Tesis de Grado

Redes complejas y el problema de reposicionamiento de fármacos

Bivort Haiek, Felipe

2017

Redes complejas y el problema de reposicionamiento de fármacos

Felipe Bivort Haiek

Tesis de Licenciatura en Ciencias Físicas

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Octubre 2017

TEMA:Física de redes/Biofísica

ALUMNO:Felipe Bivort Haiek

LU N° :609/09

LUGAR DE TRABAJO:FCEyN

DIRECTOR DEL TRABAJO: Ariel Chernomoretz

INFORME FINAL APROBADO POR

_____	_____
Autor	**Jurado**
_____	_____
Director	**Jurado**
_____	_____
Profesor de Tesis de Licenciatura	**Jurado**

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 estíimulos 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 controlando y coordinando para ello 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, pese a la cantidad de variables involucradas, hacen de ella un sistema extremadamente complejo para su estudio que ha despertado interés en distintas disciplinas científicas, entre ellas la física.

Si bien en los siglos anteriores las biologia estuvo en su mayor parte dominada por el reduccionismo, que logró identificar estructuras celulares simples y sus funciones, las células presentan múltiples fenotipos y propiedades emergentes. Estas son sistemas complejos, que dependen de intrincados mecanismos de interacción y acoplamiento entre material genético y proteínas, y presenta diversos métodos de sensado, señalización y control. En particular el mero problema de analizar el acoplamiento de una proteína y una molécula pequeña en un sitio de enlace dado, dentro del entorno celular, teniendo en cuenta el universo posible de interacciones alternativas que compiten con esta resulta demasiado costoso computacionalmente.

Uno de los métodos usados para describir los sistemas biológicos, sin perder la visión completa , pero reteniendo poder de caracterización es el de redes complejas [2]. Este enfoque, ha tenido un fuerte auge en el campo de la fíisica 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.

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, permiten relevar genomas o proteomas completos con un mínimo esfuerzo, y técnicas como Y2H(yeast-2hybrid) pueden determinar si dos proteínas interactúan dentro del entorno celular de la levadura . 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.

En los últimos años, se ha incrementado la atención, en las redes fármaco-proteína. Estas redes en general presentan dos tipos de nodos: proteínas, y drogas, donde la actividad ,es decir la interacción experimentalmente reportada de una droga sobre una proteína se representa por una arista. Para ciertas concentraciones de la droga, esta interacción afecta a la proteína y da lugar a un cambio evidente en sus propiedades.

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, para uso en humanos, 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 fructıferos y eficientes para identificar una nueva relación fármaco-diana de interés para la salud, 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.

El uso de redes biomoleculares como las redes fármaco-proteína, las redes de similitud química entre fármacos o incluso las redes de interacción de proteına 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 sıntesis conceptual y análisis integrativos para el diseño de nuevos fármacos [2] .

Un caso de especial interés son las enfermedades tropicales desatendidas (NTD) que incluyen enfermedades como malaria, enfermedad del sueño, mal de Chagas, fiebre amarilla, etc y 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 [5]. Por esta razón las estrategias de reposicionamiento, en particular las basadas en el uso de redes moleculares e integración de datos quimio genómicos, se han convertido en una herramienta fundamental para abordar el problema de identificación de dianas terapéuticas para este tipo de enfermedades [6,7,8]. De esta manera, 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 ya 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)[9,10].

Siguiendo esta línea de trabajo, en esta tesis se presenta un trabajo basado en información quimio genómica de la base de datos TDR targets , dedicada a NTD. Esta base de datos fue introducida por Aguero, Crowther y colaboradores [11,12,13] con el objeto de guiar el proceso de 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. Los autores han integrado información a esta base de datos (TDR targets [11]) 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 [12,14] . Estos trabajos han llevado actualmente la integración de datos quimio genómicos asociados a NTDs a una etapa en la cual los ejercicios de mineríia de datos de gran escala son tan factibles como prometedores.

Es particularmente relevante para el presente trabajo que las 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 ([15],[16],[2]).

En concordancia con este abordaje, en este trabajo nos propusimos analizar las propiedades topológicas de una red quimiogenómica concebida para llevar adelante un programa de priorización y reposicionamiento de drogas. Conceptualizamos una organización de los datos acumulados en TDR basada en una estructura de red compuesta por tres capas principales: una de drogas, enlazadas por relaciones de similitud estructural y subestructura (ver más abajo); una capa de proteínas, que está unida a la capa anterior por aristas de bioactividad experimentalmente reportada; y finalmente una capa con tres tipos distintos de anotaciones, que representan características moleculares y/o funciones celulares que tienen las proteínas adyacentes a estas.

Comenzamos analizando la estructura de la red en la capa de drogas. Esto nos permitió caracterizar y entender cómo se relacionan los espacios químicos, genómicos y de anotaciones de la red. En particular pudimos cuantificar en qué medida similaridades reportadas en el espacio de drogas nos hablaban de similaridades reportadas en el espacio de proteínas, propiedad fundamental sobre la que se basa el desarrollo de cualquier algoritmo de priorización que busque proponer relaciones droga-target novedosas. Posteriormente utilizamos información sobre la evolución temporal de la estructura de la red para reconocer patrones que reflejen sesgos en el desarrollo y descubrimiento de fármacos y moléculas bioactivas. Finalmente llevamos adelante una tarea de priorización para proteínas, a fin de extender la probabilidad de *drogabilidad* (i.e. la probabilidad de que alguna droga presente actividad sobre ella) de una proteína a otra, y lo validamos.

Resulta importante verificar que es posible utilizar la idea de que similaridades entre elementos de un espacio son coherentes con similaridades reportadas en otro para desarrollar algoritmos de priorización que permitan proponer nuevas relaciones droga-target. En nuestro caso esto significa que drogas parecidas tengan como targets a proteínas parecidas y viceversa: proteínas parecidas son apuntadas por drogas similares. Para llevar adelante este análisis chequeamos en qué medida similaridad entre drogas implican que comparten proteínas. Además se estableció que existe una relación entre la cantidad de dominios compartidas por par de proteínas y la asociación de este mismo par a drogas comunes. Con el mismo fin se tomó la capa de targets y anotaciones y se analizaron cantidades relacionadas con el paso de información acerca de drogabilidad entre proteínas, y se hallaron anotaciones de particular interés. Se realizó una proyección sobre la capa de anotaciones y se usaron conjuntos densamente conexos para determinar drogas relacionadas.Luego utilizamos la estructura de la red, con “timestamps” para reconocer patrones que reflejen sesgos

en el desarrollo y descubrimiento de fármacos y moléculas bioactivas, y finalmente hicimos proceso de priorización para proteínas por vecinos cercanos a en una red proyectada, a fin de extender drogabilidad de una proteína a otra, y lo validamos.

La estructura de esta tesis está organizado en 8 capítulos.En el capítulo 1 se da la introducción al trabajo. Posteriormente se describen fundamentos básicos de teoría de redes que se desarrollaran en el resto del trabajo. Por último se explican métodos de priorización y se introduce la forma de validarlos. Luego se establecen métodos de proyección en redes multipartitas y notación acorde.

En el capítulo 3 se introducen las características básicas de la red en sus tres capas y entre ellas. En el capítulo 4 se analiza la topología de la capa de drogas. En primer lugar se analiza la relación entre las semejanza entre drogas y la capa de proteínas.En el capítulo 5 se analizan las características topológicas de la capa de anotaciones en relación con la capa de proteínas.En el capítulo 6 se establece la relación entre grupos densamente conexos de anotaciones y conjuntos de drogas con espectro de acción similar.

En el capítulo 7 estudiamos la evolución temporal de la red, como motivación para la priorización de proteínas en la red, con los métodos establecidos en el capítulo 2.

Capítulo 2

Fundamentos.

2.1Introducción

En este capıtulo se introducen y discuten conceptos fundamentales de redes complejas que serán extensamente utilizados a lo largo de toda este trabajo, (una breve introducción a ideas y conceptos básicos se incluye en el apéndice A). Se describen dos técnicas para predecir enlaces faltantes 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

Formalmente, un grafo $G(\mathfrak{N}, \mathfrak{E})$ consta de un conjunto de entidades $\mathfrak{N}$, y un conjunto de interacciones o aristas $\mathfrak{E}$ establecidas entre pares de las mismas. Los elementos de $\mathfrak{N} := \{n_1, n_2, ..., n_N\}$ se denominan nodos o vértices del grafo y los elementos de $\mathfrak{E} = \{e_1, e_2, ..., e_m\}$ se denominan aristas, arcos o conexiones del mismo. El número de elementos de los conjunto $\mathfrak{N}$ y $\mathfrak{E}$ 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 notaremos al i -ésimo nodo n_i mediante la letra i. Cada arista está asociada con una dupla de dos números identificando los nodos que la misma conecta. A saber , si la k-ésima arista $e_k \in E$conecta los nodos iy$j \in N$ la denotaremos $e_k := e_{ij} = (i, j)$.En ese caso los nodos i y j son nodos adyacentes o primeros vecinos.Una arista que conecta un nodo consigo mismo (e_i) se denomina bucle. Por otro lado, si existe más de una arista e_k , e_k ambas conectando los nodos i y j

, se dice que esos nodos tienen aristas múltiples. Ambos, bucles y aristas múltiples no están incluidos en la definición estándar de grafo, que usaremos aquí.

