectivamente). En el caso de M.Musculus (5.5.A) en cambio, si bien el acuerdo entre FF y VS sigue siendo mayor al que se tiene con FASTA, el acuerdo común entre los tres algoritmos incrementa en un orden de magnitud respecto al observado en las priorizaciones sobre T.Cruzi (5.9 % para el caso de M.Musculus, y sólo un 0.6 % en el caso de T.Cruzi).

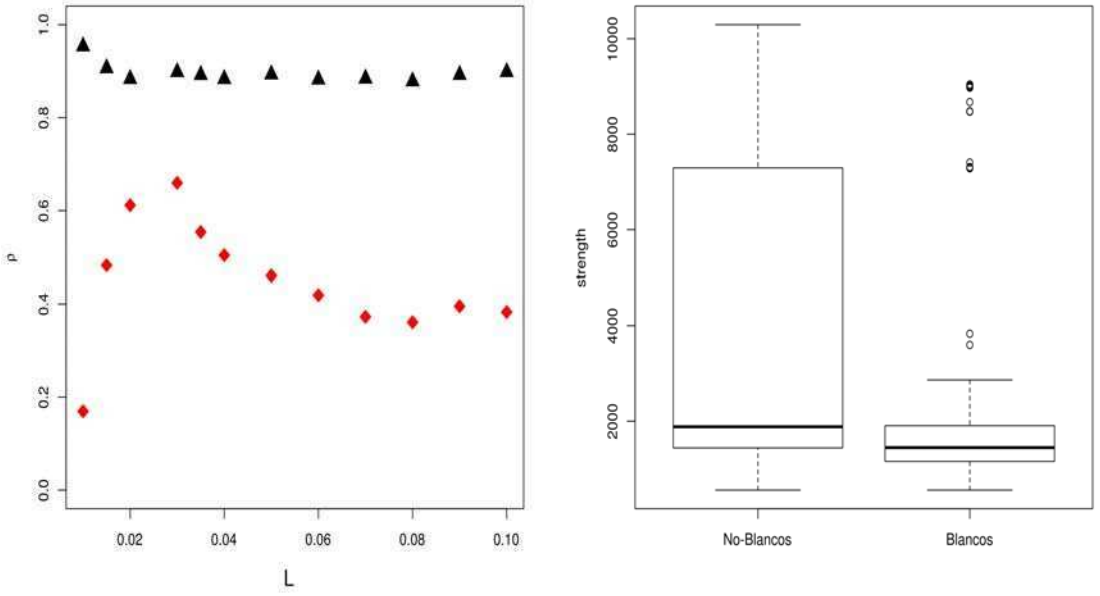
5.5.2. La corrección de grado mejora el poder predictivo de la red

Para entender por qué la corrección de grado incrementa el valor predictivo en las priorizaciones (5.5.1), resulta útil primero considerar la relación entre el puntaje obtenido por cada proteína en los algoritmos de priorización y el *strength* de las mismas en la capa $G(V = V_P, E = E_{PP})$ donde se realiza el proceso de priorización (ver ec. 2.2). En particular nos interesa analizar esta relación para el 10 % de las proteínas que alcanzan mayor puntaje en la priorización, ya

que éstas son las que tienen incidencia en el cálculo de AUC-01. Analicemos en primer lugar las priorizaciones realizadas sobre el organismo M.Musculus. La figura 5.6a consigna la correlación entre el *strength* y el puntaje de priorización alcanzado para proteínas incluidas en las primeras L posiciones de la lista generada (expresada como fracción del tamaño total del genoma considerado). En el caso de la red construida sin corrección de grado, el puntaje alcanzado en la priorización está altamente correlacionado con el strength de las proteínas, independientemente de la longitud de lista considerada (triángulos negros, fig.5.6a). En el caso de la red construida con corrección de grado esta relación no resulta a priori trivial (rombos rojos, fig. 5.6a). La correlación es menor respecto de la priorización sin corrección de grado en todo el rango analizado. Además, para las proteínas con mayor puntaje ($L < 0.03$) la correlación cae abruptamente para la red con corrección de grado mientras que en la red sin corrección esta correlación crece. En particular para L=0.01, es decir para el 1 % de las proteínas mejor puntuadas(134 proteínas), la red con corrección de grado presenta $\rho = 0.17$ mientras que en la red sin corrección se tiene $\rho = 0.95$.

Estos resultados sugieren que las priorizaciones realizadas sobre la red sin corrección de grado estarán sesgadas hacia blancos con una variable topológica particular: el *strength*. Por construcción de la red, esas proteínas son blancos involucrados en nodos de afiliación promiscuos, o bien a grupos de afiliación con alta proporción de blancos de droga. Por otro lado, la figura 5.6b consigna las distribuciones de *strength* para proteínas que son blancos de droga y para aquellas que no lo son en *mmu*. El gráfico se restringe al 10 % de los genes con mejor puntaje en las priorizaciones y el cálculo del strength se realizó en el grafo proyectado sin corrección por grado. Esta figura muestra que los blancos proteicos de *mmu* no tienen típicamente mayor *strength* que las proteínas que no son blanco de droga. Por tanto dado que el método de priorización en esta red correlaciona fuertemente con esta variable topológica, es esperable que tenga una dificultad intrínseca para recuperar blancos que no estén asociados a dominios muy promiscuos, obteniendo bajos valores de auc-01.

Notar que en el caso de T.Cruzi, (ver figura 5.7a y 5.7b) la distribución de *strength* de los blancos es aún más desfavorable respecto del resto de los genes en esa especie. Esto es consistente con el hecho de que el AUC-01 en este caso sea notablemente más bajo que en el caso de *mmu* (prácticamente, un predictor random, AUC-01=0.52, ver tabla 5.5.1.)



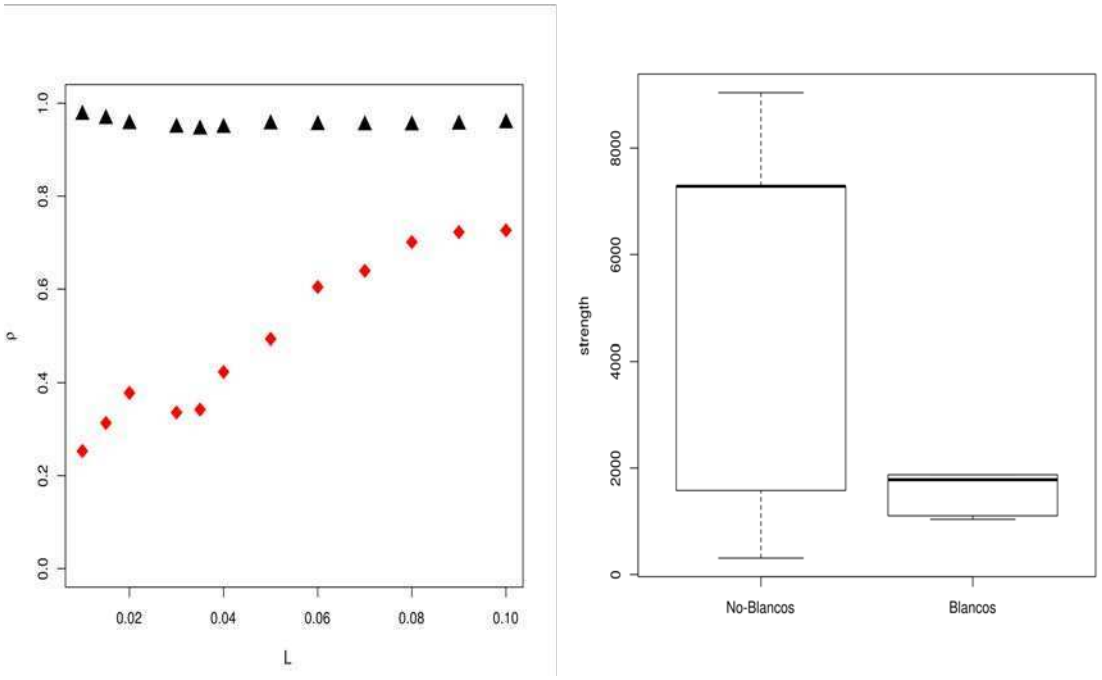
(A) Score vs Strength correlation vs size list

(B) Distribución strength

FIGURA 5.6: Análisis de priorizaciones realizadas con y sin corrección de grado para M.Musculus con el algoritmo VS sobre las redes construidas multicapa construidas. (A). Correlación de Spearman entre el puntaje asignado a cada nodo por el algoritmo *VS* (ver ecuación 2.15) y el correspondiente valor de *strength* en las redes construidas con corrección de grado (rombos rojos) y sin corrección de grado (triángulos negros). Cada valor de correlación se calculó para los primeros L candidatos de la lista de priorización, la cual se expresa en el eje de abscisas como fracción del genoma completo de M.Musculus. (B). Distribución de *strength* para dos categorías de nodos en M.Musculus: aquellos nodos que son blanco de alguna droga (Blancos) y aquellos que no tienen asignada ninguna bioactividad en la red (No-Blancos).

5.5.3. Relevancia de los distintos tipos de nodo de afiliación

Otra pregunta que resulta interesante abordar es si hay alguna clase de nodo de afiliación (Pfam, homólogos o Vías metabólicas) cuya presencia sea más relevante que otra para los procesos de priorización realizados en la red. Para analizar este punto se repitió el procedimiento de priorización quitando alternativamente los distintos tipos de nodos de afiliación. Es decir, se quitaron primero todos los nodos correspondientes a dominios Pfam, y se recalculó el grafo proyectado G_P con y sin corrección por grado, para luego priorizar blancos utilizando ambos algoritmos *VS* y *F.F*. Los correspondientes valores de AUC-01 se reportan en la tabla 5.5, columnas "Sin-Pfam", "Sin-VM", "Sin-Homol", dependiendo de que tipo de nodos de afiliación se hayan eliminado. Además se agrega como referencia los valores de AUC-01 obtenidos utilizando todas las evidencias simultáneamente (es decir los valores reportados en la tabla 5.5.1).



(A) Score vs Strength correlation vs size list

(B) Distribución strength

FIGURA 5.7: Análisis de priorizaciones realizadas con y sin corrección de grado para T.cruzi con el algoritmo VS sobre las redes construidas multicapa construidas. (A). Correlación de Spearman entre el puntaje asignado a cada nodo por el algoritmo *VS* (ver ecuación 2.15) y el correspondiente valor de *strength* en las redes construidas con corrección de grado (rombos rojos) y sin corrección de grado (triángulos negros). Cada valor de correlación se calculó para los primeros L candidatos de la lista de priorización, la cual se expresa en el eje de abscisas como fracción del genoma completo de T.Cruzi. (B). Distribución de *strength* para dos categorías de nodos en T.Cruzi: aquellos nodos que son blanco de alguna droga (Blancos) y aquellos que no tienen asignada ninguna bioactividad en la red (No-Blancos).

		Completo	Sin -Pfam	Sin-VM	Sin-Homol
Método	Especie	**1/K** | s.c	**1/K** | s.c	**1/K** | s.c	**1/K** | s.c
VS	Mus Musculus	**0.72** | 0.64	**0.57** | 0.50	**0.73** | 0.66	**0.65** | 0.63
	T.cruzi	**0.81** | 0.52	**0.73** | 0.53	**0.77** | 0.54	**0.78** | 0.52
FF	Mus Musculus	**0.66** | 0.65	**0.60** | 0.51	**0.67** | 0.64	**0.65** | 0.63
	T.cruzi	**0.72** | 0.52	**0.62** | 0.56	**0.66** | 0.53	**0.75** | 0.58

CUADRO 5.5: Comparación de desempeño para las distintas estrategias utilizadas eliminando alternativamente las distintas clases de nodos de afiliación. La primer columna se presenta a modo de referencia y contiene las priorizaciones realzadas utilizando todos los ndos de afiliación. Las restantes colunmas representan priorizaciones realziadas sobre grafos G_P que fueron construidos omitiendo el tipo de nodo de afiliación que se detalla en cada caso. Notar como cae sistemáticamente el desempeño de ambos algoritmos en ambas especies al quitar los nodos de tipo Pfam.

Cabe destacar a partir del análisis de esta tabla, que al eliminar la información correspondiente a los dominios *Pfam* se obtiene la mayor degradación en el desempeño del algoritmo en

cualquiera de las dos especies. Por lo tanto, la presencia de nodos correspondientes a dominios funcionales *Pfam* resultan claves en el proceso de priorización llevado a cabo en la red cualquiera sea la estrategia de priorización o la especie considerada. Por otro lado, consistentemente con lo visto en la sección precedente, la ausencia de corrección de grado ofrece valores de desempeño inferiores a los obtenidos utilizando la corrección en casi la totalidad de los casos. También aquí el desempeño del algoritmo VS es en casi la totalidad de los casos superior al de FF (en particular observando las priorizaciones realizadas con corrección de grado).

En suma, pese a su simplicidad, el método VS supera o iguala el desempeño del algoritmo FF en los casos analizados. A este respecto podría sugerirse que el hecho de considerar priorizaciones en la red que se expandan más allá de primeros vecinos en la red G_P(como FF lo hace), implica asumir propiedades de transitividad entre nodos de afiliación que no son necesariamente apropiadas para el problema de reposicionamiento de drogas que se está tratando. Para ejemplificar este punto, supongamos que pudiéramos adjudicar al dominio funcional $Pfam_i$ de un blanco V_P conocido la razón que un fármaco sea bioactivo sobre esta proteína. En tal caso por construcción, ninguna proteína que esté a distancia geodésica mayor a 1 de V_P en la red G_P tendrá asociado el dominio $Pfam_i$. Por tanto extender las priorizaciones a segundos, terceros o vecinos de mayor distancia podría resultar conceptualmente incorrecto en un caso de estas características. Basados en las distintas comparaciones y medidas de desempeño hasta aquí realizadas, consideramos en las siguientes secciones sólo priorizaciones basadas en la metodología VS incluyendo la corrección de grado correspondiente.

5.6. Priorización de especies patógenas y validación de literatura

Tal vez la aplicación más interesante y prometedora de las estrategias de priorización analizadas en la sección anterior, es la de proponer nuevos putativos blancos como interesantes casos de estudio.

Con esta finalidad hemos construido dos redes multicapa según el procedimiento descripto en secciones previas, pero considerando todos los blancos disponibles en todos los organismos. La diferencia sustancial entre las dos redes construidas es que una se hizo considerando la corrección de grado en la transformación 5.8 mientras que en la otra red se omitió dicha corrección. En cada caso se utilizó el algoritmos VS para generar una lista de priorización a través de toda la red. Luego en el contexto de un trabajo en colaboración con el consorcio de TDR Targets, junto al Dr. Fernan Aguero y la Dra Paula Magariños, se analizaron los principales candidatos

de tres especies patógenas de interés: Trypanosoma cruzi(TCR), Trypanosoma brucei (TBR) y Leishmania major (LMA), tres parásitos kinetoplasteas. Las 10 mejores proteínas posicionadas en las dos listas resultantes de este ejercicio de priorización (con y sin corrección de grado) se incluyen en la tabla 5.6 y y la tabla 5.7 respectivamente. Un análisis biológico detallado de los candidatos está fuera del alcance de esta tesis. Sin embargo, vale la pena mencionar que varios de estos candidatos pudieron ser validados con una trabajo de curación manual en literatura. Algunas de estas proteínas han sido ya validadas experimentalmente como blancos de droga publicados en trabajos posteriores a la integración de datos de nuestra red, o bien han sido omitidos en los esfuerzos de curación manual de las fuentes de datos empleadas para la construcción de la red multicapa. A continuación se discuten algunos de estos casos interesantes.