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 $e_{ij} \in E$ pueden ser descritos por pares no ordenados (i,j) y por tanto resulta $e_{ij} = e_{ji}$, $\forall e_{ij} \in E$. Por el contrario un grafo se dice dirigido si la naturaleza de las conexiones que representa tienen alguna dirección u orientación preferencial. En general para grafos dirigidos se tiene $e_{ij} = (i, j) \neq e_{ji}$ de manera que $e_{i,j} = (i, j)$ implica la existencia d de una conexi´on con sentido bien definido, que va del nodo i al nodo j.

2.3. Priorización

Los algoritmos de priorizacion en redes complejas son esencialmente modelos predićtivos. Supongamos que tenemos un grafo G = G(N, E), donde cada nodo del conjunto $N = \{n_1, n_2, ...n_N\}$ representa un individuo y las aristas $e_{ij} \in E$ representan relaciones entre pares de estos, 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 individuos $N_b = \{n_{k+1}, n_{k+2}, ...n_j...n_N\}$ ha adquirido aún este producto. La idea básica de algoritmo de priorización en redes complejas es utilizar la información del conjuntoN_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 [17]). En este trabajo presentaremos y utilizaremos un algoritmo de priorización. Este es una analogía de un esquema de votaciones *VS* donde cada nodo del conjunto N_a puede transmitir información a sus primeros vecinos.

2.3.1 Esquema de Votación

Sea $G = G(N, E, W)$ 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 sencilla que se puede implementar es un esquema de votación (VS), que en teoría de sistemas de recomendación se conoce como KNN con K=1, 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 [18,19] con el beneficio extra de ser extremadamente eficiente. En un esquema *VS* , dado un grafo pesado $G = G(N, E, W)$ 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

$$f_P(nj) = \sum_{l \in Nei(n_j), l \in N_a} W_{jl} \forall n_j \in N_b \qquad \textbf{(2.1)}$$

donde la suma recorre solo nodos que están simultáneamente en el conjunto de primeros vecinos de n_j, es decir $Nei(n_j)$, 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 2.20 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 conjunto de semillas N_a obtendrá en la ecuación 2.20 un puntaje nulo.

2.3.2 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 compete 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 $L = \{n_i, / n_i \in N_b\}$ ordenada según una magnitud escalar $f_P(ni)$ (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 estos 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 verifiqué

$$L(n_i, u) = \begin{cases} n_i \in P & si \quad f_P(n_i) \geq u \\ n_i \notin P & si \quad f_P(n_i) < u \end{cases} \qquad \textbf{(2.2)}$$

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 es necesario disponer entre los elementos de L algún subconjunto que se sepa a priori pertenece 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 subconjuntos N_α^T, N_α^E , de manera que se verifique $N_\alpha^T \cup N_\alpha^E = N_\alpha \quad N_\alpha^T \cap N_\alpha^E = \emptyset$. El de mayor tamaño N_α^T (supongamos un 90 % de los nodos en N_α) se denomina conjunto de entrenamiento mientras que N_α^E (el 10 % restante) se conoce como conjunto de evaluación referencia. Los nodos en N_α^I serán utilizados como semillas del algoritmo de priorización, mientras que los N_α^E nodos en 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_α^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_α^E, es decir la fracción de verdaderos positivos (TPR) o sensibilidad del predictor de nodos en el conjunto de evaluación.

$$TPR(u) = \frac{1}{|N_\alpha^E|} \sum_{n_i \in L} \delta_i^{TP} = \begin{cases} 1 & si \quad f_P(n_i) \geq u \wedge n_i \in N_\alpha^E \\ 0 & en \quad otro \quad caso \end{cases} \quad \textbf{(2.2)}$$

Por otro lado, la tasa de fallos, es decir la *fracción de falsos positivos* (FPR) viene dada por

$$FPR(u) = \frac{1}{|L|\left|N_\alpha^E\right|} \sum_{n_i \in L} \delta_i^{TP} = \begin{cases} 1 & si \quad f_P(n_i) \geq u \wedge n_i \in N_\alpha^E \\ 0 & en \quad otro \quad caso \end{cases} \quad \textbf{(2.3)}$$

es el número de elementos de la lista L que no pertenecen al conjunto de evaluación .

Notar que ambas cantidades $FPR(u)$ y $TPR(u)$ están normalizadas en el intervalo [0, 1]. La tasa FPR puede ser 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 2.1** 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_α^E un valor más alto de f_P que para cualquier otro nodo de la lista L, esto es

$$f_P(n_i) > f_P(n_j), \forall n_i \in N_a^E, n_j \in N_b \quad \textbf{(2.4)}$$

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 ni pertenece $(n_i \in P, on_i \notin P)$

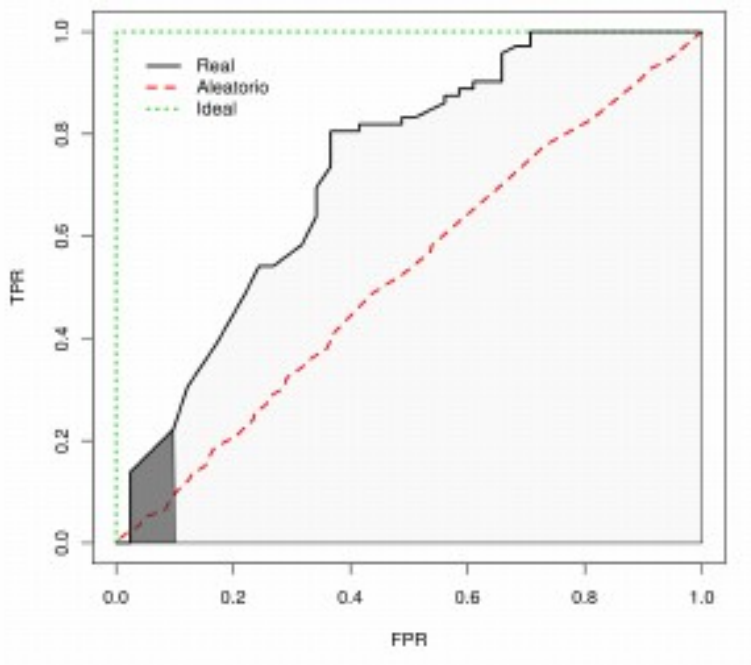


FIGURA 2.1: Curvas ROC: tasa de verdaderos positivos TPR de un predictor (o sensibilidad) en función de la tasa de falsos positivos FPR (1-especificidad del predictor). La curva de puntos verdes corresponde a un predictor ideal, cuya área encerrada es unitaria (AUC=1). La curva roja de trazos representa un predictor aleatorio cuya gréfica oscila sobre la recta identidad y presenta por tanto un AUC~ 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, AUC0.1 = 0.0137. Bajo la corrección de McClish (ver ec.2.24), se tiene AUC0c.1 = 0.546.

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 ~1/2. En general un algoritmo predictivo obtiene valores de AUC en el rango (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 $|N_\alpha^E| << |L|$, el AUC-01 equivale a considerar aproximadamente el 10 % de los nodos con mayor 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 este trabajo se utilizo para este fin la corrección de McClish [20] que se expresa según

$$AUC_\alpha^c = \frac{1}{2}(1 + \frac{AUC_\alpha - AUC_\alpha^{aleat}}{AUC_\alpha^{mx} - AUC_\alpha^{aleat}}) \quad \textbf{(2.5)}$$

$$AUC_\alpha^c = \frac{1}{2}(1 + \frac{AUC_{0.1} - 0.005}{0.1 - 0.005}) \quad \textbf{(2.6)}$$

donde α es el valor máximo de FPR considerado, AUC_c 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.2.24 se consideró el caso de interés en este trabajo, $\alpha = 0.1$.

2.4 Estructura de redes bipartitas y métodos de proyección

Redes Bipartitas

Una red G(N, E) se dice bipartita si existe una partición de N, (N_1 , N_2) que verifica N_1 ∪ N_2 = N,$N_1 \cap N_2 = \emptyset$, y además ningún arco e_i ∈ E une nodos de un mismo conjunto N_1 ó 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 [21] 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 este trabajo, consiste en tomar proteínas de diferentes especies como objetos y grupos funcionales o estructurales de las mismas como conjunto de caracterısticas.

Una extensión natural del concepto de redes bipartitas es introducir la idea de redes multipartitas. En estas últimas existe una partición $N_1, N_2, ...N_m$ que verifica las condiciones $\cup_i N_i = N$, y para cualquier par $i, j \in \{1...m\}$ se cumple $N_i \cap N_j = \emptyset$. Además, ningún par de nodos de una misma clase 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 [22,23], 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.

2.4.1.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 esta en una red de tipo monopartita que contiene uno solo de los dos tipos de nodos. De esta manera la existencia de la otra clase de nodos quedará implícita en los nuevos enlaces. En este tipo de red proyectada, dos nodos comparten una conexión solo 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 E, de manera que ei $j \in E$ con $i \in X$,$j \in Y$.Denotaremos a este grafo $G_{bip}(\mathfrak{N} = X, Y, E)$. El mismo puede ser representado por la matriz de adyacencia $A = (a_{ij})^{nXm}$.

$$a_{ij} = \begin{cases} 1 & si \quad e_{ij} \in E \\ 0 & en \quad otro \quad caso \end{cases} \qquad \textbf{(2.7)}$$

Recordemos que G_{bip} es bipartito por lo cual ningún elemento en E conecta dos nodos del conjunto X o dos nodos del conjunto Y.