La primera proteína en la lista de prioriación de Trypanosoma brucei (la sexta en la lista de Trypanosoma Cruzi) es un receptor inositol 1,4,5-trifosfato. Los receptores de inositol trifosfato son canales de liberación de calcio intracelular que juegan un rol fundamental en la señalización de Ca^{2+} en células [167]. Trabajos recientes en TBR y TCR [168, 169] muestran que este blanco es esencial para el crecimiento y establecimiento infeccioso. La tercer proteína en la lista de priorización de TCR es una fosfatidilinositol 3-quinasa (PI3K). Esta proteína tiene ortólogos en varias especies y 4 parálogos en humanos. Las PI3K pueden dividirse en 3 clases (I-III). La proteína priorizada por nuestro método es una PI3K de clase I [170]. Estas enzimas son inhibidas en concentración nanomolar por la *wortmanina* que se une al sitio de unión de ATP de PI3K. Este sitio se conserva en las tres clases de PI3K lo que posiblemente podría ser indicio de que el fármaco sea activo contra las tres clases. La vía PI3K también ha sido investigada como blanco para la tratamiento de cáncer [170, 171]. Teniendo en cuenta que el método de priorización identifica esta proteína como potencial blanco en parásitos esto puede representar una oportunidad para poner a prueba sobre parásitos potenciales compuestos encontrados en investigación de cáncer. En TCR el tratamiento con *wortmanina*, un inhibidor de PI3K, impide la entrada de parásitos a las células [172, 173]. Más aún, recientemente se caracterizó una PI3K clase III en este parásito y demostrando que que es inhibida por *wortmanina* y el ligando *LY294000* [174].

Otro blanco propuesto por el método de priorización es la proteína *demetilasa 14α lanosterol* (CYP51), tercera en la lista de LMA y quinta proteína en la lista de TBR). Esta proteína representa un interesante caso que sirve tanto para validar el método de priorización propuesto en esta tesis como para señalar fallas en los procesos de curación manual de datos de bioactividades. La enzima CYP51 pertenece a un grupo ortólogo que contiene 72 secuencias, incluyendo secuencias humanas y trypanosomátidas. Esta proteína es un citocromo

P450 que en hongos y protozoos kinetoplástidos cataliza un proceso bioquímico clave en la vía de biosíntesis de ergosterol [175]. La enzima es un blanco conocido y validado para quimioterapia contra TCR y por tanto considerado como semilla en las priorizaciones. No obstante, estudios recientes [176–178] que no estaban presentes en las bases de datos de TDR y ChEMBL utilizados para construir la red, han mostrado la inhibición sobre encimas de TBR y LMA con inhibidores de CYP51. Es decir, estos blancos fueron priorizados bajo la ausencia de bioactividad disponibles contra ellos, y dado que ya existen trabajos experimentales que validan su rol de blancos protéicos, las priorizaciones sirven sólo para identificar las fallas en los esfuerzos de curación manual de las bases de datos de bioactividades disponibles.

Cuadro 5.6: Priorizaciones sobre el grafo G_P^K. Lista de potenciales blancos de droga en tres especies patógeneas consideradas: Lesmania Major (LMA), Trypanosoma Cruzi (TCR) y Trypanosoma Brucei (TBR). La priorización se realizó con una red construida utilizando la corrección de grado (ec. 5.8)

Ranking	**Gene ID (Identificador único)**	**Descripción**
1st (TCR) 5th (LMA)	TcCLB.511277.60, TcCLB.506357.50 LmjF30.2090	Alcohol dehydrogenase
2nd (TCR) 3rd (TBR)	TcCLB.507023.120, TcCLB.511751.120 Tb927.6.3050	Aldehyde dehydrogenase familiy
3rd (TCR)	TcCLB.510167.10, TcCLB.508859.90	Phosphatidylinositol 3-kinase 2
4th (TCR) 6th (TBR) 7th (LMA)	TcCLB.510329.90, Tb927.3.4650 LmjF29.2140	C-8 sterol isomerase
5th (TCR) 8th (TBR) 9th (LMA)	TcCLB.506933.20, Tb11.02.5720, LmjF28.0890	Ribonucleoside-diphosphate reductase large chain
6th (TCR) 1st (TBR)	TcCLB.509461.90, Tb927.8.2770	Inositol 1,4,5-trisphosphate recepto
7th (TCR)	TcCLB.508183.10, TcCLB.508153.170	Ubiquinone biosynthesis protein COQ7 homolog
8th (TCR)	TcCLB.508173.100	Monooxygenase
9th (TCR) 6th (LMA)	TcCLB.511647.4, LmjF16.0590	Carbamoyl-phosphate synthase
10th (TCR)	TcCLB.509943.20	cAMP specific phosphodiesterase
2nd (TBR)	Tb927.6.4210	Aldehyde dehydrogenase, putative (ALDH)
4th (TBR)	Tb11.02.3040	Aldo/keto reductase
5th (TBR) 3rd (LMA)	Tb11.02.4080, LmjF11.1100	Sterol 14-alpha-demethylase (CYP51)
7th (TBR) 8th (LMA)	Tb927.10.2010, Tb927.10.2020, LmjF21.0240, LmjF21.0250	Hexokinase
9th (TBR) 4th (LMA)	Tb927.7.5480, LmjF06.0860	Dihydrofolate reductase thymidylate synthase (DHFR TS)
10th (TBR) 10th (LMA)	Tb927.8.7100, LmjF31.2970	Acetyl CoA carboxylase
1st (LMA)	LmjF23.0360	NADP-dependent alcohol dehydrogenase
2nd (LMA)	LmjF25.1120	Aldehyde dehydrogenase mitochorndrial (ALDH2)

CUADRO 5.7: Priorizaciones sobre el grafo G_P^{sc}.Lista de potenciales blancos de drogas en tres especies patógeneas consideradas: Lesmania Major (LMA), Trypanosoma Cruzi (TCR) y Trypanosoma Brucei (TBR). La priorización se realizó con una red construida sin utilizar la corrección de grado

Ranking	Identificador único	Descripción	Grupo	Familia
1st (TCR) 7th (TBR) 1st (LMA)	Tcr.40411, tcr.45645, tbr.14943, tbr.15196, tbr.20396, lma.21622, lma.23193, lma.23398	protein kinase A catalytic subunit isoforms 1 and 2	AGC	PKA
2nd (TCR) 8th (TBR)	Tcr.53544, tbr.13606	mitogen-activated protein kinase 3, putative	CMGC	MAPK
3rd (TCR), 9th (TBR), 2nd (LMA)	Tcr.50210, tbr.16775, lma.29160	casein kinase II, putative	Other	CK2
4th (TCR), 3rd (LMA)	Tcr.31102, tcr.38167, tcr.48086, lma.27570, lma.27653	mitogen activated protein kinase, putative	CMGC	MAPK
5th (TCR), 4th (LMA)	Tcr.54789, lma.27232	protein kinase, putative	Other	Aurora
6th (TCR), 5th (LMA)	Tcr.31281, tcr.32580, tcr.35570, lma.21536	casein kinase, putative	CK1	CK1
7th (TCR)	Tcr.45558	protein kinase, putative;	Other	ULK
8th (TCR), 7th (LMA)	Tcr.43421, lma.23730	mitogen-activated protein kinase, putative	CMGC	MAPK
9th (TCR), 8th (LMA)	Tcr.39469, lma.21051	serine/arginine-rich protein specific kinase SRPK, putative; serine/threonine-protein kinase	CMGC	SRPK
10th (TCR), 9th (LMA)	Tcr.35784, tcr.38523, lma.26724	protein kinase, putative	CMGC	MAPK
1st (TBR)	Tbr.14801	protein kinase, putative	STE	STE11
2nd (TBR)	tbr.10954	protein kinase, putative, NEK family, HsNEK1-like	Other	NEK

Continua en la página siguiente

Cuadro 5.7 – viene de la página anterior

Ranking	Gene ID (Identificador único)	Descripción	Grupo	Familia
2nd (TBR)	tbr.15022	serine/threonine-protein kinase, putative	Other	NEK
2nd (TBR)	tbr.15541	protein kinase, putative	-	-
2nd (TBR)	tbr.18941	serine/threonine-protein kinase NrkA	Other	NEK
2nd (TBR)	Tbr.321685, tbr.321748	protein kinase, putative,NEK family, HsNEK1-like	Other	NEK
3rd (TBR)	tbr.19596	protein kinase, putative,serine/threonine protein kinase, putative	STE	STE11
4th (TBR)	tbr.18292	protein kinase, putative	STE	-
5th (TBR)	tbr.12066	protein kinase,zinc finger protein kinase	AGC	-
6th (TBR)	tbr.10341	protein kinase, putative	CAMK	-
8th (TBR)	Tbr.15920	rac serine-threonine kinase, putative,protein kinase, putative	AGC	-
10th (TBR)	tbr.12774	protein kinase, putative	STE	STE11
6th (LMA)	lma.21108	mitogen-activated protein kinase kinase 2	Other	ULK
7th (LMA)	lma.26156	protein kinase, putative	Other	PEK
10th (LMA)	lma.23701	serine/threonine-protein kinase Nek3, putative	Other	NEK
10th (LMA)	lma.24345	serine/threonine-protein kinase, putative	Other	NEK
10th (LMA)	lma.25051	casein kinase 1 isoform 2, putative	?	?
10th (LMA)	lma.25629	casein kinase II, alpha chain, putative	Other	CK2

Continua en la página siguiente

Cuadro 5.7 – viene de la página anterior

Ranking	Gene ID (Identificador único)	Descripción	Grupo	Familia
10th (LMA)	lma.27151	serine/threonine-protein kinase, putative	Other	NEK
10th (LMA)	lma.27427	serine/threonine-protein kinase, putative	Other	Aurora
10th (LMA)	lma.28321	protein kinase, putative	Other	PEK
10th (LMA)	lma.29063	protein kinase, putative	Other	VPS15

En cuanto a la segunda clase de priorización de proteínas, aquella realizada sin corrección de grado, los blancos obtenidos en las listas de priorización son mayoritariamente quinasas. El primer blanco obtenido en TCR ha sido mostrado que interactúa y fosforila varias proteínas del parásito [179], incluyendo algunos de la familia *transialidasa* [180]. La transfección con una construcción de DNA superenrollado que contiene PKI (inhibidor de PKA) mata tripanosomátidos en etapa epimastigotes (mostrado con un experimento genético) mientras que el tratamiento con H89, un compuesto que también es inhibidor de PKA, mata el 98 % de los parásitos dentro de las 48 h (mostrado con un experimento farmacológico) [179] La proteína de TCR obtenida en el décimo lugar de la lista de priorización, TcMAPK2, ya ha sido estudiada y caracterizada. Esta proteína no puede ser inhibida con FR180204, un inhibidor en mamíferos de la proteína ERK2. Esto sugiere que TcMAPK2 puede resultar una potencial diana de fármacos ya que difiere significativamente de la proteína de mamífero [180]. Los blancos en quinta y sexta posición en las listas de LMA y TCR respectivamente, son la misma caseína quinasa I isoforma 2. Esta proteína ha sido probada ser blanco para 4 inhibidores en LMA [181] . En este trabajo se demostró que estos 4 compuestos también inhiben el crecimiento de los cultivos de promastigotes de LMA y tripomastigotes de TBR. En otro trabajo, la proteína de LMA se encontró inhibida por tres *piridinas 2,3-diarylimidazo[1,2-a]* [182]. También se estudió en TCR [183, 184] donde se encontró ligada al compuesto *purvalanol B*.

5.7. Priorización de blancos para drogas huérfanas.

5.7.1. Estrategia de priorización y validación in-sílico

En la búsqueda de nuevos fármacos es usual realizar estudios fenotípicos de alto rendimiento llevados a cabo en organismos completos o células en cultivo. Esta es una buena estrategia para filtrar compuestos e identificar candidatos razonables. Así mismo, para desarrollar estos fármacos resultaría mucho más ventajoso conocer los blancos de acción del compuesto bajo estudio y obtener así una mejor comprensión del mecanismo de acción del fármaco. En lo sucesivo, cuando un fármaco muestre actividad sobre un dado organismo pero sus blancos de acción resulten desconocidos, diremos que estamos en presencia de un compuesto *huérfano*
En esta sección, se busca obtener para un dado compuesto *huérfano* bajo estudio, una lista de potenciales blancos de acción. Para ello, se propone utilizar la información embebida en la red multicapa construida en combinación con los algoritmos de priorización descriptos en secciones previas (en particular con el uso de VS).
Cada compuesto *huérfano* estará representado por un nodo V_{D_i} de la *capa-DD* del grafo $G(V = \{V_D, V_P\}, E = \{E_{DD}, E_{DP}, E_{PP}\})$ y carecerá de conexiones directas hacia la *capa-PP* por medio de conexiones de tipo E_{DP} (es decir, carece de bioactividades conocidas). Para un compuesto de este tipo se quiere obtener una lista de proteínas V_{P*} de la red, que sean potenciales blancos de acción.
Partiendo del compuesto *huérfano* V_{D_i} se identifica el conjunto de moléculas químicamente similares, es decir, el conjunto de nodos que son primeros vecinos V_{D_j} en la *capa-DD*. Esto es, para cada nodo V_{D_j} existe una conexión E_{DD} que lo une de forma directa al nodo V_{D_i} de interés. Luego, consideramos el conjunto de proteínas en la *capa-PP* V_{P_k}, asociados a algún nodo del conjunto V_{D_j} a través de bioactividades E_{DP}. Los nodos en el conjunto V_{P_k} fueron utilizados como semillas para el procedimiento de priorización mediante el esquema VS en la *capa-PP*. Cabe destacar que cada semilla tiene distinto nivel de relevancia inicial en base a la intensidad y cantidad de conexiones que lo unen a los compuestos del conjunto V_{D_j}. La intensidad de la k-ésima semilla V_{P_k}, en relación a la droga *huérfana* V_{D_i} que notaremos $I(V_{P_k}|_i)$ queda definida según:

$$I(V_{P_k}|_i) = \sum_{e_{jk} \in E_{DP}} w_{ij} \tag{5.9}$$

donde e_{jk} es el conjunto de conexiones que unen la proteína V_{P_k} con la j-ésima droga en el conjunto V_{D_j}, y w_{ij} pertenece al conjunto de conexiones $\in E_{DD}$ y es el peso correspondiente a la

similitud de Tanimoto o relación de subestructura entre los compuestos V_{D_j} y la droga *huérfana* V_{D_i}.

Para validar esta estrategia se consideró un conjunto de 1.000 moléculas V_{D_i} seleccionadas al azar de un conjunto de 10^5 nodos V_D con exactamente un blanco conocido en la red. La idea general es, para cada uno de estos 1000 compuestos eliminar su bioactividad conocida $e_{ik} \in E_{DP}$ trasformándolos artificialmente en compuestos *huérfanos*, y poder así evaluar la capacidad de nuestro algoritmo para recuperar esta conexión en las correspondientes listas de priorización. Para cada uno de estos compuestos *artificialmente huérfanos* V_{D_i}, se obtiene una lista de putativos blancos V_P^* que denominaremos L_i. Notar que al igual que en los casos de drogas *huérfanas* reales, conocemos la especie de interés donde buscamos el blanco de acción (en el caso de esta validación, lo conocemos por haber borrado intencionalmente la conexión entre V_{D_i} y su blanco V_P). Por lo tanto, es posible calcular dos observables que permitan evaluar precisión en nuestro ejercicio de validación. Primero, podemos calcular la posición global de la proteína buscada V_P en nuestra lista L_i, la cual denotaremos como AR (absolute ranking). En segundo lugar, podemos considerar un subconjunto $L_i^{\#}$ de L_i restringido a la especie de interés (la especie correspondiente a la proteína V_P buscada), y calcular allí la posición del blanco de interés que denotaremos como LR (local ranking). Notar que por construcción, $LR \geq AR$.

El gráfico de la figura 5.8, resume los resultados obtenidos para el ejercicio de validación propuesto. Dadas las 1000 listas L_i podemos fijar un nivel de profundidad o longitud, λ, y calcular la cantidad de oportunidades en las que se pudo identificar el blanco buscado en cada lista (RP). El panel izquierdo de la figura 5.8 consigna en línea de trazo discontinuo (escala izquierda), el valor de RP como como función del nivel de profundidad λ considerado. Evidentemente la relación $RP(\lambda)$ debe ser una función monótona creciente. También se observa que la tasa de crecimiento de $RP(\lambda)$ disminuye con el aumento de λ. Por lo tanto, para caracterizar el desempeño del método propuesto buscaremos obtener una profundiad característica λ^* del mismo, a partir de analizar la tasa de variación de $RP(\lambda)$ que definimos según:

$$RPR(\lambda) := \frac{\partial}{\partial \lambda} RP \tag{5.10}$$

El panel A de la figura 5.8 consigna en línea de trazo azul continuo (escala derecha) la función $RPR(\lambda)$. Esta función tiene un comportamiento asintótico cercano a cero, que se puede interpretar como un punto donde los incrementos en λ no se ven reflejados en una ganancia significativa de RP. Por lo tanto, estimaremos un umbral óptimo λ^* donde la tasa de variación RPR sea significativamente mayor al nivel asintótico de referencia. Para definir ese valor, podemos

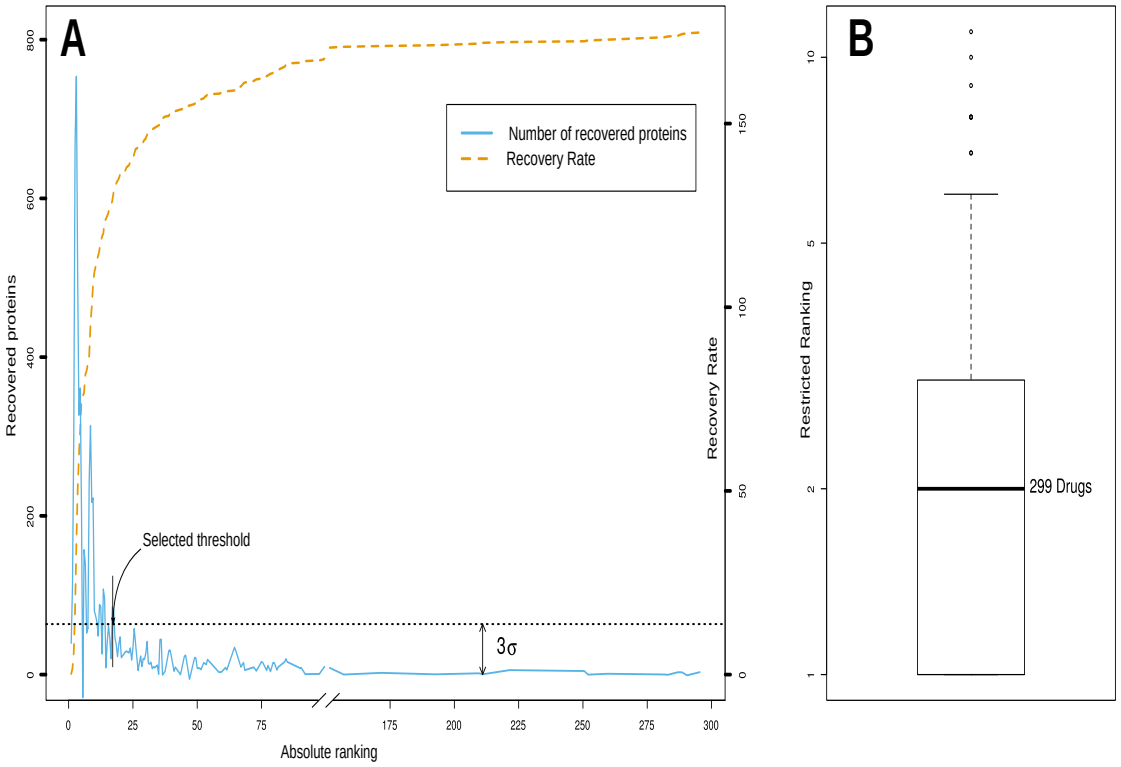


FIGURA 5.8: Resultados de validación sobre las 1000 drogas artificialmente huérfanas generadas. Panel A: Cantidad de blancos recuperados *RP* (escala izquierda, línea discontinua naranja) y tasa de recuperación *RPR* definida según la ec. 5.10. (escala derecha, línea continua azul) ambas, como función del nivel de profundidad en las listas L_i consideradas (λ). La cota punteada representa 3 desvíos del nivel de ruido medio, estmado del comportamiénto asintótico de la función $RPR(\lambda)$. B: Distribución de la posición del blanco buscado en el subconjunto $L_i^{\#}$ restringido al organismo específico de interés en cada caso, *LR* (ver texto). En 598 casos se ha podido recuperar el blanco buscado en un nivel de profundidad $\lambda \leq \lambda^* = 17$. Notablemente en el 50 % de los casos, el blanco buscado se halló en primer o segunda posición de *LR*.

interpretar a *RPR* como una variable aleatoria y calcular el nivel donde RPR esté a 3 desvíos (3σ) del valor asintótico. Finalmente, tomaremos λ^* según:

$$\lambda^* := \max_{\lambda}\{RPR(\lambda) \leq RPR_\infty + 3\sigma\} \tag{5.11}$$

donde RPR_∞ representa el valor asintótico de $RPR(\lambda)$ y σ es la desviación estándar de la variable aleatoria *RPR*.

La longitud de lista definida en la ec.5.11 equivale a $\lambda^* = 17$. A continuación restringimos las 1000 listas a esta longitud y calculamos el valor de *LR* para el blanco V_P buscado en cada caso . El panel B de la figura 5.8 consigna la distribución de *LR*, para el subconjunto de 598 casos donde se halló el blanco buscado a una profundidad menor a λ^*. Notablemente se observa que el 50 % de los casos (299 blancos), el blanco buscado se halla en la primer o segunda posición

LR (relativa a la especie de interés en cada caso). y más del 96 % fueron clasificados dentro de los 6 primeros candidatos. Este resultado tiene un notable valor desde el punto de vista experimental, dado que estos resultados involucran longitudes de listas completamente factibles de analizar en estudios experimentales de bioactividades para los cuales la tasa de éxito alcanzada resulta aceptable. La figura 5.9 esquematiza los dos posibles mecanismos de priorización por

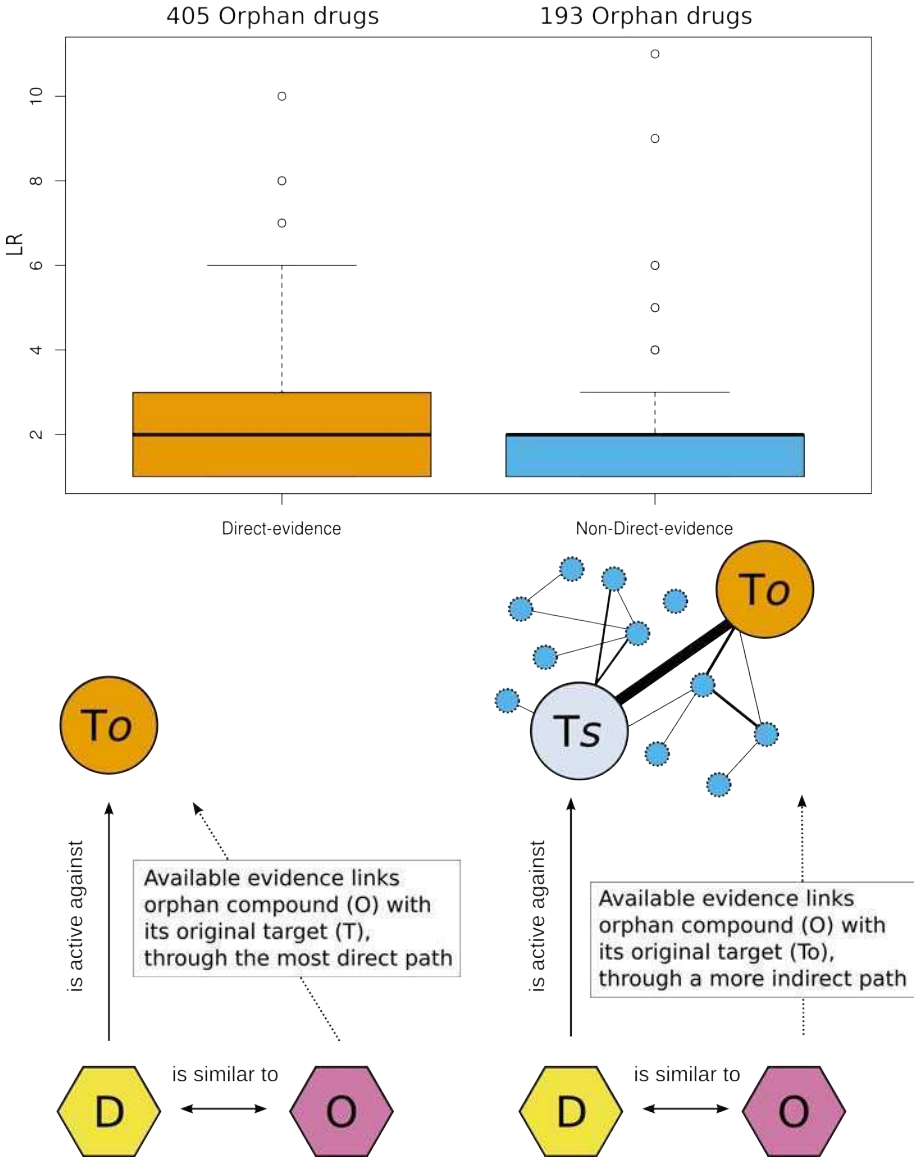


FIGURA 5.9: Inferencia de blancos para compuestos huérfanos. Vista esquemática de diferentes formas en la cual el algoritmo VS puede encontrar blancos correctos para un compuesto *artificialmente huérfano*. O := Compuesto artificialmente huérfano; D:= Compuesto químicamente similar a O, con bioactividad probada; To:= blanco conocido del compuesto artificialmente huérfano O; Ts= Blanco del compuesto D químicamente similar a O. Notar que, Ts y To deben compartir algún nodo de afiliación (PFam, Ortologo o V.Metabólica). Las flechas representan conexiones de tipo E_{DD}, E_{DP} o E_{PP}. Las flechas de trazo punteado representan conexiones E_{DP} removidas en el proceso de validación. Los paneles superiores consignan la distribución estadística de la posición de To (*LR*) en las listas de priorización L_i.

los cuales podemos recuperar el blanco de un compuesto huérfano. La primera es a través de un camino relativamente corto en la red, (panel izquierdo figura 5.9), cuando hay una conexión directa entre algún compuesto bioactivo en la vecindad de la droga huérfana. Es decir cuando algunos de los primeros vecinos directos de la droga huérfana V_{D_j} y el blanco buscado están conectados. Este mecanismo de acción se dió en el 67 % de los casos (405 de los 598 blancos recuperados). No obstante, los restantes 193 blancos recuperados (33 %) carecen de de enlaces directos a moléculas en la vecindad de la droga huérfana, En estos casos, el blanco recuperado se localiza exclusivamente por las relaciones impuestas en la *capa-pp*, es decir por dominios PFam, Ortólogos y vías metabólicas en conjunto al algoritmo VS implementado (panel derecho, Figura 5.9). Estos resultados muestran la utilidad de la metodología basada en la red multicapa para revelar blancos correctos con alta especificidad en la ausencia de conexiones directas de bioactividad, sugiriendo que la metodología planteada puede resultar de gran utilidad para proponer estudios experimentales sobre compuestos huérfanos.

5.7.2. Algunos casos interesantes en organismos patógenos

Como caso de estudio, se utilizó la red para inferir blancos en 19.124 compuestos *huérfanos* de Plasmodium falciparum, es decir compustos con actividad probada contra el organismo patógeno mencionado, pero en todos los casos, los ensayos correspondientes no reportan ningún blanco específico de acción para el compuesto. La mayor parte de estos compuestos se derivan de ensayos masivos de alto rendimiento contra P. falciparum [136–138]. Basados en la estrategia de priorización presentada en la sección anterior se han sugerido candidatos para 5903 de estos compuestos.

Luego en el contexto de un trabajo en colaboración con el consorcio de TDR Targets, junto al Dr. Fernan Aguero y la Dra Paula Magariños, se analizaron algunos casos que se presentan a continuación.

Un ejemplo de interacción fármaco-diana propuesto se muestra en la Figura 5.10. En el panel superior de esta figura se mustra el compuesto huérfano para el que se ha buscado blancos de acción,una *benzotiazolina* que se sabe activa contra Plasmodium falciparum (cepa W2). El mecanismo de acción de este compuesto es actualmente desconocido. En la red multicapa construida, el esquema de conectividad del compuesto a través de la capa $capa - DD$ y las bioactividades E_{DP} conduce a la proteína *N-miristoiltransferasa* de la especie *C. albicans*. Esta

enzima cataliza la *N-Miristoilación* de proteínas, en el que se añade una molécula de miristato (ácido graso saturado 14-C) a la N-terminal de un residuo de glicina en diana protéicas específicas [185, 186]. Esta modificación post-traduccional afecta a la asociación de la proteína myristoylated a las membranas[185]. Un análisis detallado de la literatura disponible muestra que esta proteína es de hecho un blanco prometedor para el desarrollo de nuevos antimaláricos [187–189]. Sin embargo, a pesar de que varios compuestos de benzotiazol han sido probados contra la enzima de Plasmodium [189], ninguno de los compuestos presentados en esa publicación fueron parte de nuestro conjunto de datos, y por lo tanto no se incluyeron en la red multicapa construida.