La proyección más simple de una red monopartita en una red monopartita quedará dada por

$$w_{i,j} = \sum_{l=1}^{m} a_{il} a_{jl} \qquad \textbf{(2.8)}$$

que escrita matricialmente es

$$W = AA^t \qquad \textbf{(2.9)}$$

Esta proyección resulta en un grafo monopartito, pesado y no dirigido $G_x(X = \{x_1, x_2, ...x_n\}, E_x, W_x)$ de nodos $X = x_1, x_2, ...x_n$, aristas $E_x = e_{i,j}$ con $i, j \in 1, 2...n$ que tienen pesos asociados $W_x = w_{i,j}$. Cada arista $e_{i,j} \in E_x$ El peso entre nodos i y j es simplemente la cantidad de vecinos tipo Y que comparten en la red original

Otra posible proyección bipartita del grafo G_{bip} sobre nodos X, fue definida por Zhou [24], y es conocida como ProbS. Esta proyección resulta en un grafo monopartito, pesado y dirigido $G_x(X = \{x_1, x_2, ...x_n\}, E_x, W_x)$ de nodos $X = x_1, x_2, ...x_n$, aristas $E_x = e_{i,j}$ con $i, j \in 1, 2...n$ que tienen pesos asociados $W_x = w_{i,j}$. Cada arista $e_{i,j} \in E_x$ toma valores

$$\begin{cases} 1 & si \quad w_{i,j} \neq 0 \\ 0 & si \quad w_{i,j} = 0 \end{cases} \qquad \textbf{(2.10)}$$

y los pesos $w_{i,j} \in W_x$ se definen según

$$w_{i,j} = \frac{1}{k_{x_j}} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k_{y_l}} \qquad \textbf{(2.11)}$$

donde la suma corre sobre todos los nodosy_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 W. Notemos que en general en este grafo proyectado ,$w_{ij} = w_{ji}$.

Esta matriz de pesos 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−esima columna y lo notaremos $\hat{a}$. Es decir que para cada elemento$\hat{a_{ij}} \in \hat{A}$ tenemos

$$\hat{a_{ij}} = \frac{a_{ij}}{\sum_j a_{ij}} \tag{2.12}$$

Notar que el denominador de la ec. 2.32 es el grado del j-esimo nodo k_{y_j} . Entonces podemos reescribir la ecuación 2.31 como

$$w_{ij} = \sum_{l=1}^{m} \frac{a_{il}a_{jl}}{k_{y_l}k_{x_j}} = \sum_{l=1}^{m} \frac{a_{il}}{k_{y_l}} \frac{a_{jl}}{k_{y_j}} \tag{2.13}$$

$$= \sum_{l=1}^{m} \hat{a_{il}} \frac{a_{jl}}{k_{y_j}} = \sum_{l=1}^{m} \hat{a_{il}} \frac{(a_{lj})^t}{k_{y_j}} \tag{2.14}$$

$$w_{ij} = \sum_{l=1}^{m} \hat{a_{il}} \hat{a_{ij}^t} \tag{2.15}$$

$$W = \hat{A}\hat{A}^t \tag{2.16}$$

donde el exponente $(a_{ij})^t$ en la ec. 2.35 indica la operación de transposición. La expresión matricial de la ec. 2.37 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 2.37 puede ser trivialmente extendida a la expresión

$$W = \hat{A}I\hat{A}^t \tag{2.17}$$

donde I representa la matriz identidad I ∈ mxm.

2.4.2.Proyección por validación estadística

El método de proyección por validación estadística (StatVal)[25] consiste en asignar enlaces entre aquellos vértices del tipo A que tengan una cantidad estadísticamente significativa de enlaces hacia vértices B de un grado dado. A saber, sea el subgrafo bipartito S con N_A vértices de tipo A y N_B vértices de tipo B, para cada k en la distribución de grado de los vértices de tipo B se construye el subgrafo

inducido S_k de los N_B^k ,elementos B de grado k, y de todos los elementos A unidos a estos. Por tanto la única heterogeneidad que existe en S_k es la que se debe a los nodos de tipo A. Si i y j son nodos de tipo A en S_k , $N_{i,j}^k$ es la cantidad de primeros vecinos en común, y N_i^k , N_j^k sus grados respectivos en S_k. En estas condiciones, si los enlaces entre nodos A y B se produjeran de manera aleatoria, la probabilidad de que i y j compartan $N_{i,j}^k$ vecinos queda dada por una distribución geométrica. Esto permite definir un p-valor para cada par i, j. Los pares i j cuyos p-valores (ajustado por el método FDR) superen un umbral serán considerados como pares validados estadísticamente y se asignará un enlace entre ambos. Este método se repite para todo k , si algún par i, j resulta validado para más de un subgrafo S_k se le asignará un peso correspondiente a la cantidad de veces que resultó validado.

Sea el grafo bipartito inducido S_2, donde los nodos sobre los que proyecto tienen grado 2 ,como el de la figura 2. Los nodos del conjunto en el presente subgrafo presentan variedad de grados, en particular los nodos 4 y 5 presentan cantidades $N_4^2 = 6$, $N_5^2 = 5$ y $N_{4,5}^2 = 5$. Con lo cual la distribución hipergeométrica $p(N_{4,5}^2; N_4^2, N_5^2, N_B^2) = 0.0004$, y si se elige el umbral de valor 0.0005 la arista $e_{4,5}$ queda validada y será parte del grafo proyectado.

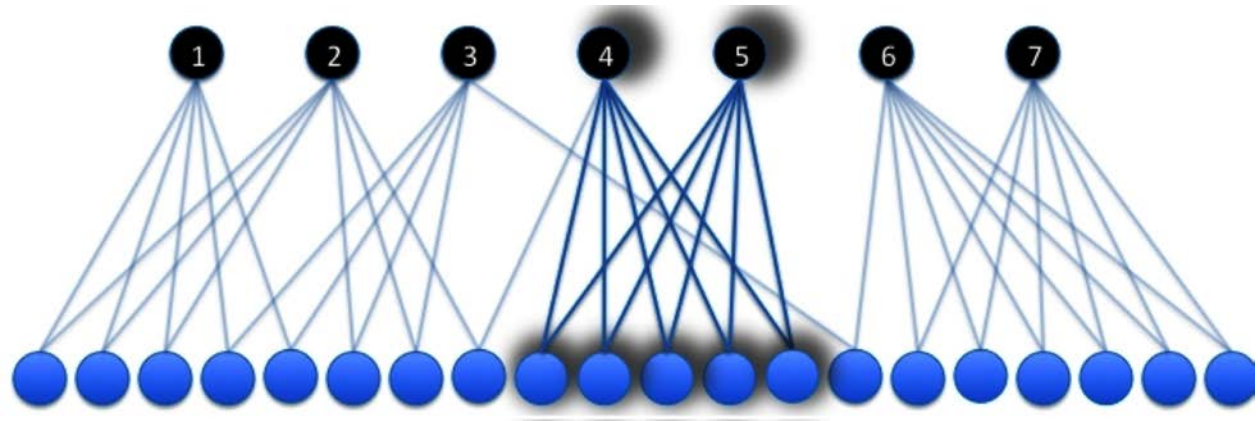


Figura 2.2. Subgrafo S_2 para la proyección por validación estadística. Los nodos azules son sobre los cuales se proyecta, y tienen todos grado 2, mientras que los nodos negros son los que integrarán la red proyectada.Los nodos 4 y 5 presentan cantidades $N_4^2 = 6$ $N_5^2 = 5$(grado propio del nodo 4 y 5 en S_2) y $N_{4,5}^2 = 5$ (vecinos compartidos por los nodos 4 y 5) . Con lo cual la distribución hipergeométrica queda dada por $p(N_{4,5}^2) = 0.0004$, y si se elige el umbral de valor 0.0005 la arista $e_{4,5}$ queda validada y será parte del grafo proyectado.

Alternativamente, existe otra forma de calcular StatVal que llamamos StatVal restricto. La mayor restricción es un resultado de que la estadística usada aumenta los pvalores obtenidos. Esto se logra notando que cada subgrafo, S_k tenía a su vez una separación en p componentes disconexos,S_k^p. Al validar los enlaces dentro de estos componentes en vez de en el subgrafo, el número a incorporar en la función hipergeométrica es N_B^{kp} (cantidad de enlaces en todo el componente del subgrafo),en vez de N_B^k (cantidad de enlaces en todo el subgrafo), que es un número mayor, efectivamente corriendo la distribución hacia la derecha, por tanto aumentando los p-valores y reduciendo la cantidad de enlaces que quedan validados al final.

2.5 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 [26]. 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 representan 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 co-expresión 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 tratarse 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 proteico 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.