Otro caso interesante es el compuesto huérfano (TDR Targets ID 599594) que ha demostrado

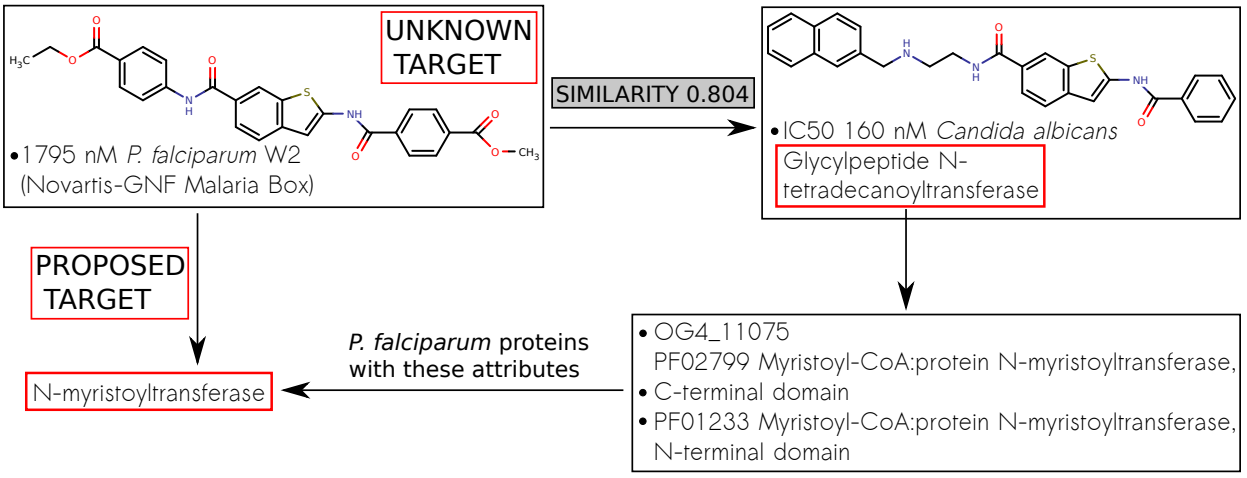


FIGURA 5.10: Blancos sugeridos para compuestos huérfanos. El compuesto mostrado en el panel superior izquierdo (TDR Targets ID 606689, ChEMBL ID 688510) es un compuesto huérfano cuyo blanco de acción permanece desconocido. Este compuesto ha mostrado actividad contra Plasmodium Falciparum. El compuesto mostrado a su derecha, químicamente similar a éste (Coeficiente de Tanimoto = 0.804), resulta activo contra el péptido N-myristoyltransferase de *Candida albicans* [190], el cual pertenece al mismo grupo de otrólogos y comparte dos dominios funcionales Pfam con la N-myristoyltransferase de P. Falciparum, el que fue sugerido como potencial blanco de del compuesto huérfano bajo estudio.

ser activo a $2\mu M$ contra la cepa salvaje 3d7 de *P.Falciparum* y la cepa multirresistente Dd2 (inhibición del crecimiento al 100 % y 97 % respectivamente). En la red multicapa este compuesto está conectado con otros fármacos activos a diferentes niveles de similitud estructural, pero en todos los casos con bioactividades sobre péptidos bacterianos deformilase, y contienen el grupo ácido hidroxámico funcional que es esencial para la actividad contra estos blancos [191]. El inhibidor más frecuentemente utilizado para péptidos deformilase, es la *Actinonin*, que también ha demostrado ser activa contra P. falciparum [192]. Aunque queda pendiente probar si estos compuestos son activos contra la malaria humana en ensayos *in-vivo*, los ejemplos señalados sirven

para mostrar que las conexiones sugeridas por el modelo de priorización propuesto otorga resultados coherentes. Otros cinco compuestos huérfanos fueron propuestos en las priorizaciones realizadas como potencialmente activos sobre la proteína *fabl* (enoyl-acyl carrier reductase). Esta enzima está implicada en biosíntesis de ácidos grasos de tipo II, una vía que es de esencial para el correcto desarrollo del hígado en parásitos [193]. La proteína fabI ha sido validada como diana farmacológica para antibacterianos y antimaláricos, tales como *triclosan*, un fármaco que inhibe esta enzima en varias especies, incluyendo *Escherichia coli*, *Mycobacterium Tuberculosis* ,*Staphylococcus Aureus* y *P. Falciparum*. Varios compuestos han sido probados recientemente como potenciales inhibidores de este blanco en P. falciparum [193–196] y en otros parásitos [197]. Sin embargo ninguno de ellos son los 5 que se proponen en este trabajo, haciendo de éstos candidatos originales para nuevos experimentos.

Algunos compuestos huérfanos se predicen en nuestras priorizaciones actuando a través de varias enzimas aminoacil-ARNt sintetasa, en particular *isoleucil*,*metionina*, *fenilalanilo* y sintetasas *tirosil-ARNt*. Las vías de traducción de proteínas es un blanco validado para compuestos anti-infecciosos [198]. Las aminoacil-tRNA sintetasas catalizan la unión de aminoácidos a sus correspondientes ARNt (ARN de transferencia). Por otro lado, la *mupirocina*, es un fármaco antibacteriano está actualmente en uso contra *Staphylococcus Aureus* y su modo de acción es mediante la inhibición de la sintetasa isoleucil-tRNA [198]. Se ha demostrado que la *mupirocina* inhibe el crecimiento de la etapa sanguínea de P. falciparum, y que los parásitos Plasmodium resistentes a *mupirocina* tienen mutaciones en la enzima apicoplastica *IleRSs* [199]. Estos resultados sugieren que la *isoleucil-tRNA* sintetasa propuesta puede ser un blanco interesante en Plasmodium falciparum. Con respecto a la *metionina*, y a la *fenilalanilo tirosilo-ARNt* sintetasas, a nuestro conocimiento no se han estudiado en P. falciparum como potenciales blancos de drogas, por lo que podría resultar interesante probar experimentalmeente si estos compuestos huérfanos están actuando a través de las proteínas propuestas.

5.8. Discusión

El trabajo propuesto en este capítulo resume una forma original de integrar una amplia gama de datos quimioinformáticos. Los mismos abarcan datos de similitud estructural entre compuestos químicos, una amplia gama ensayos experimentales entre compuestos y proteínas (bioactividades), así como también distintos criterios de similitud entre proteínas. Estos últimos incluyen relaciones de homología, presencia de dominios estructurales y funcionales así como también la participación en diversas vías metabólicas. Encontrar una forma conveniente de

integrar una cantidad tan numerosa y variada de datos involucra en sí mismo un desafío. En particular, si la integración se hace mediante el enfoque de redes multicapa, se hace evidente que existen múltiples formas de combinar la información disponible, y cada uno de ellos conduce a un modelo de red diferente. La integración de la información disponible sobre relaciones entre proteínas fue pensada originalmente en dos capas con distintos tipos de nodos V_P (proteínas) y V_B (nodos de afiliación). La convergencia de estas dos capas a un único grafo de proteínas involucra algún tipo de proyección bipartita. Un punto sobre el que se ha trabajado a lo largo del capítulo es sobre el tipo de proyección bipartita realizada, ya que de ésta dependerán las características finales de la red multicapa resultante. En este capítulo se propuso una generalización de la proyección propuesta por Zhou y colaboradores [200], de manera que la misma pueda considerar una importancia relativa ente los nodos de afiliación compartidos por las proteínas de la red. Este modelo nos permite considerar ya desde la proyección bipartita la importancia que cada dominio funcional, via metabólica o grupo de homólogos tiene para el problema de reposicionamiento de drogas, y constituye en sí mismo un aporte original del trabajo. Para evaluar la importancia relativa de estos nodos de afiliación realizamos una prueba de enriquecimiento que considera la proporción de proteínas de cada nodo de afiliación que resultan blancos de droga conocidos. Los valores de significancia estadística obtenidos para los distintos nodos de afiliación deben llevarse a un valor de peso relativo entre éstos. La forma funcional planteada se consigna en la ecuación 5.8, y depende básicamente de un parámetro libre α y de la inclusión o no de un factor de peso adicional ($\frac{1}{K_i}$) que penaliza el grado de promiscuidad de cada nodo. Como se ha mostrado, la construcción de la red multicapa, y en particular el plano de proteínas G_P depende tanto de α como del factor de corrección $\frac{1}{K_i}$. En adelante referiremos al grafo generado sin corrección de grado G_P^{sc}, y al grafo con corrección lo denotaremos G_P^k. Hemos visto mediante un procedimiento de validación cruzada de 10 iteraciones que la construcción de la red es robusta frente a la elección de α seleccionada en un rango de valores $\alpha \in [0.2, 1]$. Más importante aún, se ha mostrado en la figura 5.4, que la proyección bipartita propuesta presenta mejoras estadísticamente significativas ($pv \lesssim 10^{-24}$) respecto a la proyección original de Zhou [200] que se corresponde con la elección de $\alpha = 0$.

Por otro lado se ha mostrado que la inclusión o no del factor de corrección $\frac{1}{K_i}$ resulta determinante en las propiedades emergentes de la red G_P. En particular, los procesos de priorización de blancos de drogas realzados sobre las capas G_P^k y G_P^{sc} muestran resultados diametralmente opuestos en cuanto a la capacidad que presentan para detectar posibles blancos de drogas (ver tabla 5.5.1 y tabla 5.5). Mientras que en algunos casos G_P^k provee altos valores predictivos

(AUC-01 $\sim$ 0.8) la red G_P^{sc} provee valores de desempeño que no se diferencian de un predictor aleatorio (AUC-01$\sim$ 0.5). En este punto, una mirada superficial bastaría para descartar por completo la construcción de la red sin corrección de grado G_P^{sc}. No obstante, una análisis crítico y en mayor profundidad permite observar que en el caso de G_P^{sc} la lista de proteínas propuesta como potenciales candidatos de la red tiene una enorme correlación con una propiedad topológica emergente de la red G_P^{sc}: el grado generalizado de sus nodos (*strength*). El hecho de que la red G_P^{sc} alcance valores de $AUC-01$ cercanos a un predictor aleatorio o no, (ver diferencias de AUC-01 en tabla 5.5.1 para *T.Cruzi* y *M.Musculus*) depende exclusivamente de si los blancos de droga en cada especie son nodos en G_P^{sc} con alto o bajo valor de $Strength$ en la red G_P construida (ver figuras 5.6b y 5.7b). A este respecto, el estudio de curación manual con material bibliográfico reciente u omitido accidentalmente en la construcción de la red, muestra que las priorizaciones en G_P^{sc} son mayoritariamente proteínas con dominios *kinasas*, los cuales resultan usualmente muy promiscuos en la red, en el sentido que tienen una gran cantidad de proteínas anotadas. Estas proteínas con dominios *kinasas* en el dominio del problema de búsqueda de blancos de droga, se muestran como candidatos que resultan generalmente razonables y atractivos casos de estudio. Esto sugiere que la red G_P^{sc} tiene por si misma propiedades emergentes no triviales relacionadas al problema de búsqueda de potenciales blancos de droga. Esto resulta lógico si consideremos que el *strength* de las proteínas está vinculado a la promiscuidad y significancia estadística de los nodos de afiliación a los que pertenecen (ec. 5.4) y podría entonces estar reflejando un sesgo en el tipo de evidencia acumulada hacia el estudio de proteínas con dominios kinasas. Esta característica de la red se desprende del tipo de proyección bipartita planteada, dado que la misma contempla para la construcción de G_P^{sc} datos de bioactividad en el cálculo de los pesos relativos de los distintos nodos de afiliación, lo que resulta en sí mismo otra ventaja adicional del modelo propuesto.

5.9. Conclusiones

El modelo de red multicapa propuesto en esta tesis proporciona una forma original de abordar problemas de búsqueda de nuevos blancos proteicos y el reposicionamiento de fármacos existentes. En particular, proporciona una manera eficiente de integrar grandes conjuntos de datos quimiogenómicos de muy variada naturaleza. A su vez es posible mediante el uso de estrategias de priorización, sacar provecho de los distintos patrones de conectividad para abordar dos problemáticas biológicas de relevante interés. En primer lugar, el enfoque ofrece una forma original de identificar potenciales candidatos de droga en especies con escasa o nula cantidad de

datos de bioactividades integrando información disponible en organismos más estudiados. En segundo lugar, es posible proponer listas reducidas de potenciales blancos proteicos para una droga específica de interés. En particular se hizo énfasis en el estudio de compuestos *huérfanos*, es decir, compuestos con algún tipo de actividad probada sobre un dado organismo, pero cuyo mecanismo y blanco de acción permanecen desconocidos. En ambos casos, se mostró haciendo uso de validaciones computacionales, así como por medio de una minuciosa búsqueda de curación manual en literatura, que las listas de potenciales blancos proveen candidatos razonables y con gran potencial para guiar nuevos ensayos experimentales.

Este resultado es particularmente importante en el caso de las enfermedades tropicales desatendidas, dado que es posible guiar el reposicionamiento de drogas a partir de esfuerzos en organismos modelo, abaratando enormemente los costos y tiempos de investigación farmacológica.

Capítulo 6

Conclusiones

En la primer parte de esta tesis se analizó y caracterizó una red de interacción de proteínas (PIN) estudiando en que medida la estructura topológica y modular correlaciona con grupos de proteínas que llevan a cabo funciones biológicas específicas. Estudiar la conformación y calidad de los experimentos que dan lugar a la construcción de la PIN considerada resulta de fundamental importancia, dado que las conclusiones que pueden extraerse usualmente son sensibles a la calidad de los datos originales. Por lo tanto, se comenzó realizando un análisis detallado de la composición de la PIN utilizada (capítulo 3), y se mostró que la red considerada provee interacciones soportadas por múltiples evidencias, con al menos un experimento de mediana o alta calidad para cada interacción. Estos resultados sugieren que la PIN empleada reduce la tasa de falsos positivos en las interacciones reportadas.

Se mostró también que las aristas soportadas por mayor cantidad de ensayos experimentales tienen tendencia a formar triángulos en la PIN, incrementando la fiabilidad sobre este tipo de patrones de conectividad de la red. Se mostró también la existencia de correlaciones de segundo orden o superiores que diferencian la estructura topológica de la PIN de los distintos modelos de redes aleatorias considerados. En particular, la PIN utilizada presenta una cola pesada en su distribución de grado, una gran cantidad de nodos de bajo grado y alto betweenness y otros con elevado coeficiente de agrupamiento. Además se mostró explícitamente cómo los patrones de conectividad de segundo orden, cuantificados mediante el estudio del grado medio a primeros vecinos, presentan comportamientos que no pueden ser explicados mediante un modelo configuracional de idéntica distribución de grado. En suma, la PIN considerada está soportada por vasta información experimental y presenta propiedades colectivas emergentes que denotan la existencia de patrones de organización global subyacentes.

Una vez caracterizada esta PIN, en el capítulo 4 se analizaron relaciones existentes entre

la estructura modular y grupos funcionales de proteínas de interés: por un lado proteínas asociadas a procesos de envejecimiento celular y por otro lado grupos de proteínas involucrados en distintas secciones de vías de señalización celular. En particular se hizo especial énfasis en comprender cómo puede afectar el uso de distintas técnicas de reconocimiento de estructura modular a los análisis y conclusiones biológicas subsecuentes. Se consideraron dos algoritmos de detección de comunidades ampliamente reconocidos y utilizados, *Infomap* que se basa en conceptos de teoría de información y *CNM* que optimiza de forma directa la modularidad Q de la red. Ambos algoritmos basados en principios funcionales cualitativamente diferentes son capaces de detectar particiones con similares niveles de modularidad, pero que difieren radicalmente en el nivel de granularidad provisto para analizar la PIN.

No obstante, ambas particiones provistas por los algoritmos son compatibles en el sentido que las estructuras de mayor tamaño halladas con *CNM* se subdividen en estructuras más pequeñas en la descripción provista por *Infomap*. En otras palabras, al nivel de resolución provisto por *Infomap* se generan superficies internas en las estructuras de mayor tamaño de *CNM*. Esto se refleja en una tendencia general de los nodos de la red a aumentar el grado de *participación* observado y por lo tanto a una diferente distribución en los correspondientes roles cartográficos. En particular, los roles cartográficos de alta participación, *kinless* (R4) y *kinless hubs* (R7) se observan poblados sólo bajo la prescripción de *Infomap*. Es importante destacar que este comportamiento no es propio de la PIN ni del organismo considerado, ya que el mismo se observa en otras redes de interacción de proteínas en levaduras empleadas en bibliografía reciente. Este hecho ciertamente relativiza la afirmación original de Guimera respecto a la ausencia de nodos *kinless* en redes reales y hace evidente que tal afirmación es consecuencia directa de la metodología de agrupamiento utilizada y no una característica intrínseca de las redes complejas bajo estudio.

Por otro lado, la discrepancia en los distintos niveles de resolución implícitos a cada algoritmo tiene consecuencias directas en los niveles de congruencia biológica alcanzados en cada caso. En particular, la forma en que *Infomap* particiona los módulos de mayor tamaño en *CNM* da lugar a estructuras con mayor congruencia biológica, y el incremento observado no puede ser explicado meramente por las evidentes diferencias de tamaño.

Otra consecuencia directa del nivel de resolución proporcionado por cada algoritmo es el tipo de correlaciones observadas entre los roles cartográficos producidos por cada algoritmo y los grupos funcionales de proteínas estudiados. En el caso de proteínas asociadas a procesos de envejeciminto celular (ARG), se observa un enriquecimiento no trivial de éstas en la categoría *kinless Infomap*. Es importante destacar que ésta fue la única categoría cartográfica capaz de

revelar patrones de conectividad intramodular e intermodular en genes ARG que no puedan ser explicados meramente por la distribución de grado de esas proteínas. En particular, los niveles de *participación* de estos genes calculada al nivel de resolución provisto por *Infomap*, resultan en un descriptor topológico para este conjunto de genes con mejor desempeño que la *participación CNM* u otros observables topológicos considerados (*betweenness*, *bridging-centrality*), en especial cuando proteínas de bajo grado son consideradas. Estas consideraciones son compatibles con la hipótesis de que genes del conjunto ARG podrían coordinar y transmitir información entre los módulos topológicos detectados bajo la descripción *Infomap*. Además, esto último es compatible con la observación realizada por Xue y colaboradores [102] de que proteínas ARG tienen tendencia a situarse en las interfases modulares.

Consistentemente, las proteínas de tipo *corss-talk* que participan en diversas vías de señalización y se suponen coordinando e interconectando procesos entre distintos módulos funcionales, se encontraron significativamente enriquecidas en categorías *kinless Infomap*. Otro conjunto de proteínas enriquecido en esta categoría cartográfica es el de proteínas asociadas a *endocitosis*, las cuales fueron recientemente propuestas como agentes claves en la intercomunicación de vías metabólicas [115]. En ambos casos, (*cross-talk* y *endocitosis*), así como en proteínas asociadas a vías de *receptores*, fue *Infomap* la única descripción modular capaz de revelar asociaciones cartográficas significativas más allá de la distribución de grado de las proteínas bajo estudio. Por otro lado en el análisis de *ligandos*, *factores de transcripción* y *proteínas de andamiaje* ambas descripciones modulares fueron consistentes en el tipo de enriquecimientos cartográficos que revelaron.

En líneas generales los resultados expuestos hacen una llamada de atención respecto al uso de herramientas técnicas empleadas en el estudio de PINs para analizar y extraer conclusiones biológicas. En particular, cuando los análisis de PINs involucran el uso de una descripción modular, no resulta suficiente optar por particiones de óptima modularidad, sino que es imprescindible considerar también el nivel de resolución provisto por la técnica empleada.

Por otra parte, en el capítulo 5, se abordó el problema de búsqueda de blancos de droga y reposicionamiento de compuestos existentes en el contexto de enfermedades tropicales desatendidas (NTD). Se propuso una forma original de integrar una vasta cantidad y variedad de datos quimiogenómicos mediante el uso de redes multicapa. Este tipo de enfoque permitió abordar dos problemas biológicos de interés. En primer lugar se abordó la predicción de blancos proteicos en una especie patógena dada con escasa o nula cantidad de evidencias de bioactividades. En segundo lugar, se trabajó sobre la predicción de listas reducidas de blancos de acción para un compuesto químico de interés. En particular, se focalizó en el estudio de compuestos *huérfanos*,

es decir aquellos que se saben activos contra una especie patógena dada pero cuyos mecanismos y blancos de acción permanecen desconocidos. Ambos problemas fueron abordados con el mismo formalismo de redes multicapa, validando los resultados de priorizaciones realizadas tanto desde una perspectiva computacional (utilizando técnicas de validación cruzada) como por curación manual de literatura reciente u omitida en los datos considerados.

Respecto a la construcción de la red multicapa se mostró que un punto especialmente delicado es la manera de definir criterios de similitud entre proteínas. En el contexto de la red multicapa propuesta, eso se refleja en la forma particular de proyectar la red bipartita conformada por proteínas de distintas especies y los nodos de afiliación, es decir los criterios de similitud considerados (dominios Pfam, vías metabólicas y grupos de homología). Se propuso en este punto una forma de generalizar de la metodología propuesta por Zhou y colaboradores [46]. La versión propuesta en este manuscrito permite adicionalmente regular el grado de importancia relativa de los nodos en la capa a proyectar. En el contexto del problema de reposicionamiento de drogas, esto implica poder asignar mayor o menor nivel de similitud entre proteínas según la relevancia de los nodos de afiliación compartidos. Se mostró explícitamente mediante validación cruzada de 10 iteraciones que este tipo de proyección es robusta respecto a su parámetro libre α en un rango $\alpha \in [0.2,1]$, superando en todo este rango la capacidad predictiva de la red generada con la proyección original de Zhou.

Adicionalmente para la proyección bipartita propuesta se introdujo un factor de corrección que penaliza la promiscuidad de los nodos de afiliación compartidos entre proteínas. Se mostró que la consideración u omisión de esta corrección por grado de los nodos de afiliación, conduce a la construcción de redes multicapa de características y capacidades predictivas muy diferentes. Se mostró además que esas diferencias en la capacidad predictiva son independientes del algoritmo de priorización y los criterios de similitud (nodos de afiliación) utilizados. En particular, se mostró que el orden de proteínas en las priorizaciones realizadas sobre redes construidas sin emplear la penalización por grado (G_P^{sc}) pueden ser explicados esencialmente con el *strength* de las proteínas en la capa proyectada G_P^{sc}. Este hecho sugiere que las predicciones realizadas sobre la red G_P^{sc} pueden ser poco sensibles a las condiciones iniciales, ya que el orden resultante es esencialmente dado por un observable topológico de la red: el *strength*.

Desde el punto de vista biológico sin embago, las priorizaciones sobre este tipo de red (G_P^{sc}) resultan en interesantes casos de estudio para el problema de búsqueda de blancos de droga. Las proteínas priorizadas con esta red en tres especies patógenas fueron en su extensa mayoría proteínas con dominios kinasas, y varios de ellos fueron validados por literatura bajo curación manual.

Respecto a la metodologías de priorización utilizadas, se encontró que el algoritmo de priorización a primeros vecinos VS supera en desempeño a la estrategia más sofisticada de simulación de flujo funcional FF e incluso al desempeño obtenido utilizando una técnica de alineamiento de secuencias, FASTA. Más allá de las diferencias de desempeño en los valores de AUC-01 encontrado en uno u otro caso, se mostró que las priorizaciones basadas en la red multicapa ofrecen listas de priorización que difieren cualitativamente de los blancos propuestos por alineamiento de secuencias.

Por otro lado, para el problema de búsqueda de blancos en drogas huérfanas se encontró mediante validación computacional que la metodología propuesta permite inferir correctamente una alta cantidad de blancos para este tipo de compuestos. En el 60 % de un total de 1000 moléculas elegidas al azar se halló exitosamente el único blanco a buscado dentro de las primeras 10 proteínas propuestas. Más áun, en un 30 % de los casos, los blancos se hallaron dentro de las dos primeras posiciones. De hecho, analizando los mecanismos por los cuales se logró inferir correctamente estos blancos, se observa que en el 33 % de los casos los candidatos propuestos se detectaron haciendo uso extensivo de las conectividades dentro de la red multicapa. Por último también se realizaron priorizaciones de drogas huérfanas en Plasmodium Falciparum, encontrando numerosos casos interesantes de estudio (la N-miristoiltransfrerasa como blanco de una benzotiazolina, TDR ID:599594 contra péptidos bacterianos deformilasas, la proteína enoyl-acyl carrier reductase como blanco de 5 compuestos diferentes, etc). En suma, los resultados obtenidos muestran un gran potencial de la metodología propuesta para ser utilizada en la guía de nuevos ensayos experimentales relacionados con la búsqueda de nuevas dianas terapéuticas en NTDs.

Apéndice A

Trabajos científicos producto de este manuscrito

Mining the modular structure of protein interaction networks

Berenstein $A.$[1]., Piñero $J.$[1], Furlong L.I., Chernomoretz A.

[1]These authors contributed equally to this work.

Status: Minor Revision. PLOS ONE

A multilayer network approach for guiding drug repositioning in neglected diseases

Berenstein $A.$[1], Magariños P[1], Chernomoretz A., Aguero F.

[1]These authors contributed equally to this work.

Próximo a enviarse

Bibliografía

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition, 2002. ISBN 0815332181. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0815332181.

[2] Peter Csermely, Tamás Korcsmáros, Huba J M Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, June 2013. ISSN 1879-016X. doi: 10.1016/j.pharmthera.2013.01.016. URL http://www.ncbi.nlm.nih.gov/pubmed/23384594.

[3] D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. ISSN 0028-0836. doi: 10.1038/30918.

[4] A. L. Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272:173–187, 1999. ISSN 03784371. doi: 10.1016/S0378-4371(99)00291-5.

[5] Hannah Carter, Matan Hofree, and Trey Ideker. Genotype to phenotype via network analysis. *Current opinion in genetics & development*, 23:611–21, 2013. ISSN 1879-0380. doi: 10.1016/j.gde.2013.10.003. URL http://www.ncbi.nlm.nih.gov/pubmed/24238873.

[6] L H Hartwell, J J Hopfield, S Leibler, and A W Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999. ISSN 0028-0836. doi: 10.1038/35011540.

[7] Roger Guimerà and Luís A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005. ISSN 0028-0836. doi: 10.1038/nature03288.

[8] Roger Guimerà and Luís A Nunes Amaral. Cartography of complex networks: modules and universal roles, 2005. ISSN 1742-5468.

[9] Sumeet Agarwal, Charlotte M. Deane, Mason A. Porter, and Nick S. Jones. Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6:1–12, 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000817.

[10] Yuri Pritykin and Mona Singh. Simple Topological Features Reflect Dynamics and Modularity in Protein Interaction Networks. *PLoS Computational Biology*, 9, 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003243.

[11] Xiao Chang, Tao Xu, Yun Li, and Kai Wang. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Scientific reports*, 3:1691, 2013. ISSN 2045-2322. doi: 10.1038/srep01691. URL http://www.nature.com/srep/2013/130422/srep01691/full/srep01691.html.

[12] Aaron Clauset, M. Newman, and Cristopher Moore. Finding community structure in very large networks, 2004. ISSN 1539-3755.

[13] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1118–1123, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0706851105.

[14] Stephanie A. Robertson and Adam R. Renslo. Drug discovery for neglected tropical diseases at the Sandler Center. *Future Med Chem*, 3(10):1279–1288, Aug 2011. doi: 10.4155/fmc.11.85. URL http://dx.doi.org/10.4155/fmc.11.85.

[15] Juuso A. Parkkinen and Samuel Kaski. Probabilistic drug connectivity mapping. *BMC Bioinformatics*, 15(1):113, Apr 2014. doi: 10.1186/1471-2105-15-113. URL http://dx.doi.org/10.1186/1471-2105-15-113.

[16] Murat Iskar, Georg Zeller, Peter Blattmann, Monica Campillos, Michael Kuhn, Katarzyna H. Kaminska, Heiko Runz, Anne-Claude Gavin, Rainer Pepperkok, Vera van Noort, and Peer Bork. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biol*, 9:662, 2013. doi: 10.1038/msb.2013.20. URL http://dx.doi.org/10.1038/msb.2013.20.

[17] Dorothea Emig, Alexander Ivliev, Olga Pustovalova, Lee Lancashire, Svetlana Bureeva, Yuri Nikolsky, and Marina Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, 8(4):e60618, 2013. doi: 10.1371/journal.pone.0060618. URL http://dx.doi.org/10.1371/journal.pone.0060618.

[18] Persi Diaconis and Bela Bollobas. *Modern Graph Theory*, volume 95. 1998. ISBN 0387984887. doi: 10.2307/2669801. URL http://www.jstor.org/stable/2669801?origin=crossref.

[19] Stanley Wasserman and Katherine Faust. *Social network analysis*, volume 8. 1994. ISBN 9780521387071.

[20] S BOCCALETTI, V LATORA, Y MORENO, M CHAVEZ, and D HWANG. Complex networks: Structure and dynamics, 2006. ISSN 03701573.

[21] Bo Hu, Xin Yu Jiang, Jun Feng Ding, Yan Bo Xie, and Bing Hong Wang. A weighted network model for interpersonal relationship evolution. *Physica A: Statistical Mechanics and its Applications*, 353:576–594, 2005. ISSN 03784371. doi: 10.1016/j.physa.2005.01.052.

[22] V Latora and M Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87: 198701, 2001. ISSN 0031-9007. doi: 10.1103/PhysRevLett.87.198701.

[23] V. Latora and M. Marchiori. Economic small-world behavior in weighted networks. *European Physical Journal B*, 32:249–263, 2003. ISSN 14346028. doi: 10.1140/epjb/e2003-00095-5.

[24] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001. doi: 10.1103/PhysRevE.64.016132. URL http://link.aps.org/doi/10.1103/PhysRevE.64.016132.

[25] A Barrat, M Barthélemy, R Pastor-Satorras, and A Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:3747–3752, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0400087101.

[26] Santo Fortunato. Community detection in graphs, 2010. ISSN 03701573.

[27] M. Newman and M. Girvan. Finding and evaluating community structure in networks, 2004. ISSN 1539-3755.

[28] M MOLLOY and B REED. A CRITICAL-POINT FOR RANDOM GRAPHS WITH A GIVEN DEGREE SEQUENCE. *Random Structures and Algorithms*, 6:161–179, 1995. ISSN 1098-2418. doi: 10.1002/rsa.3240060204. URL papers2://publication/uuid/4BE5456F-A1A8-40A8-B519-A1045DAE0F34.

[29] MICHAEL MOLLOY and BRUCE REED. The Size of the Giant Component of a Random Graph with a Given Degree Sequence, 1998. ISSN 09635483.

[30] Erik Volz. Random networks with tunable degree distribution and clustering. *Phys. Rev. E*, 70: 056115, Nov 2004. doi: 10.1103/PhysRevE.70.056115. URL http://link.aps.org/doi/10.1103/PhysRevE.70.056115.

[31] P Erdös and A Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959. ISSN 00029947. doi: 10.2307/1999405.

[32] Z. Burda, J. Jurkiewicz, and A. Krzywicki. Statistical mechanics of random graphs. *Physica A*, 344:56–61, 2004. ISSN 03784371. doi: 10.1016/j.physa.2004.06.087. URL http://linkinghub.elsevier.com/retrieve/pii/S0378437104009069.

[33] B. Bollobás. *Random graphs*. Academic Press, 1985. ISBN 9780121117559. URL http://books.google.com.ar/books?id=2uvuAAAAMAAJ.

[34] P. Erdős and A Rényi. On the evolution of random graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960.