2.5.1 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 $M = (G, C)$, donde $G = G_\alpha, \alpha \in \{1, 2, ...M\}\}$ es una familia de grafos $G_\alpha = G(X_\alpha, E_\alpha)$ (dirigidos, no dirigidos, pesados o no pesados) que denominaremos capas de M, y $C = \{E_\alpha, \beta \subseteq X_\alpha \times X_\beta; \alpha, \beta \in 1, 2, ...M, \alpha, \beta\}$ 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 $X_\alpha = x_1^\alpha, x_2^\alpha, ..., x_{N_\alpha}$ y sus conexiones se pueden representar con una matriz de adyacencia $A^{[\alpha]} = (a_{ij}^\alpha) \in \Re^{N_\alpha X N_\alpha}$ cuyos elementos se definen

$$a_{ij}^\alpha = \begin{cases} 1 & si \quad (x_i^\alpha, x_j^\alpha) \in E_\alpha \\ 0 & en \qquad otro \qquad caso \end{cases} \tag{2.18}$$

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

$$a_{ij}^{\alpha\beta} = \begin{cases} 1 & si \quad (x_i^\alpha, x_j^\beta) \in E_{\alpha\beta} \\ 0 & en \qquad otro \qquad caso \end{cases} \tag{2.19}$$

En suma, cada nodo $x_i^\alpha \in X_\alpha$ vive en una capa y puede conectarse a través de dos clases de aristas, unas denominadas conexiones intra-capa E_α que las unen a nodos de su misma capa, y otras denominadas conexiones inter-capa E, 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. Un ejemplo que resulta de particular interés para esta tesis son las redes multipartitas definidas en la sección anterior.

Capítulo 3

Estructura de la red

3.1.Introducción

En este capítulo nos ocuparemos de describir y caracterizar la red multicapa que utilizaremos en este estudio. La misma está constituida por una capa de drogas, una de proteinas (que denotaremos equivalentemente como targets o blancos) y una capa que incluye nodos que representan tres tipos de conceptos que pensamos relevantes para nuestro problema. Estos involucran las ideas de: Dominios PFAM, grupos ortólogos y vías metabólicas, relacionadas con la presencia de subunidades estructuras de proteínas, similaridad de secuencia debido a la existencia de un origen evolutivo en común y la participación en reacciones químicas similares, respectivamente. Características básicas y un esquema de la organización de nuestra red se incluyen en la **tabla 3.1** y en la **figura 3.1** respectivamente.

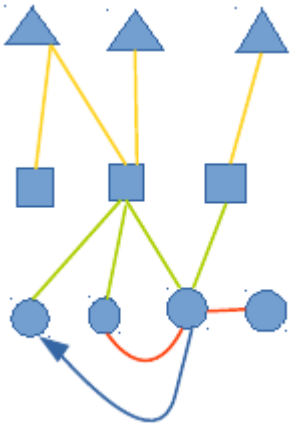


FIGURA 3.1: Esquema de la red multicapa. Los triángulos representan nodos en la capa de anotaciones que pueden ser o bien dominios pfam , grupos de ortología o caminos metabólicos. Los cuadrados representan las o proteínas, que están unida a nodos de la capa anterior. Las proteínas que están conectadas a la capa de drogas (círculos) se dicen drogables y el enlace representa bioactividad. Las drogas están relacionados a través de dos tipos de similaridad subestructura(flecha azul) y Tanimoto (flecha roja).

	Nodos	Enlaces
Drogas	14,888,034	15.950.829(subestructura) 23.028.356(Tanimoto)
Drogas bioactivas	177.506	576.147(subestructura) 2.067.256(Tanimoto)
Proteínas	168.622	325.843 (a drogas)
Proteínas drogables	6.051	325.843 (a drogas)
PFAM	2252	219.313
Vías metabólicas	145	77.389
Grupos ortólogos	2.789	51.702

TABLA 3.1: Cantidad de nodos por tipo y cantidad de enlaces en nuestra red**.**

3.2.1.Similitud entre drogas

La capa de drogas tiene 14,888,034 nodos asociados a drogas de efecto conocido y otros compuestos químicos, caracterizados por su peso molecular y su estructura química. Al mismo tiempo, poseemos información sobre la semejanza de compuestos a partir de dos nociones complementarias de similitud estructural: similitud de tipo Tanimoto y de subestructura. Las relaciones de Tanimoto quedan dadas por similaridad de huellas moleculares que comparten las drogas. La huella molecular es un conjunto de bits que describen en una biomolécula la presencia de estructuras o grupos químicos característicos de manera que si está presente una característica el correspondiente bit será un 1, de lo contrario será un 0. Las relaciones de subestructura, por otra parte, quedan definidas cuando una molécula contiene a otra como subconjunto estructural.

La **figura 3.2** , muestra un conjunto de cuatro drogas con Tanimoto igual a 1 , es decir comparten todas las características moleculares marcadas con círculos, entre todas, Sin embargo como se puede apreciar ninguna de las moléculas es un subconjunto estructural de la otra. Esto ilustra el aspecto complementario entre ambas relaciones de similitud que abordaremos más abajo.

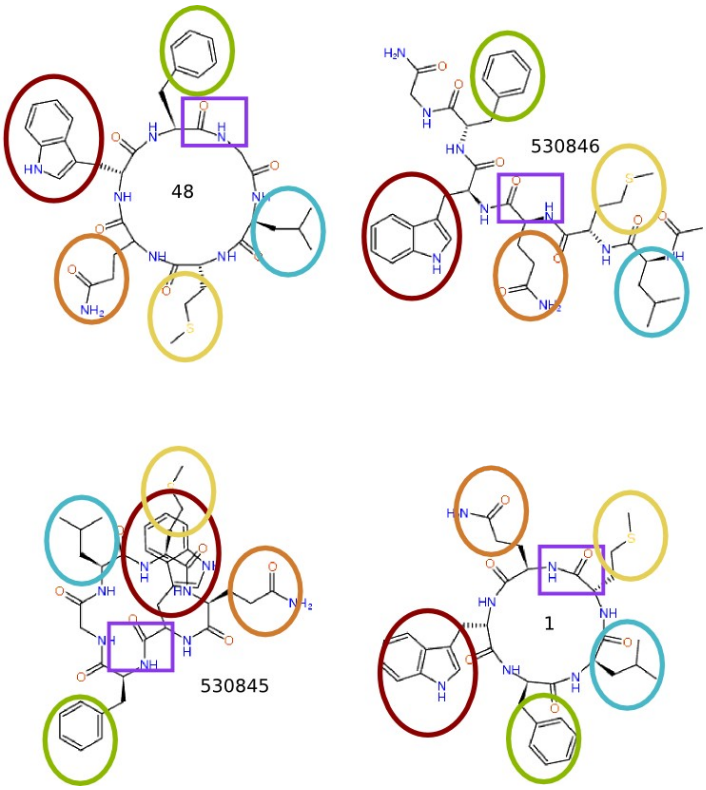


FIGURA 3.2: En la figura se observan moléculas con similitud de Tanimoto 1. En distintos colores se marcan grupos químicos utilizados en las huellas moleculares. Se ve que en todos los compuestos están los mismos grupos químicos lo que da como resultado la similaridad tanimoto igual a 1. Cómo se puede apreciar no son subestructuras unas de otras.

3.2.2.Tanimoto y subestructura

Como se introdujo antes, la capa de drogas está compuesta por enlaces de dos tipos: Tanimoto y subestructura. Para poder calcular la relación de Tanimoto primero se establecen cuantas y cuales caracteristicas incluir en la huella molecular. En el caso de la red TDR las huellas sólo contienen información sobre estructuras bidimensionales. El tamaño de la huella asociada a cada compuesto es de 512 bits.

Una vez definido el tipo de huella para representar a la molécula es necesario definir una noción de semejanza. Estas dependen de la cantidad de 1s en la huella de cada compuesto, siendo el par de compuestos AB, denotamos a y b a la cantidad de unos en A y B, respectivamente. A su vez, llamaremos *c* a la cantidad de unos en común(i.e. caracteristicas compartidas). Para el cálculo de similaridad Tanimoto se considera el índice de Jaccard establecido entre un par de huellas, que se define $sim(A,B) = \frac{c}{a+b-c}$.Para el caso de nuestra red sólo se tuvieron en cuenta relaciones

Tanimoto con un valor mayor a 0,8 que es un valor razonable según fue reportado en [27] . Esto es importante, ya que garantiza que el tamaño de la red sea apto para trabajar, y permita hacer estudios de su estructura , evitando cuellos de botella computacionales. Con este criterio quedan establecidos 23.028.356 de enlaces entre drogas (**tabla 3.1**).

Respecto a la relación de subestructura entre dos drogas, A y B, toma un valor 1 droga A es subestructura del complejo B. Es interesante notar que, contrariamente a lo que sucede con la similitud de Tanimoto éste tipo de asociación es direccionada, y en principio no tiene asociado un peso. En la red, pudimos establecer 12.076.297 relaciones de sub-estructura entre compuestos químicos.

Dadas las métricas de similitud adoptadas, nos propusimos identificar dos tipos de estructuras relevantes: conjuntos fuertemente *conexos* definidos a partir de enlaces de subestructura (i.e. conjuntos donde la relación dirigida de es-subestructura-de permitía establecer un camino entre cualesquiera par de drogas del conjunto) y *cliques* de drogas para enlaces con Tanimoto =1.

Llamamos a estas estructuras clusters-S y clusters-T respectivamente. Estos conforman clusters *identitarios* , en el sentido que sus componentes resultan indistiguibles desde el punto de vista de una u otra similitud. Es importante señalar, sin embargo, que al estar basadas ambas relaciones en representaciones simplificadas de la estructura química real, las mismas pueden no contemplar la totalidad de grados de libertad moleculares. En particular, el hecho de que exista un identidad entre dos moléculas según la métrica de Tanimoto no implica necesariamente que una sea subestructura de la otra (ver **Figura 3.2**).