[35] Béla Bollobás. Degree sequences of random graphs. *Discrete Mathematics*, 33(1):1 – 19, 1981. ISSN 0012-365X. doi: http://dx.doi.org/10.1016/0012-365X(81)90253-3. URL http://www.sciencedirect.com/science/article/pii/0012365X81902533.

[36] Fan Chung and Linyuan Lu. The diameter of random sparse graphs. *Advances in Applied Math*, 26(4): 257–279, 2001.

[37] Roger Guimerà and Luís A Nunes Amaral. Cartography of complex networks: modules and universal roles, 2005. ISSN 1742-5468.

[38] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1–49, 2012.

[39] Jesse Gillis and Paul Pavlidis. The impact of multifunctional genes on "guilt by association.ªnalysis. *PLoS One*, 6(2):e17258, 2011. doi: 10.1371/journal.pone.0017258. URL http://dx.doi.org/10.1371/journal.pone.0017258.

[40] Jesse Gillis and Paul Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–1866, Jul 2011. doi: 10.1093/bioinformatics/btr288. URL http://dx.doi.org/10.1093/bioinformatics/btr288.

[41] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl

1:i302–i310, Jun 2005. doi: 10.1093/bioinformatics/bti1054. URL http://dx.doi.org/10.1093/bioinformatics/bti1054.

[42] Donna K. McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, August 1989. ISSN 1552-681X. doi: 10.1177/0272989x8900900307. URL http://dx.doi.org/10.1177/0272989x8900900307.

[43] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks, 2000.

[44] Renaud Lambiotte and Marcel Ausloos. Collaborative tagging as a tripartite network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3993 LNCS - III, pages 1114–1117, 2006. ISBN 3540343830. doi: 10.1007/11758532_152.

[45] Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D P Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network Properties of Folksonomies. *Ai Communications*, 20:245–262, 2007. ISSN 09217126. doi: citeulike-article-id:1473536. URL http://iospress.metapress.com/index/8X34022X43427K3H.pdf.

[46] Tao Zhou, Jie Ren, Matús Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76(4 Pt 2):046115, October 2007. ISSN 1539-3755. URL http://www.ncbi.nlm.nih.gov/pubmed/17995068.

[47] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks, 2014. ISSN 03701573.

[48] Martin H. Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2):e31826, 02 2012. doi: 10.1371/journal.pone.0031826. URL http://dx.doi.org/10.1371%2Fjournal.pone.0031826.

[49] Michael Albers, Harald Kranz, Ingo Kober, Carmen Kaiser, Martin Klink, Jörg Suckow, Rainer Kern, and Manfred Koegl. Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Molecular & cellular proteomics : MCP*, 4:205–213, 2005. ISSN 1535-9476. doi: 10.1074/mcp.M400169-MCP200.

[50] Russell Bell, Alan Hubbard, Rakesh Chettier, Di Chen, John P. Miller, Pankaj Kapahi, Mark Tarnopolsky, Sudhir Sahasrabuhde, Simon Melov, and Robert E. Hughes. A human protein interaction network shows

conservation of aging processes between human and invertebrate species. *PLoS Genetics*, 5, 2009. ISSN 15537390. doi: 10.1371/journal.pgen.1000414.

[51] Heike Goehler, Maciej Lalowski, Ulrich Stelzl, Stephanie Waelter, Martin Stroedicke, Uwe Worm, Anja Droege, Katrin S. Lindenberg, Maria Knoblich, Christian Haenig, Martin Herbst, Jaana Suopanki, Eberhard Scherzinger, Claudia Abraham, Bianca Bauer, Renate Hasenbank, Anja Fritzsche, Andreas H. Ludewig, Konrad Buessow, Sarah H. Coleman, Claire Anne Gutekunst, Bernhard G. Landwehrmeyer, Hans Lehrach, and Erich E. Wanker. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Molecular Cell*, 15:853–865, 2004. ISSN 10972765. doi: 10.1016/j.molcel.2004.09.016.

[52] Linda S. Kaltenbach, Eliana Romero, Robert R. Becklin, Rakesh Chettier, Russell Bell, Amit Phansalkar, Andrew Strand, Cameron Torcassi, Justin Savage, Anthony Hurlburt, Guang Ho Cha, Lubna Ukani, Cindy Lou Chepanoske, Yuejun Zhen, Sudhir Sahasrabudhe, James Olson, Cornelia Kurschner, Lisa M. Ellerby, John M. Peltier, Juan Botas, and Robert E. Hughes. Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genetics*, 3:689–708, 2007. ISSN 15537390. doi: 10.1371/journal.pgen.0030082.

[53] Ben Lehner and Christopher M. Sanderson. A protein interaction framework for human mRNA degradation. *Genome Research*, 14:1315–1323, 2004. ISSN 10889051. doi: 10.1101/gr.2122004.

[54] Janghoo Lim, Tong Hao, Chad Shaw, Akash J. Patel, Gábor Szabó, Jean François Rual, C. Joseph Fisk, Ning Li, Alex Smolyar, David E. Hill, Albert László Barabási, Marc Vidal, and Huda Y. Zoghbi. A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell*, 125:801–814, 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.03.032.

[55] Manabu Nakayama, Reiko Kikuno, and Osamu Ohara. Protein-protein interactions between large proteins: Two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Research*, 12:1773–1784, 2002. ISSN 10889051. doi: 10.1101/gr.406902.

[56] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick,

David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005. ISSN 0028-0836. doi: 10.1038/nature04209.

[57] Frédéric Colland, Xavier Jacq, Virginie Trouplin, Christelle Mougin, Caroline Groizeleau, Alexandre Hamburger, Alain Meil, Jérôme Wojcik, Pierre Legrain, and Jean Michel Gauthier. Functional proteomics mapping of a human signaling pathway. *Genome Research*, 14:1324–1332, 2004. ISSN 10889051. doi: 10.1101/gr.2334104.

[58] Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amélie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-László Barabási, and Marc Vidal. An empirical framework for binary interactome mapping. *Nature methods*, 6:83–90, 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1280.

[59] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 37:D767–D772, 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn892.

[60] Andrew Chatr-Aryamontri, Bobby Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, Teresa Reguly, Ashton Breitkreutz, Adnane Sellam, Daici Chen, Christie Chang, Jennifer Rust, Michael Livstone, Rose Oughtred, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2013 Update. *Nucleic Acids Research*, 41, 2013. ISSN 03051048. doi: 10.1093/nar/gks1158.

[61] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40, 2012. ISSN 03051048. doi: 10.1093/nar/gkr1088.

[62] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Research*, 40, 2012. ISSN 03051048. doi: 10.1093/nar/gkr930.

[63] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic acids research*, 32:D449–D451, 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh086.

[64] Gary D. Bader, Doron Betel, and Christopher W V Hogue. BIND: The Biomolecular Interaction Network Database, 2003. ISSN 03051048.

[65] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21:832–834, 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti115.

[66] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro, Andrew R. Jones, Pierre Alain Binz, Eric W. Deutsch, Matthew Chambers, Marius Kallhardt, Fredrik Levander, James Shofstahl, Sandra Orchard, Juan Antonio Vizcaíno, Henning Hermjakob, Christian Stephan, Helmut E. Meyer, and Martin Eisenacher. The HUPO proteomics standards initiativemass spectrometry controlled vocabulary. *Database*, 2013, 2013. ISSN 17580463. doi: 10.1093/database/bat009.

[67] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40:35, 1977. ISSN 00380431. doi: 10.2307/3033543. URL http://www.jstor.org/stable/3033543?origin=crossref.

[68] Haiyuan Yu, Philip M. Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3:713–720, 2007. ISSN 1553734X. doi: 10.1371/journal.pcbi.0030059.

[69] Martin H. Schaefer, Jean Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0031826.

[70] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *PNAS*, 99: 7821–7826, 2002.

[71] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5:101–113, 2004. ISSN 1471-0056. doi: 10.1038/nrg1272.

[72] Réka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118:4947–4957, 2005. ISSN 0021-9533. doi: 10.1242/jcs.02714.

[73] Elena Nabieva, Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21, 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti1054.

[74] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3:140, 2007. ISSN 1744-4292. doi: 10.1038/msb4100180.

[75] Andreas Zanzoni, Montserrat Soler-L??pez, and Patrick Aloy. A network medicine approach to human disease, 2009. ISSN 00145793.

[76] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68, 2011. ISSN 1471-0056. doi: 10.1038/nrg2918.

[77] M Vidal, M E Cusick, and A L Barabasi. Interactome networks and human disease. *Cell*, 144:986–998, 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.02.016. URL http://www.ncbi.nlm.nih.gov/pubmed/21414488.

[78] Antonio del Sol, Rudi Balling, Lee Hood, and David Galas. Diseases as network perturbations, 2010. ISSN 09581669.

[79] Laura I. Furlong. Human diseases through the lens of network biology, 2013. ISSN 01689525.

[80] H Jeong, S P Mason, A L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. ISSN 0028-0836. doi: 10.1038/35075138.

[81] Haiyuan Yu, Dov Greenbaum, Hao Xin Lu, Xiaowei Zhu, and Mark Gerstein. Genomic analysis of essentiality within protein networks, 2004. ISSN 01689525.

[82] Nizar N. Batada, Laurence D. Hurst, and Mike Tyers. Evolutionary and physiological importance of hub proteins. *PLoS Computational Biology*, 2:0748–0756, 2006. ISSN 1553734X. doi: 10.1371/journal.pcbi.0020088.

[83] Elena Zotenko, Julian Mestre, Dianne P. O'Leary, and Teresa M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 4, 2008. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000140.

[84] Jimin Song and Mona Singh. From Hub Proteins to Hub Modules: The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization. *PLoS Computational Biology*, 9, 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002910.

[85] Roger Guimerà, Marta Sales-Pardo, and Luís A Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76, 2007. ISSN 15393755. doi: 10.1103/PhysRevE.76.036102.

[86] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104:36–41, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0605965104.

[87] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis, 2009. ISSN 1539-3755.

[88] Renaud Lambiotte. Multi-scale modularity in complex networks. *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, 2010.

[89] Rodrigo Aldecoa and Ignacio Marín. Exploring the limits of community detection strategies in complex networks. *Scientific reports*, 3:2216, 2013. ISSN 2045-2322. doi: 10.1038/srep02216. URL http://www.nature.com/srep/2013/130717/srep02216/full/srep02216.html.

[90] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84, 2011. ISSN 15393755. doi: 10.1103/PhysRevE.84.066122.

[91] Ju Xiang and Ke Hu. Limitation of multi-resolution methods in community detection. *Physica A: Statistical Mechanics and its Applications*, 391:4995–5003, 2012. ISSN 03784371. doi: 10.1016/j.physa.2012.05.006.

[92] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification

of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL http://dx.doi.org/10.1038/75556.

[93] Susmita Datta and Somnath Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC bioinformatics*, 7:397, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-397.

[94] Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta. clValid : An R Package for Cluster Validation. *Journal Of Statistical Software*, 25:1–28, 2008. ISSN 15487660. doi: citeulike-article-id:2574494. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.3573&rep=rep1&type=pdf.

[95] Nicolas Bertin, Nicolas Simonis, Denis Dupuy, Michael E Cusick, Jing-Dong J Han, Hunter B Fraser, Frederick P Roth, and Marc Vidal. Confirmation of organized modularity in the yeast interactome. *PLoS Biol*, 5(6):e153, 06 2007. doi: 10.1371/journal.pbio.0050153. URL http://dx.doi.org/10.1371%2Fjournal.pbio.0050153.

[96] Adam Antebi. Genetics of aging in Caenorhabditis elegans, 2007. ISSN 15537390.

[97] Tarynn M. Witten and Danail Bonchev. Predicting aging/longevity-related genes in the nematode Caenorhabditis elegans. *Chemistry and Biodiversity*, 4:2639–2655, 2007. ISSN 16121872. doi: 10.1002/cbdv.200790216.

[98] Joris Deelen, Marian Beekman, Miriam Capri, Claudio Franceschi, and P. Eline Slagboom. Identifying the genomic determinants of aging and longevity in human population studies: Progress and challenges. *BioEssays*, 35:386–396, 2013. ISSN 02659247. doi: 10.1002/bies.201200148.

[99] Richard Weindruch, Tsuyoshi Kayo, Cheol Koo Lee, and Tomas A Prolla. Gene expression profiling of aging using DNA microarrays. *Mechanisms of ageing and development*, 123:177–193, 2002. ISSN 0047-6374. doi: 10.1016/S0047-6374(01)00344-X.

[100] Tao Lu, Ying Pan, Shyan-Yuan Kao, Cheng Li, Isaac Kohane, Jennifer Chan, and Bruce A Yankner. Gene regulation and DNA damage in the ageing human brain. *Nature*, 429:883–891, 2004. ISSN 0028-0836. doi: 10.1038/nature02661.

[101] Jerome D Boyd-Kirkup, Christopher D Green, Gang Wu, Dan Wang, and Jing-Dong J Han. Epigenomics and the regulation of aging. *Epigenomics*, 5:205–27, 2013. ISSN 1750-192X. doi: 10.2217/epi.13.5. URL http://www.ncbi.nlm.nih.gov/pubmed/23566097.

[102] Huiling Xue, Bo Xian, Dong Dong, Kai Xia, Shanshan Zhu, Zhongnan Zhang, Lei Hou, Qingpeng Zhang, Yi Zhang, and Jing-Dong J Han. A modular network model of aging. *Molecular systems biology*, 3:147, 2007. ISSN 1744-4292. doi: 10.1038/msb4100189.

[103] James West, Martin Widschwendter, and Andrew E Teschendorff. Distinctive topology of age-associated epigenetic drift in the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 110:14138–43, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1307242110. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3761591&tool=pmcentrez&rendertype=abstract.

[104] Jiguang Wang, Shihua Zhang, Yong Wang, Luonan Chen, and Xiang Sun Zhang. Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. *PLoS Computational Biology*, 5, 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000521.

[105] Jo??o Pedro de Magalh??es, Arie Budovsky, Gilad Lehmann, Joana Costa, Yang Li, Vadim Fraifeld, and George M. Church. The Human Ageing Genomic Resources: Online databases and tools for biogerontologists, 2009. ISSN 14749718.

[106] E R DeLong, D M DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845, 1988. ISSN 0006-341X. doi: 10.2307/2531595.

[107] D Fazekas, M Koltai, D Turei, D Modos, M Palfy, Z Dul, L Zsakai, M Szalay-Beko, K Lenti, I J Farkas, T Vellai, P Csermely, and T Korcsmaros. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol*, 7:7, 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-7. URL http://www.ncbi.nlm.nih.gov/pubmed/23331499.

[108] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(Database issue), January 2007. ISSN 1362-4962. URL http://view.ncbi.nlm.nih.gov/pubmed/17148475.

[109] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL http://dx.doi.org/10.2307/2346101.

[110] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. XThe hallmarks of aging, 2013. ISSN 00928674.

[111] William R. Jeck, Alex P. Siebold, and Norman E. Sharpless. Review: A meta-analysis of GWAS and age-associated diseases, 2012. ISSN 14749718.

[112] D Fazekas, M Koltai, D Turei, D Modos, M Palfy, Z Dul, L Zsakai, M Szalay-Beko, K Lenti, I J Farkas, T Vellai, P Csermely, and T Korcsmaros. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol*, 7:7, 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-7. URL http://www.ncbi.nlm.nih.gov/pubmed/23331499.

[113] Roby P Bhattacharyya, Attila Reményi, Brian J Yeh, and Wendell A Lim. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annual review of biochemistry*, 75:655–680, 2006. ISSN 0066-4154. doi: 10.1146/annurev.biochem.75.103004.142710.

[114] Tom D. Bunney, Diego Esposito, Corine Mas-Droux, Ekatarina Lamber, Rhona W. Baxendale, Marta Martins, Ambrose Cole, Dmitri Svergun, Paul C. Driscoll, and Matilda Katan. Structural and functional integration of the PLC?? interaction domains critical for regulatory mechanisms and signaling deregulation. *Structure*, 20:2062–2075, 2012. ISSN 09692126. doi: 10.1016/j.str.2012.09.005.

[115] M??t?? P??lfy, Attila Rem??nyi, and Tam??s Korcsm??ros. Endosomal crosstalk: Meeting points for signaling pathways, 2012. ISSN 09628924.

[116] Peter J. Hotez, David H. Molyneux, Alan Fenwick, Jacob Kumaresan, Sonia Ehrlich Sachs, Jeffrey D. Sachs, and Lorenzo Savioli. Control of neglected tropical diseases. *N Engl J Med*, 357(10):1018–1027, Sep 2007. doi: 10.1056/NEJMra064142. URL http://dx.doi.org/10.1056/NEJMra064142.

[117] Paul G. Wyatt, Ian H. Gilbert, Kevin D. Read, and Alan H. Fairlamb. Target validation: linking target and chemical properties to desired product profile. *Curr Top Med Chem*, 11(10):1275–1283, 2011.

[118] Joseph A. DiMasi, Ronald W. Hansen, and Henry G. Grabowski. The price of innovation: new estimates of drug development costs. *J Health Econ*, 22(2):151–185, Mar 2003. doi: 10.1016/S0167-6296(02)00126-1. URL http://dx.doi.org/10.1016/S0167-6296(02)00126-1.

[119] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*, 3(8):711–715, Aug 2004. doi: 10.1038/nrd1470. URL http://dx.doi.org/10.1038/nrd1470.

[120] Patrice Trouiller, Piero Olliaro, Els Torreele, James Orbinski, Richard Laing, and Nathan Ford. Drug development for neglected diseases: a deficient market and a public-health policy failure. *Lancet*, 359 (9324):2188–2194, Jun 2002. doi: 10.1016/S0140-6736(02)09096-7. URL http://dx.doi.org/10.1016/S0140-6736(02)09096-7.

[121] Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*, 3(8):673–683, Aug 2004. doi: 10.1038/nrd1468. URL http://dx.doi.org/10.1038/nrd1468.

[122] Curtis R. Chong and David J Sullivan, Jr. New uses for old drugs. *Nature*, 448(7154):645–646, Aug 2007. doi: 10.1038/448645a. URL http://dx.doi.org/10.1038/448645a.

[123] Natalia Novac. Challenges and opportunities of drug repositioning. *Trends Pharmacol Sci*, 34(5):267–272, May 2013. doi: 10.1016/j.tips.2013.03.004. URL http://dx.doi.org/10.1016/j.tips.2013.03.004.

[124] Steve K. Teo, Ken E. Resztak, Michael A. Scheffler, Karin A. Kook, Jerry B. Zeldis, David I. Stirling, and Steve D. Thomas. Thalidomide in the treatment of leprosy. *Microbes Infect*, 4(11):1193–1202, Sep 2002.

[125] V. J. Haupt and Michael Schroeder. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief Bioinform*, 12(4):312–326, Jul 2011. doi: 10.1093/bib/bbr011. URL http://dx.doi.org/10.1093/bib/bbr011.

[126] Michael P. Pollastri and Robert K. Campbell. Target repurposing for neglected diseases. *Future Med Chem*, 3(10):1307–1315, Aug 2011. doi: 10.4155/fmc.11.92. URL http://dx.doi.org/10.4155/fmc.11.92.

[127] Guangxu Jin and Stephen T C. Wong. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today*, 19(5):637–644, May 2014. doi: 10.1016/j.drudis.2013.11.005. URL http://dx.doi.org/10.1016/j.drudis.2013.11.005.

[128] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, Jul 2008. doi: 10.1126/science.1158140. URL http://dx.doi.org/10.1126/science.1158140.

[129] Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, Nov 2009. doi: 10.1038/nature08506. URL http://dx.doi.org/10.1038/nature08506.

[130] Francesco Iorio, Timothy Rittman, Hong Ge, Michael Menden, and Julio Saez Rodriguez. Transcriptional data: a new gateway to drug repositioning? *Drug Discov Today*, 18(7-8):350–357, Apr 2013. doi: 10.1016/j.drudis.2012.07.014. URL http://dx.doi.org/10.1016/j.drudis.2012.07.014.

[131] Jamel Meslamani, Ricky Bhajun, Francois Martz, and Didier Rognan. Computational profiling of bioactive compounds using a target-dependent composite workflow. *J Chem Inf Model*, 53(9):2322–2333, Sep 2013. doi: 10.1021/ci400303n. URL http://dx.doi.org/10.1021/ci400303n.

[132] Felix A. Kruger and John P. Overington. Global analysis of small molecule binding to related protein targets. *PLoS Comput Biol*, 8(1):e1002333, Jan 2012. doi: 10.1371/journal.pcbi.1002333. URL http://dx.doi.org/10.1371/journal.pcbi.1002333.

[133] Fernán Agüero, Bissan Al Lazikani, Martin Aslett, Matthew Berriman, Frederick S. Buckner, Robert K. Campbell, Santiago Carmona, Ian M. Carruthers, A. W EdithA. W Edith Chan, Feng Chen, Gregory J. Crowther, Maria A. Doyle, Christiane Hertz Fowler, Andrew L. Hopkins, Gregg McAllister, Solomon Nwaka, John P. Overington, Arnab Pain, Gaia V. Paolini, Ursula Pieper, Stuart A. Ralph, Aaron Riechers, David S. Roos, Andrej Sali, Dhanasekaran Shanmugam, Takashi Suzuki, Wesley C. Van Voorhis, and Christophe L. M. J. Verlinde. Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov*, 7(11):900–907, Nov 2008. doi: 10.1038/nrd2684. URL http://dx.doi.org/10.1038/nrd2684.

[134] Gregory J. Crowther, Dhanasekaran Shanmugam, Santiago J. Carmona, Maria A. Doyle, C. Hertz Fowler, Matthew Berriman, Solomon Nwaka, Stuart A. Ralph, David S. Roos, Wesley C. Van Voorhis, and Fernán Agüero. Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. *PLoS Negl Trop Dis*, 4(8):e804, 2010. doi: 10.1371/journal.pntd.0000804. URL http://dx.doi.org/10.1371/journal.pntd.0000804.

[135] María P. Magariños, Santiago J. Carmona, Gregory J. Crowther, Stuart A. Ralph, David S. Roos, Dhanasekaran Shanmugam, Wesley C. Van Voorhis, and Fernán Agüero. TDR Targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res*, 40(Database issue):D1118–D1127, Jan 2012. doi: 10.1093/nar/gkr1053. URL http://dx.doi.org/10.1093/nar/gkr1053.

[136] Francisco-Javier Gamo, Laura M. Sanz, Jaume Vidal, Cristina de Cozar, Emilio Alvarez, Jose-Luis Lavandera, Dana E. Vanderwall, Darren V. S. Green, Vinod Kumar, Samiul Hasan, James R. Brown, Catherine E. Peishoff, Lon R. Cardon, and Jose F Garcia Bustos. Thousands of chemical starting points for antimalarial lead identification. *Nature*, 465(7296):305–310, May 2010. doi: 10.1038/nature09107. URL http://dx.doi.org/10.1038/nature09107.

[137] W. Armand Guiguemde, Anang A. Shelat, David Bouck, Sandra Duffy, Gregory J. Crowther, Paul H. Davis, David C. Smithson, Michele Connelly, Julie Clark, Fangyi Zhu, María B Jiménez Díaz, María S. Martinez, Emily B. Wilson, Abhai K. Tripathi, Jiri Gut, Elizabeth R. Sharlow, Ian Bathurst, Farah El Mazouni, Joseph W. Fowble, Isaac Forquer, Paula L McGinley, Steve Castro, Iñigo Angulo Barturen, Santiago Ferrer, Philip J. Rosenthal, Joseph L. Derisi, David J. Sullivan, John S. Lazo, David S. Roos, Michael K. Riscoe, Margaret A. Phillips, Pradipsinh K. Rathod, Wesley C. Van Voorhis, Vicky M. Avery, and R. Kiplin Guy. Chemical genetics of Plasmodium falciparum. *Nature*, 465(7296):311–315, May 2010. doi: 10.1038/nature09099. URL http://dx.doi.org/10.1038/nature09099.

[138] Thomas Spangenberg, Jeremy N. Burrows, Paul Kowalczyk, Simon McDonald, Timothy N. C. Wells, and Paul Willis. The open access malaria box: a drug discovery catalyst for neglected diseases. *PLoS One*, 8(6):e62906, 2013. doi: 10.1371/journal.pone.0062906. URL http://dx.doi.org/10.1371/journal.pone.0062906.

[139] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology*, 8(5):e1002503, January 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002503. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3349722&tool=pmcentrez&rendertype=abstract.

[140] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics (Oxford, England)*, 29(16):2004–8, August 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt307. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3722516&tool=pmcentrez&rendertype=abstract.

[141] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Globalized bipartite local model for drug-target interaction prediction. *Proceedings of the 11th International Workshop on Data Mining in Bioinformatics - BIOKDD '12*, pages 8–14, 2012. doi: 10.1145/2350176.2350178. URL http://dl.acm.org/citation.cfm?doid=2350176.2350178.

[142] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug-target interaction prediction by random walk on the heterogeneous network, 2012. ISSN 1742-206X.

[143] Peter Csermely, Tamás Korcsmáros, Huba J M Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138:333–408, 2013. ISSN 1879-016X. doi: 10.1016/j.pharmthera.2013.01.016. URL http://www.ncbi.nlm.nih.gov/pubmed/23384594.

[144] W M Fitch. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19:99–113, 1970. ISSN 00397989. doi: 10.2307/2412448.

[145] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39:309–338, 2005. ISSN 0066-4197. doi: 10.1146/annurev.genet.39.073003.114725.

[146] Li Li, Christian J Stoeckert, and David S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003. doi: 10.1101/gr.1224503. URL http://dx.doi.org/10.1101/gr.1224503.

[147] Steve Fischer, Brian P. Brunk, Feng Chen, Xin Gao, Omar S. Harb, John B. Iodice, Dhanasekaran Shanmugam, David S. Roos, and Christian J Stoeckert, Jr. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics*, Chapter 6:Unit 6.12.1–Unit 6.1219, Sep 2011. doi: 10.1002/0471250953.bi0612s35. URL http://dx.doi.org/10.1002/0471250953.bi0612s35.

[148] A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30:1575–1584, 2002. ISSN 1362-4962. doi: 10.1093/nar/30.7.1575.

[149] Kratz Rene Fester. *Molecular & Cell Biology for Dummies*. For Dummies Series. Wiley Publishing Inc, 2009.

[150] J MONOD, J WYMAN, and J P CHANGEUX. ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *Journal of molecular biology*, 12:88–118, 1965. ISSN 00222836. doi: 10.1016/S0022-2836(65)80285-6.

[151] Timothy J. Peters Peter Nelson Campbell. *Bioquímica ilustrada: bioquímica y biología molecular en la era posgenómica*. Masson 2007, 5 edition, 2006.

[152] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: The protein families database, 2014. ISSN 03051048.

[153] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014. doi: 10.1093/bioinformatics/btu031. URL http://dx.doi.org/10.1093/bioinformatics/btu031.

[154] Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7, 2011. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002195.

[155] M Kanehisa and S Goto. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28:27–30, 2000. ISSN 03051048. URL http://www.genome.jp/kegg/.

[156] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, 42, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1076.

[157] Raul O Cosentino, Patricio Diosque, and Fernán Agüero. Genetic profiling of the isoprenoid and sterol biosynthesis pathways of trypanosoma cruzi. *PeerJ PrePrints*, 1:e44v1, 7 2013. ISSN 2167-9843. doi: 10.7287/peerj.preprints.44v1. URL http://dx.doi.org/10.7287/peerj.preprints.44v1.

[158] Anna Gaulton, Louisa J. Bellis, A. P. Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, S. McGlinchey, David Michalovich, Bissan Al Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 40(Database issue):D1100–D1107, Jan 2012. doi: 10.1093/nar/gkr777. URL http://dx.doi.org/10.1093/nar/gkr777.

[159] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Sociète Vaudense des Sciences Naturelles*, 44:223–270, 1908.

[160] Evan E. Bolton, Yanli Wang, Paul A. Thiessen, and Stephen H. Bryant. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities, 2008. ISSN 15741400.

[161] Maureen E. Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E. Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P. St Onge, Mike Tyers, Daphne Koller, Russ B. Altman, Ronald W. Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874):362–365, Apr 2008. doi: 10.1126/science.1150021. URL http://dx.doi.org/10.1126/science.1150021.

[162] David E. Patterson, Richard D. Cramer, Allan M. Ferguson, Robert D. Clark, and Laurence E. Weinberger. Neighborhood behavior: A useful concept for validation of 'molecular diversity' descriptors. *Journal of Medicinal Chemistry*, 39:3049–3059, 1996. ISSN 00222623. doi: 10.1021/jm960290n.

[163] Yvonne C. Martin, James L. Kofron, and Linda M. Traphagen. Do structurally similar molecules have similar biological activity? *J Med Chem*, 45(19):4350–4358, Sep 2002.

[164] Katherine Faust. Centrality in affiliation networks. *Social Networks*, 19(2):157 – 191, 1997. ISSN 0378-8733. doi: http://dx.doi.org/10.1016/S0378-8733(96)00300-0. URL http://www.sciencedirect.com/science/article/pii/S0378873396003000.

[165] William R. Pearson. BLAST and FASTA similarity searching for multiple sequence alignment. *Methods Mol Biol*, 1079:75–101, 2014. doi: 10.1007/978-1-62703-646-7_5. URL http://dx.doi.org/10.1007/978-1-62703-646-7_5.

[166] W. James Kent. BLAT - The BLAST-like alignment tool. *Genome Research*, 12:656–664, 2002. ISSN 10889051. doi: 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002.

[167] Randen L Patterson, Darren Boehning, and Solomon H Snyder. Inositol 1,4,5-trisphosphate receptors as signal integrators. *Annual review of biochemistry*, 73:437–465, 2004. ISSN 0066-4154. doi: 10.1146/annurev.biochem.73.071403.161303.

[168] Guozhong Huang, Paula J Bartlett, Andrew P Thomas, Silvia N J Moreno, and Roberto Docampo. Acidocalcisomes of Trypanosoma brucei have an inositol 1,4,5-trisphosphate receptor that is required for growth and infectivity. *Proceedings of the National Academy of Sciences of the United States of America*, 110:1887–92, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1216955110. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3562765&tool=pmcentrez&rendertype=abstract.

[169] Muneaki Hashimoto, Masahiro Enomoto, Jorge Morales, Nagomi Kurebayashi, Takashi Sakurai, Tetsuo Hashimoto, Takeshi Nara, and Katsuhiko Mikoshiba. Inositol 1,4,5-trisphosphate receptor regulates replication, differentiation, infectivity and virulence of the parasitic protist Trypanosoma cruzi. *Mol Microbiol*, 87(6):1133–1150, Mar 2013. doi: 10.1111/mmi.12155. URL http://dx.doi.org/10.1111/mmi.12155.