Es importante destacar que así definidos los clusters se corrobora que los valores de las similitudes entre nodos de diferentes clusters son los mismos para cualquier dupla. Por ejemplo, sean $x_A x_A$ nodos en el cluster A y $x_B x_B$ nodos en el cluster B, siempre se comprueba $T(x_A, x_B) = T(x_A, x_B)$. De esta manera la relación de Tanimoto entre dos clusters-T queda bien definida a partir de la asociación Tanimoto de cualquiera de los pares inter-clusters. Algo similar ocurre para los clusters-S en el sentido que las relaciones de subestructura de elementos de un mismo cluster-S hacia elementos externos se mantienen idénticas. Finalmente asumimos que drogas no incluidas en clusters identitarios tipo-S o tipo-T conformaban su propio cluster de tamaño 1.

La descripción en términos de clusters-S y clusters-T permite implementar una reducción importante del tamaño de la red sin pérdida de información. Un estudio más detallado de las características de estas estructuras se incluye en la siguiente sección. Aquí simplemente mencionamos que al reemplazar a todos los elementos de un cluster por un único nodo, resulta un grafo de subestructura con una reducción en el número de nodos del casi 4 % mientras que la reducción en el número de aristas es del 17 %. a 103.550.339 reducción Con respecto a Tanimoto resultan 1.330.235 clusters y 23.028.356 aristas. Esto implica una reducción en el número de nodos del 10 % y una reducción en el número de aristas del 47 %.

Finalmente, para seguir adelante decidimos asignar un peso a cada enlace del espacio de subestructura, que tenga en cuenta la asimetría en el peso molecular de los vértices asociados. Definimos entonces

$$Sub(X,Y) = \frac{\acute{m}in(w_x, w_y)}{máx(w_x, w_y)} \quad \textbf{(3.1)}$$

donde *x* e *y* son clusters-S de pesos moleculares w_x y w_y respectivamente.

Se toman valores máximos y mínimos ya que debido a los problemas de representación incompleta de compuestos químicos podría ocurrir que $w_X > w_Y$, aún cuando $x \subset y$. De esta manera estamos dando más pesos a una relación de similitud por subestructura, cuanto más parecidos sean los tamaños de las moléculas asociadas.

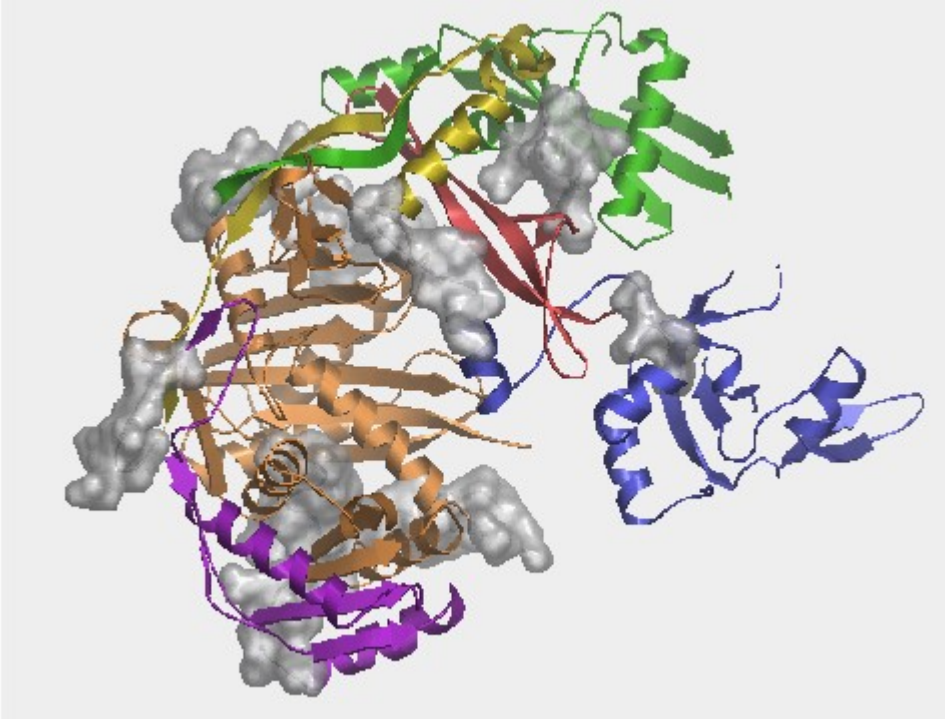
3.2.2.Anotaciones

En esta subsección introduciremos con algún detalle las componentes de la capa de anotaciones consideradas para caracterizar estructuralmente a las proteínas de nuestra red.

Dominios pfam:

Las proteínas son secuencias de aminoácidos que presentan una estructura que en general puede caracterizarse de acuerdo a cuatro niveles de descripción. El primer nivel involucra a la cadena de aminoácidos, es decir la secuencia. Al plegarse la misma da origen a la aparición de estructura secundaria que incluye arreglos en hélices y láminas (ver figura 3.3 , en particular la estructura roja está marcadamente formada por láminas). La estructura terciaria corresponde a la cadena ya completamente plegada. Los dominios PFAM , se pueden identificar en la estructura terciaria (porciones coloreadas de la proteina de la figura 3.3) y se corresponden con

unidades estructurales que se pliegan de forma más o menos independiente una de otra y que en ciertos casos pueden asociarse con funciones proteicas específicas que se preservan a lo largo de distintas especies [28] .En nuestro caso, consideramos los 2252 dominios PFAMs descriptos en http://PFAM.xfam.org/.



Pfam family	Colour
ArgoN (PF16486)	
ArgoL1 (PF08699)	
PAZ (PF02170)	
ArgoL2 (PF16488)	
ArgoMid (PF16487)	
Piwi (PF02171)	

FIGURA 3.3: Proteína *ago1_human* , en colores se ilustran sus dominios PFAM mientras que en gris se muestran las zonas de la proteína que no han sido asignadas a dominios. Los dominios PFAM son estructuras que se repiten en varias proteínas. Se puede observar también la estructura secundaria de las proteínas, formada por hélices y láminas.

ortologia:

Dos proteínas son homólogas si descienden de una misma proteína ancestral. Usualmente se puede concluir que dos secuencias son homólogas sobre la base de la alta similitud en secuencia que las mismas presentan. Cuando acontece que las proteínas *homólogas* pertenecen a la misma especies se dice que son *paralogas*. Proteínas *homólogas* presentes en diferentes especies se denominan *ortólogas*. Nosotros consideramos 2789 grupo de ortología definidas en la version 4 de orthoMCL (http://orthomcl.org/orthomcl/home.do) que utiliza métodos de clustering markoviano sobre las secuencias de cada proteína para reconocer grupos de proteinas de secuencias similares.

Finalmente consideramos 145 vías metabólicas, que también denotaremos *pathways*. Las mismas corresponden a reacciones enzimáticas encadenadas que ocurren dentro de una célula. En ellas un sustrato inicial se transforma y da lugar a productos finales que a su vez pueden ser utilizados como sustrato de reacciones subsiguientes. La información de pertenencia a una dada vía metabólica (pathway) corresponde a la extraída de http://www.genome.jp/kegg/pathway.html.

Capítulo 4

Caracterización de la capa de drogas

En este capítulo nos ocuparemos de describir la capa de drogas y su relación con la de proteínas. Se establecen los vínculos que hay entre los distintos tipos de similitudes en las capa de drogas, Tanimoto y subestructura. Se comparan los clusters que las dos relaciones de similitud generan y se analiza su homogeneidad, a fin de establecer si estos clusters fueron bien elegidos y si representan distintos tipos de información sobre las drogas. Además se estudian las relaciones entre esos enlaces de similitud y los enlaces de bioactividad sobre proteínas.

4.1.Parejas de subestructura vs. parejas Tanimoto en clusters

Como se mencionó antes dado que las relaciones subestructura y Tanimoto están basadas en simplificaciones de la estructura molecular, resulta apropiado estudiar el solapamiento entre los dos tipos de clusters formados para así analizar las coincidencias y diferencias encontradas.

Para este fin, consideraremos a la persistencia P_{er} , una medida que permite cuantificar la relación entre la composición de clusters-S y clusters-T. Dado un cluster de subestructura *clusA* de la partición S, definimos el coeficiente de persistencia por cluster, $P_{er}(clusA)$, como la fracción de pares de drogas que se encuentran en clusA que también aparecen juntas en un mismo cluster-T :

$$P_{er}(ClusA) = \sum_{i,j \in ClusA} \sum_{ClusB} \frac{\delta_{i,j}^{ClusB}}{N_{ClusA}(1-N_{ClusA})/2} \qquad \textbf{(4.1)}$$

con $\delta_{i,j}^{ClusB}$ la delta que es uno si el par i,j aparece en el cluster B y N_{ClusA} la cantidad de drogas en el Cluster A. Así mismo, realizamos este cálculo para estimar la persistencia de clusters-T en agrupamientos-S.

En los paneles de la figura **figura 4.1** se muestran los resultados de persistencia de clusters-S en clusters-T y viceversa.

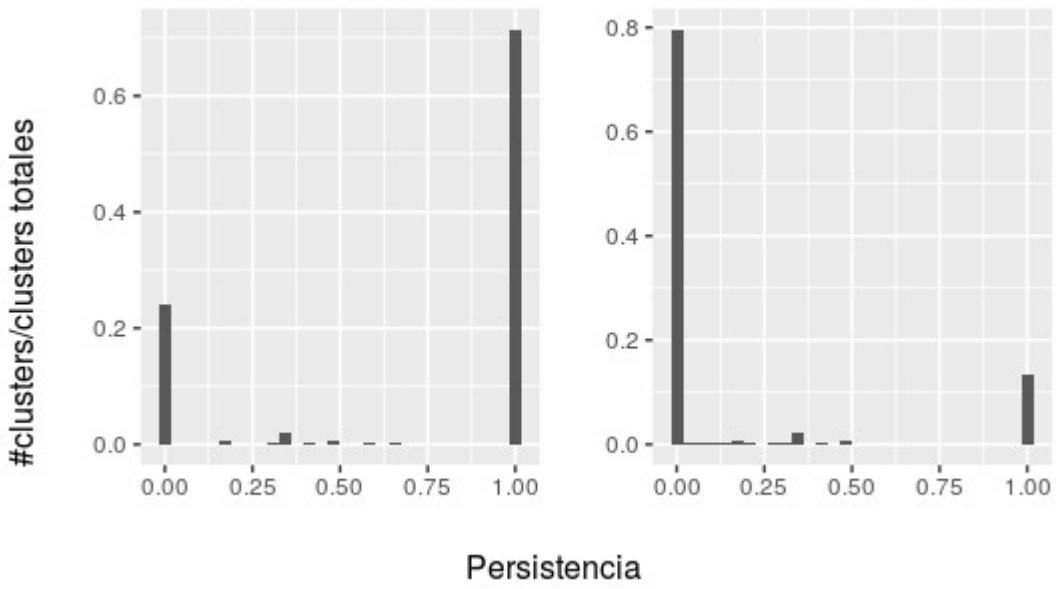


FIGURA 4.1: Panel derecho gráfico de barras con la fracción de clusters S que presentan un dado pair ratio S en T, según se definió en la ecuaci 3.1.Lado izquierdo proporción de clusters S para cada persistencia de S en cluster T. Los valores de persistencia grandes indican que varios de los pares de drogas en un dado cluster S también están juntos en un cluster T.

Conociendo los valores de persistencia de todos los clusters-S fue posible estimar un valor de persistencia promedio (o persistencia de la partición) sobre todos ellos, para cuantificar en qué medida se conserva la estructura de la partición en Cluster S al hacer la de cluster T. Análogamente se calculó la persistencia promedio de clusters T en S. Los valores obtenidos fueron P(S->T)= 0,73 y P(T->S)= 0.15 lo que implicaría que los clusters tipo S tienden a permanecer juntos en la partición inducida por los clusters Tanimoto.

Esto tiene sentido pues el requerimiento de identidad por subestructura es más fuerte que el inducido por niveles máximos del índice de Tanimoto, en el que sólo importa que grupos funcionales se repitan, sin importar por ejemplo orden o posición espacial. La condición de "subestructura" , en cambio, es una condición fuerte que implica que varios elementos constitutivos aparecen de forma contigua. De hecho, es sabido que compuestos isómeros presentan las mismas características de cadena (es decir tendrán Tanimoto 1) pero distinta configuración espacial con lo que no son en general subestructura uno de otro. Dado que es esta misma diferencia la que muchas veces es responsable de que compuestos posean actividades biológicas significativamente distintas, desde un punto de vista químico y biológico es deseable,

y de hecho ocurre en la red, que los clusters de subestructura separen drogas isoméricas en clusters diferentes.

4.2.Homogeneidad de clusters S y T

Hasta aquí hemos agrupado los tipos de relaciones Tanimoto y subestructura en clusters de identidad por similaridad bioquímica. Como ya se ha indicado ninguno de los dos criterios captura la totalidad de los grados de libertad que existen en una molécula. Por eso es interesante estudiar algunas propiedades de los clusters S y T e investigar qué tan homogéneos son dichas estructuras respecto a otras condiciones como por ejemplo peso molecular y bioactividades compartidas.

4.2.1.Analisis de peso molecular

La **figura 4.2** muestra la distribución de pesos moleculares de estructuras T y S. Se observa que las distribuciones son similares, con una cola pesada hacia pesos grandes. Para los cluster-T se obtuvo un peso molecular promedio de 427 g/mol con desviación estándar de 251g/mol , mientras que para la distribución de pesos de clusters S se obtuvo un promedio de 442 g/mol y una desviación estándar de 262g/mol. Las desviaciones estándar encontradas implican que hay una gran variedad de clusters con masas promedio muy disímiles. Notamos además que los tamaños de las drogas que estamos considerando (reportados por pesos promedios) son típicamente mucho menores que el de las proteínas, que se encuentra en el orden de 10.000 g/mol.

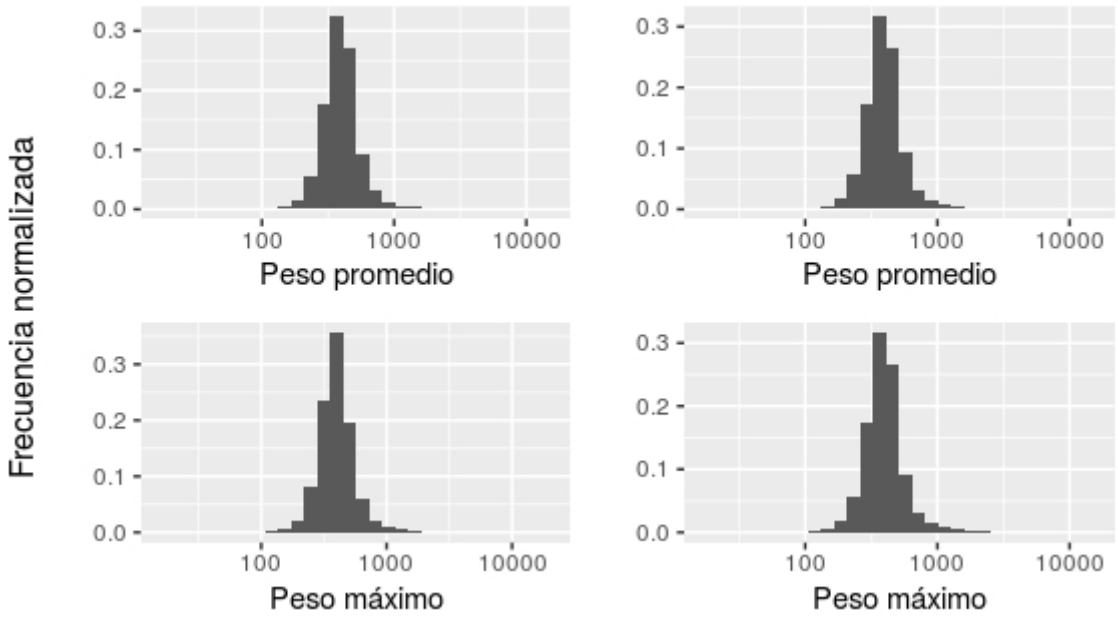


FIGURA 4.2: En el panel de la Izquierda corresponde a los pesos de clusters Tanimoto derecha corresponde a los pesos de clusters S.

Encontramos así mismo un alto grado de homogeneidad de los clusters en cuanto al peso molecular de los elementos que los componen. El 98% de los cluster S tiene diferencias máxima de de peso menores a 1, y la situación de clusters con mayores diferencias se explica como resultado de que la relación de subestructura es inexacta ya que parte de una representación simplificada de la molécula . La situación para los clusters Tanimoto es diferente. La diferencia de pesos entre el mayor y menor elemento es mayor a 1 g/ml, para más de 3/4 de los clusters T, y mayor a 50 g/ml para 1/4 de los mismos. En este caso encontramos que típicamente la diferencia entre el peso máximo y mínimo encontrado para moléculas de un mismo cluster T es del orden del 10% del peso promedio del mismo (ver **figura 4.3**). En [29,30,31] se afirma que la similaridad de Tanimoto tiende a ser mayor para relaciones que involucran moléculas grandes, debido a que en general poseen más variedad de características y tienen más chances de compartir bits con otra proteína. Este sesgo implica que $Tani(A,B)$ será mayor si A o B son un par de moléculas de tamaños disímiles, en contraste con el caso en que A y B son moléculas ambas de tamaños pequeños y similares . Se observa que el peso máximo correlaciona con la desviación estándar de pesos, como se observa en la **figura 4.3**. Las distancia intercuartil corresponden al 10% de la mediana lo que habla que en todos los casos los clusters incluyen drogas con pesos moleculares similares.