[170] Diana Bahia, Luciana Márcia Oliveira, Fabio Mitsuo Lima, Priscila Oliveira, José Franco da Silveira, Renato Arruda Mortara, and Jerônimo Conceição Ruiz. The TryPIKinome of five human pathogenic trypanosomatids: Trypanosoma brucei, Trypanosoma cruzi, Leishmania major, Leishmania braziliensis and Leishmania infantum–new tools for designing specific inhibitors. *Biochem Biophys Res Commun*,

390(3):963–970, Dec 2009. doi: 10.1016/j.bbrc.2009.10.086. URL http://dx.doi.org/10.1016/j.bbrc.2009.10.086.

[171] Daniel P. Sutherlin, Linda Bao, Megan Berry, Georgette Castanedo, Irina Chuckowree, Jenna Dotson, Adrian Folks, Lori Friedman, Richard Goldsmith, Janet Gunzner, Timothy Heffron, John Lesnick, Cristina Lewis, Simon Mathieu, Jeremy Murray, Jim Nonomiya, Jodie Pang, Niel Pegg, Wei Wei Prior, Lionel Rouge, Laurent Salphati, Deepak Sampath, Qingping Tian, Vickie Tsui, Nan Chi Wan, Shumei Wang, Binqing Wei, Christian Wiesmann, Ping Wu, Bing-Yan Zhu, and Alan Olivero. Discovery of a potent, selective, and orally available class I phosphatidylinositol 3-kinase (PI3K)/mammalian target of rapamycin (mTOR) kinase inhibitor (GDC-0980) for the treatment of cancer. *J Med Chem*, 54(21):7579–7587, Nov 2011. doi: 10.1021/jm2009327. URL http://dx.doi.org/10.1021/jm2009327.

[172] Aaron M Woolsey, Lisa Sunwoo, Christine A Petersen, Saskia M Brachmann, Lewis C Cantley, and Barbara A Burleigh. Novel PI 3-kinase-dependent mechanisms of trypanosome invasion and vacuole maturation. *Journal of cell science*, 116:3611–3622, 2003. ISSN 0021-9533. doi: 10.1242/jcs.00666.

[173] Luciana O Andrade and Norma W Andrews. Lysosomal fusion is essential for the retention of Trypanosoma cruzi inside host cells. *The Journal of experimental medicine*, 200:1135–1143, 2004. ISSN 0022-1007. doi: 10.1084/jem.20041408.

[174] Alejandra C. Schoijet, Kildare Miranda, Wendell Girard Dias, Wanderley de Souza, Mirtha M. Flawiá, Héctor N. Torres, Roberto Docampo, and Guillermo D. Alonso. A Trypanosoma cruzi phosphatidylinositol 3-kinase (TcVps34) is involved in osmoregulation and receptor-mediated endocytosis. *J Biol Chem*, 283(46):31541–31550, Nov 2008. doi: 10.1074/jbc.M801367200. URL http://dx.doi.org/10.1074/jbc.M801367200.

[175] Raúl O. Cosentino and Fernán Agüero. Genetic Profiling of the Isoprenoid and Sterol Biosynthesis Pathway Genes of Trypanosoma cruzi. *PLoS One*, 9(5):e96762, 2014. doi: 10.1371/journal.pone.0096762. URL http://dx.doi.org/10.1371/journal.pone.0096762.

[176] Galina I. Lepesheva, Natalia G. Zaitseva, W. David Nes, Wenxu Zhou, Miharu Arase, Jialin Liu, George C. Hill, and Michael R. Waterman. CYP51 from Trypanosoma cruzi: A phyla-specific residue in the B??? helix defines substrate preferences of sterol 14??-demethylase. *Journal of Biological Chemistry*, 281: 3577–3585, 2006. ISSN 00219258. doi: 10.1074/jbc.M510317200.

[177] Galina I. Lepesheva, Hee-Won Park, Tatiana Y. Hargrove, Benoit Vanhollebeke, Zdzislaw Wawrzak, Joel M. Harp, Munirathinam Sundaramoorthy, W David Nes, Etienne Pays, Minu Chaudhuri, Fernando Villalta, and Michael R. Waterman. Crystal structures of Trypanosoma brucei sterol 14alpha-demethylase

and implications for selective treatment of human infections. *J Biol Chem*, 285(3):1773–1780, Jan 2010. doi: 10.1074/jbc.M109.067470. URL http://dx.doi.org/10.1074/jbc.M109.067470.

[178] Valter Viana Andrade-Neto, Herbert Leonel de Matos-Guedes, Daniel Cláudio de Oliveira Gomes, Marilene Marcuzzo do Canto-Cavalheiro, Bartira Rossi-Bergmann, and Eduardo Caio Torres-Santos. The stepwise selection for ketoconazole resistance induces upregulation of C14-demethylase (CYP51) in Leishmania Amazonensis. *Memorias do Instituto Oswaldo Cruz*, 107:416–419, 2012. ISSN 00740276. doi: 10.1590/S0074-02762012000300018.

[179] Yi Bao, Louis M. Weiss, Vicki L. Braunstein, and Huan Huang. Role of protein kinase A in trypanosoma cruzi. *Infection and Immunity*, 76:4757–4763, 2008. ISSN 00199567. doi: 10.1128/IAI.00527-08.

[180] Yi Bao, Louis M. Weiss, Yan Fen Ma, Stuart Kahn, and Huan Huang. Protein kinase A catalytic subunit interacts and phosphorylates members of trans-sialidase super-family in Trypanosoma cruzi. *Microbes and Infection*, 12:716–726, 2010. ISSN 12864579. doi: 10.1016/j.micinf.2010.04.014.

[181] John J. Allocco, Robert Donald, Tanya Zhong, Anita Lee, Yui Sing Tang, Ronald C. Hendrickson, Paul Liberator, and Bakela Nare. Inhibitors of casein kinase 1 block the growth of Leishmania major promastigotes in vitro. *International Journal for Parasitology*, 36:1249–1259, 2006. ISSN 00207519. doi: 10.1016/j.ijpara.2006.06.013.

[182] Sophie Marhadour, Pascal Marchand, Fabrice Pagniez, Marc Antoine Bazin, Carine Picot, Olivier Lozach, Sandrine Ruchaud, Maud Antoine, Laurent Meijer, Najma Rachidi, and Patrice Le Pape. Synthesis and biological evaluation of 2,3-diarylimidazo[1,2-a]pyridines as antileishmanial agents. *European Journal of Medicinal Chemistry*, 58:543–556, 2012. ISSN 02235234. doi: 10.1016/j.ejmech.2012.10.048.

[183] Carmenza Spadafora, Yolanda Repetto, Cristina Torres, Laura Pino, Carlos Robello, Antonio Morello, Francisco Gamarro, and Santiago Castanys. Two casein kinase 1 isoforms are differentially expressed in Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, 124:23–36, 2002. ISSN 01666851. doi: 10.1016/S0166-6851(02)00156-1.

[184] M. Knockaert, N. Gray, E. Damiens, Y. T. Chang, P. Grellier, K. Grant, D. Fergusson, J. Mottram, M. Soete, J. F. Dubremetz, K. Le Roch, C. Doerig, P. G. Schultz, and L. Meijer. Intracellular targets of cyclin-dependent kinase inhibitors: Identification by affinity chromatography using immobilised inhibitors. *Chemistry and Biology*, 7:411–422, 2000. ISSN 10745521. doi: 10.1016/S1074-5521(00)00124-1.

[185] Edward W. Tate, Andrew S. Bell, Mark D. Rackham, and Megan H. Wright. N-Myristoyltransferase as a potential drug target in malaria and leishmaniasis. *Parasitology*, 141(1):37–49, Jan 2014. doi: 10.1017/S0031182013000450. URL http://dx.doi.org/10.1017/S0031182013000450.

[186] Chunquan Sheng, Jie Zhu, Wannian Zhang, Min Zhang, Haitao Ji, Yunlong Song, Hui Xu, Jianzhong Yao, Zhenyuan Miao, Youjun Zhou, Jü Zhu, and Jiaguo Lü. 3D-QSAR and molecular docking studies on benzothiazole derivatives as Candida albicans N-myristoyltransferase inhibitors. *Eur J Med Chem*, 42 (4):477–486, Apr 2007. doi: 10.1016/j.ejmech.2006.11.001. URL http://dx.doi.org/10.1016/j.ejmech.2006.11.001.

[187] Mark D. Rackham, James A. Brannigan, Kaveri Rangachari, Stephan Meister, Anthony J. Wilkinson, Anthony A. Holder, Robin J. Leatherbarrow, and Edward W. Tate. Design and synthesis of high affinity inhibitors of Plasmodium falciparum and Plasmodium vivax N-myristoyltransferases directed by ligand efficiency dependent lipophilicity (LELP). *J Med Chem*, 57(6):2773–2788, Mar 2014. doi: 10.1021/jm500066b. URL http://dx.doi.org/10.1021/jm500066b.

[188] Megan H. Wright, Barbara Clough, Mark D. Rackham, Kaveri Rangachari, James A. Brannigan, Munira Grainger, David K. Moss, Andrew R. Bottrill, William P. Heal, Malgorzata Broncel, Remigiusz A. Serwa, Declan Brady, David J. Mann, Robin J. Leatherbarrow, Rita Tewari, Anthony J. Wilkinson, Anthony A. Holder, and Edward W. Tate. Validation of N-myristoyltransferase as an antimalarial drug target using an integrated chemical biology approach. *Nat Chem*, 6(2):112–121, Feb 2014. doi: 10.1038/nchem.1830. URL http://dx.doi.org/10.1038/nchem.1830.

[189] Paul W. Bowyer, Ruwani S. Gunaratne, Munira Grainger, Chrislaine Withers Martinez, Sasala R. Wickramsinghe, Edward W. Tate, Robin J. Leatherbarrow, Katherine A. Brown, Anthony A. Holder, and Deborah F. Smith. Molecules incorporating a benzothiazole core scaffold inhibit the N-myristoyltransferase of Plasmodium falciparum. *Biochem J*, 408(2):173–180, Dec 2007. doi: 10.1042/BJ20070692. URL http://dx.doi.org/10.1042/BJ20070692.

[190] Kazuo Yamazaki, Yasushi Kaneko, Kie Suwa, Shinji Ebara, Kyoko Nakazawa, and Kazuhiro Yasuno. Synthesis of potent and selective inhibitors of Candida albicans N-myristoyltransferase based on the benzothiazole structure. *Bioorg Med Chem*, 13(7):2509–2522, Apr 2005. doi: 10.1016/j.bmc.2005.01.033. URL http://dx.doi.org/10.1016/j.bmc.2005.01.033.

[191] Sarah K Volkman, Pardis C Sabeti, David DeCaprio, Daniel E Neafsey, Stephen F Schaffner, Danny A Milner, Johanna P Daily, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Manoj T Duraisingh, Amanda Lukens, Alan Derr, Nicole Stange-Thomann, Skye Waggoner, Robert Onofrio, Liuda Ziaugra, Evan Mauceli, Sante Gnerre, David B Jaffe, Joanne Zainoun, Roger C Wiegand, Bruce W Birren, Daniel L Hartl, James E Galagan, Eric S Lander, and Dyann F Wirth. A genome-wide map of diversity in Plasmodium falciparum. *Nature genetics*, 39:113–119, 2007. ISSN 1061-4036. doi: 10.1038/ng1930.

[192] J. Wiesner, S. Sanderbrand, B. Altincicek, E. Beck, and H. Jomaa. Seeking new targets for antiparasitic agents. *Trends Parasitol*, 17(1):7–8, Jan 2001.

[193] Federica Belluti, Remo Perozzo, Leonardo Lauciello, Francesco Colizzi, Dirk Kostrewa, Alessandra Bisi, Silvia Gobbi, Angela Rampa, Maria Laura Bolognesi, Maurizio Recanatini, Reto Brun, Leonardo Scapozza, and Andrea Cavalli. Design, synthesis, and biological and crystallographic evaluation of novel inhibitors of Plasmodium falciparum enoyl-ACP-reductase (PfFabI). *Journal of medicinal chemistry*, 56: 7516–26, 2013. ISSN 1520-4804. doi: 10.1021/jm400637m. URL http://www.ncbi.nlm.nih.gov/pubmed/24063369.

[194] Ramanuj P. Samal, Vijay M. Khedkar, Raghuvir R. S. Pissurlenkar, Angela Gono Bwalya, Deniz Tasdemir, Ramesh A. Joshi, P. R. Rajamohanan, Vedavati G. Puranik, and Evans C. Coutinho. Design, synthesis, structural characterization by ir, 1h, 13c, 15n, 2d-nmr, x-ray diffraction and evaluation of a new class of phenylaminoacetic acid benzylidene hydrazines as pfenr inhibitors. *Chemical Biology Drug Design*, 81(6):715–729, 2013. ISSN 1747-0285. doi: 10.1111/cbdd.12118. URL http://dx.doi.org/10.1111/cbdd.12118.

[195] Florian C Schrader, Serghei Glinca, Julia M Sattler, Hans-Martin Dahse, Gustavo A Afanador, Sean T Prigge, Michael Lanzer, Ann-Kristin Mueller, Gerhard Klebe, and Martin Schlitzer. Novel type II fatty acid biosynthesis (FAS II) inhibitors as multistage antimalarial agents. *ChemMedChem*, 8:442–61, 2013. ISSN 1860-7187. doi: 10.1002/cmdc.201200407. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3633473&tool=pmcentrez&rendertype=abstract.

[196] Akhtar Muhammad, Itrat Anis, Zulfiqar Ali, Sufyan Awadelkarim, Ajmal Khan, Asaad Khalid, Muhammad Raza Shah, M. Galal, Ikhlas A. Khan, and M. Iqbal Choudhary. Methylenebissantin: A rare methylene-bridged bisflavonoid from Dodonaea viscosa which inhibits Plasmodium falciparum enoyl-ACP reductase. *Bioorganic and Medicinal Chemistry Letters*, 22:610–612, 2012. ISSN 0960894X. doi: 10.1016/j.bmcl.2011.10.072.

[197] Stephen P. Muench, Jozef Stec, Ying Zhou, Gustavo A. Afanador, Martin J. McPhillie, Mark R. Hickman, Patty J. Lee, Susan E. Leed, Jennifer M. Auschwitz, Sean T. Prigge, David W. Rice, and Rima McLeod. Development of a triclosan scaffold which allows for adaptations on both the A- and B-ring for transport peptides. *Bioorganic and Medicinal Chemistry Letters*, 23:3551–3555, 2013. ISSN 0960894X. doi: 10.1016/j.bmcl.2013.04.035.

[198] James S. Pham, Karen L. Dawson, Katherine E. Jackson, Erin E. Lim, Charisse Flerida A Pasaje, Kelsey E C Turner, and Stuart A. Ralph. Aminoacyl-tRNA synthetases as drug targets in eukaryotic parasites, 2014. ISSN 22113207.

[199] Eva S Istvan, Neekesh V Dharia, Selina E Bopp, Ilya Gluzman, Elizabeth A Winzeler, and Daniel E Goldberg. Validation of isoleucine utilization targets in Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, 108:1627–1632, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1011560108.

[200] Tao Zhou, Jie Ren, Matús Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76(4 Pt 2):046115, October 2007. ISSN 1539-3755. URL http://www.ncbi.nlm.nih.gov/pubmed/17995068.