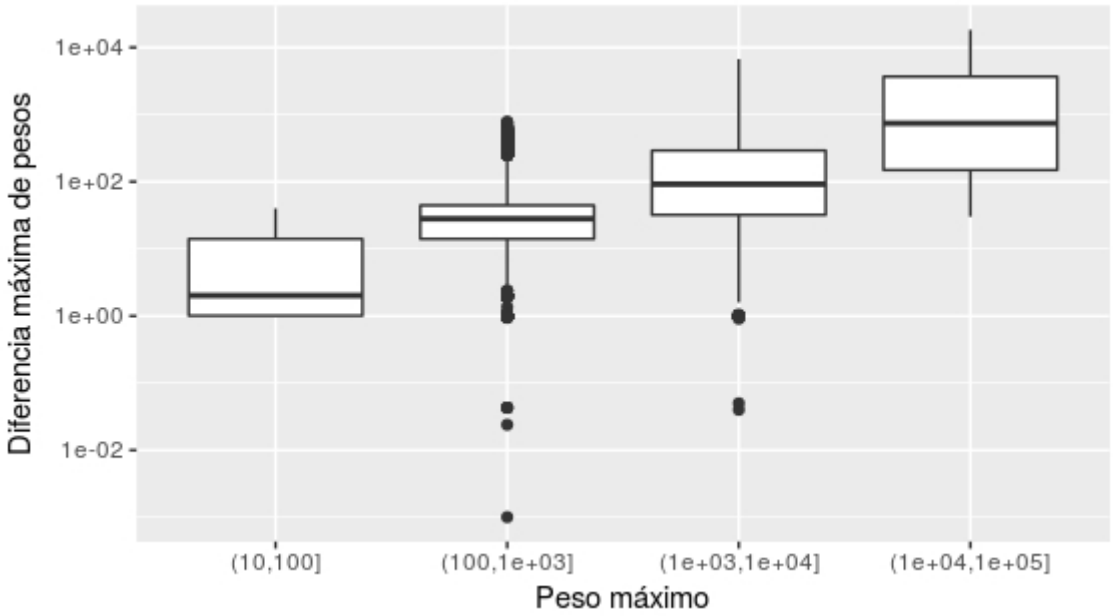


FIGURA 4.3:.Gráfico que caracteriza la desviación estándar de los pesos en los clusters T en función de su peso máximo Se observan con ejes logarítmicos que las medidas de variación de peso incrementan con el peso máximo del cluster.

4.2.2.Homogeneidad de blancos

Como se mencionó anteriormente hay 177506 drogas en la red con actividades sobre proteínas. Dado que los clusters reflejan algún tipo de identidad química entre drogas, es deseable investigar si esto se refleja en la existencia de bioactividades similares para drogas pertenecientes al mismo cluster. Esto nos permitirá, a partir de las asociaciones entre compuestos químicos, inferir nuevas bioactividades y será particularmente útil a la hora de recomendar nuevas drogas en las que ya se conoce alguna actividad.

Con esto en mente, se analiza la tasa de blancos compartidos de todas las drogas (con al menos un enlace reportado hacia un *target* proteico) de un mismo cluster. Esto se cuantifica tomando el índice de Jaccard de pares de drogas en los clusters, y promediando esta magnitud sobre todos los pares. Es decir para cada par de drogas, A B, con blancos $T_A T_B$ se toma $\frac{T_A \cap T_B}{T_A \cup T_B}$ y luego se promedia sobre el total de pares de drogas del cluster.

Encontramos que si bien más de un tercio de los clusters Tanimoto presentan Jaccard nulo, el 50% de los clusters Tanimoto presenta un Jaccard promedio de valor 1. Para clusters S esto esto ocurre para el 80% de los clusters. Este primer resultado sugiere que la coaparición en un cluster S puede asociarse fuertemente al hecho de compartir targets proteicos.

Para analizar la significancia estadística de las asociaciones encontradas entre coaparición en clusters identitarios y número de coincidencias en targets asociados, se hizo el análisis de frecuencia del Jaccard entre todos los pares posibles en un mismo cluster y se contrastó con pares elegidos al azar. Para ello se considerò pares de drogas seleccionadas al azar de todas las disponibles, respetando que tengan el mismo número de bioactividades que las de los datos de pares intra-clusters. En particular por cada par de drogas bioactivas dentro de un cluster se tomaron 10 otros pares formados al azar con drogas de mismos grados que los pares originales, pero en distintos clusters.

Para clusters Tanimoto hay 62.241 pares de drogas en un mismo cluster. Entre todos ellos(**figura 4.4** panel superior), 86 % tiene índice de Jaccard mayor que 0 (al menos un blanco compartido), 65 % mayor que 0.5, y 60 % igual a 1 (exactamente los mismos blancos). Estos valores resultaron significativamente superiores a los obtenidos a partir de pares de drogas tomadas al azar, respetando el número de bioactividades original. Para este control se obtuvo que 42 % de pares control tiene en promedio un índice de Jaccard mayor que 0, 9.99 % mayor 0.5, y sólo el 4% igual a 1.

De forma análoga, para los 9.271 pares de drogas pertenecientes a clusters de subestructura (**figura 4.4** panel inferior), encontramos que el 85 % tiene índice de Jaccard mayor que 0, 54 % mayor que 0.5, y 42 % igual a , mientras que el control random reportó que el 78 % tenía en promedio un índice de Jaccard mayor que 0 pero ninguno mayor que 0.5.

Estas observaciones indican que, si bien el hecho de pertenecer al mismo cluster no garantiza que dos drogas tengan exactamente las mismas bioactividades, si garantiza que estas, en promedio, tendrán más bioactividades en común que cualquier otras dos tomadas al azar de distintos clusters. De esta forma vemos que las estructuras generadas a partir de estas nociones complementarias de similaridad estructural, demuestran ser relevantes con respecto a las bio-actividades reportadas. En particular, el hecho de que pares al azar en distintos clusters S no superen el Jaccard de 0,5, refuerza la importancia de subestructuras identitarias basadas en esta noción de similaridad para definir bioactividades.

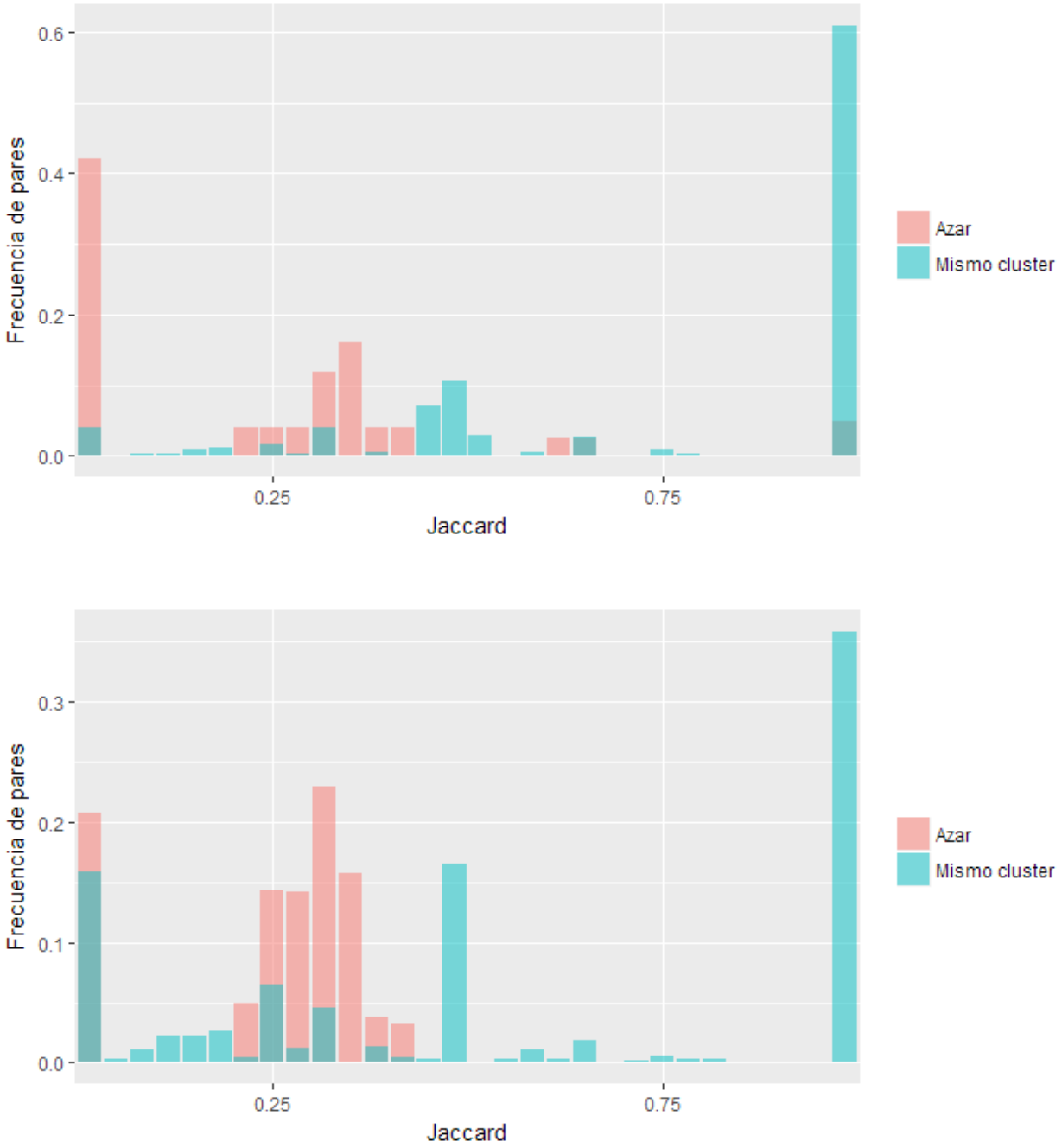


FIGURA 4.4: Frecuencia de pares normalizada para el índice de Jaccard entre drogas bioactivas de un mismo cluster y tomadas al azar.En el panel superior clusters Tanimoto, debajo clusters subestructura

4.3.Grado de Similaridad Química y bioactividades compartidas

Hasta aquí el análisis nos ha mostrado que hay una relación positiva entre blancos compartidos y la pertenencia a un mismo cluster tipo S o tipo T, o dicho de otra manera entre drogas con similaridades de valor 1. Abordaremos ahora, la cuestión de si en general, el grado de similaridad química provisto por los índices de Tanimoto y/o subestructura, correlaciona con la tendencia a presentar targets en común.

Para analizar esto, estimamos la fracción de targets compartidos (utilizando índices de Jaccard) entre drogas vinculadas a través de enlaces de Subestructura o Tanimoto de un valor dado. Claramente, observamos en la figura 3.9 una relación monótona creciente entre los índices de similaridad química entre pares de compuestos quımicos y el número de blancos compartidos. (panel izquierdo y central). Observamos que en ambos casos, Tanimoto y sub-estructura, un valor de similaridad entre compuestos cercano a 0.7 - 0.8, se corresponde con una fracciòn de targets compartidos del orden del 40%. El panel derecho muestra finalmente la relación lineal que actúa como curva de calibración entre los dos índices de similaridad.

Estos resultados muestran en particular que ambas nociones de similitud, aunque alternativas complementarias, son coherentes entre sí. Más aún esto refuerza lo visto antes sobre que la relación subestructura es más fuerte que la de Tanimoto, ya que para alcanzar un mismo valor de Jaccard se necesita un menor peso de la relación de subestructura. Se debe notar que ya que las similitudes por sub−estructura son dirigidas, es necesario considerar aristas en ambas direcciones, solo aristas de salida, o solo aristas de entrada. Debido a que no se observan diferencias significativas entre estos casos, mostramos el resultado para el caso en que no se consideran las direcciones.

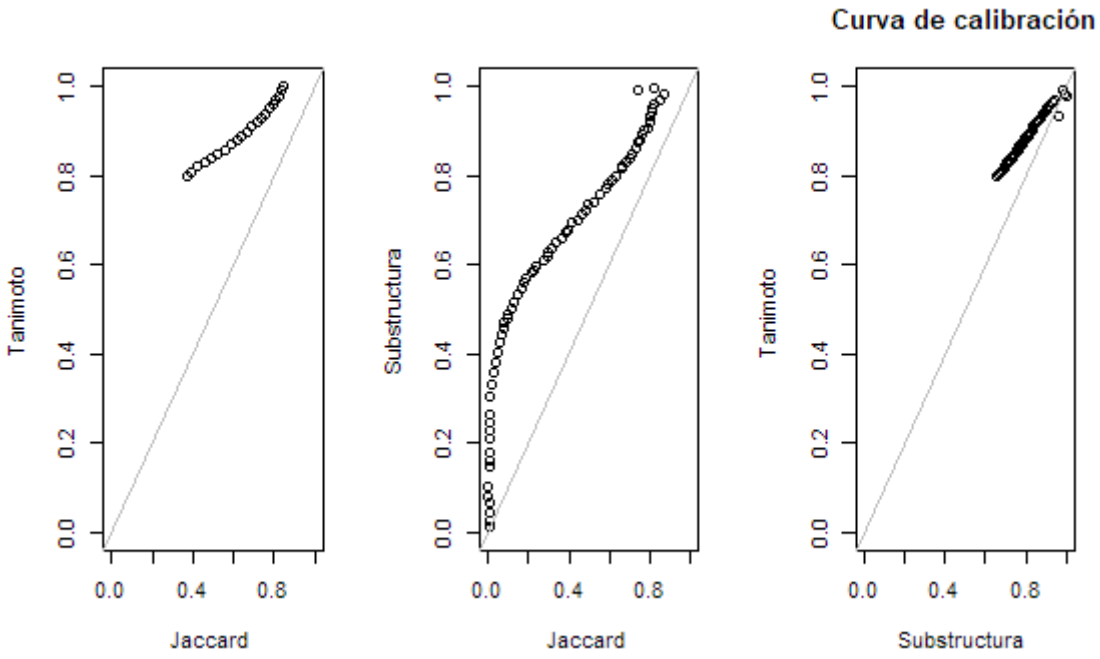


FIGURA 4.5: En la izquierda se encuentra la relación Tanimoto promediada para pares de drogas con un cierto Jaccard de blancos. El gráfico del centro es igual al anterior pero para la similaridad de subestructura. El gráfico de la derecha muestra el resultado de calibrar los dos gráficos en función del Jaccard de blancos.

4.4. Conclusión

Como se vió anteriormente, el análisis de estructuras identitarias a partir de links con $Tani(A, B) = 1$, nos permitió reconocer clusters (clusters-T, clusters-S) que permiten realizar una descripción en términos de estructuras 'coarse grained' y reducir así el tamaño efectivo de la red en ambos tipos de métricas de similaridad. En este capítulo establecimos las relaciones entre estos dos tipos de clusters. Observamos que en promedio la relación de subestructura es más fuerte que la de tanimoto, en el sentido de que la persistencia es mucho mayor al pasar de una partición de la capa en clusters S a T que al revés, lo que indica que los clusters-S están contenidos en mayor grado dentro de clusters-T que a la inversa. Finalmente establecimos un criterio de comparación y calibración entre el valor numérico de la similitud por subestructura y Tanimoto. La escala de calibración entre ambas medidas (panel derecho de la **figura 4.5**) permite apreciar que para un mismo valor de similaridad la relación de subestructura se corresponde con una mayor coincidencia de blancos compartidos.

Capítulo 5

Capa de targets y anotaciones

En este capítulo estudiaremos la capa de proteínas, y su relación con la capa de anotaciones. En el mismo sentido se analizará la capacidad de las anotaciones de PFAM, vías metabólicas y ortologia, para mediar información de drogabilidad entre proteínas en distintas especies. Finalmente se estudiará la capa de anotaciones proyectada, a fin de encontrar grupos de anotaciones específicos, densamente conectados, que reflejen propiedades de proteinas drogables.

5.1.caracterización de distribución de grados de conjuntos de filiación a afiliados

Comenzamos el análisis de la red bipartita: targets-anotaciones estudiando aspectos básicos de conectividad a partir de las distribuciones de grado observada. Se obtuvieron las distribuciones de la cantidad de dominios por blanco y la cantidad de blancos por dominio, **figura 5.1** . Como se observa en el panel de la izquierda ,cada target tiene a lo sumo asociados 10 dominios PFAMs. Esto es natural ya que los dominios representan subestructuras moleculares y la existencia de un límite a la cantidad de dominios es una consecuencia natural de la finitud de las cadenas proteicas.

En el panel derecho de la figura se observa que la mayor parte de los PFAMS presentan pocas proteínas asociadas. Sin embargo existen algunos dominios PFAM particularmente promiscuos, presentes en más de 500 proteínas. Dichos PFAM corresponden a dominios funcionales generales asociados muchas veces a estructuras que llevan adelante funciones biológicas tales como: transducción de señales, metabolismo y reconocimiento de motivos de RNA, entre otros.

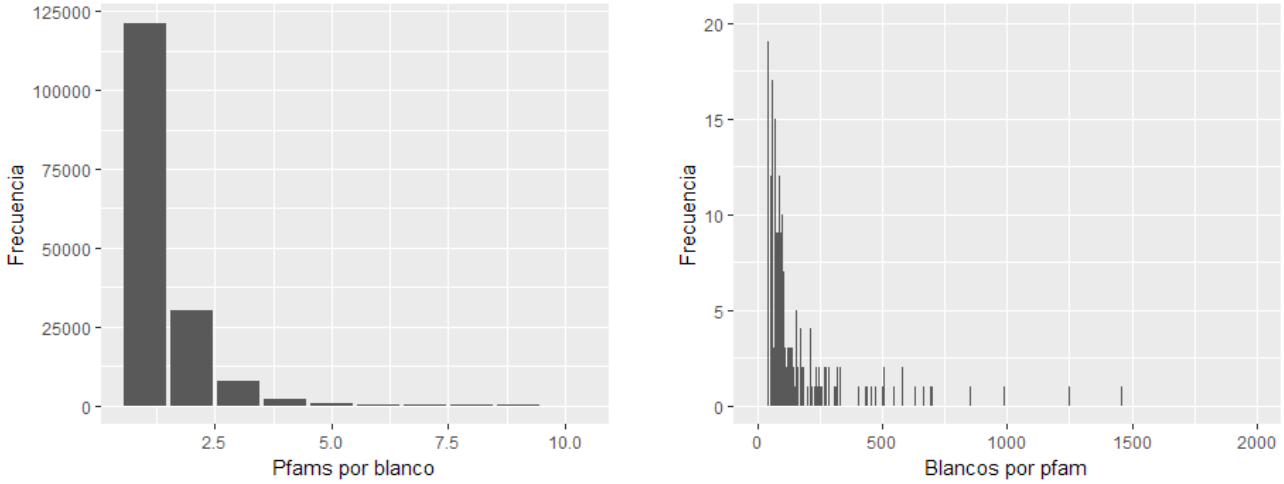


FIGURA 5.1: Izquierda histograma de la cantidad de PFAMs a los que están afiliados las proteínas.Derecha histograma cantidad de proteínas afiliadas por PFAM.

Para el caso de grupos de ortólogos, por definición cada target está asociado a 1 sólo grupo. Así mismo, la **figura 5.2** muestra que la mayoría de los grupos ortólogos no tienen más de algunas decenas de proteínas . Por otro lado existen grupos con más de una centena de proteínas, y que deben corresponder a las funciones celulares más antiguas y esenciales para la vida. El grupo ortólogo con más proteínas es OG4_126558 que está relacionado con el metabolismo del ATP, de fundamental importancia para la respiración celular en todos los seres vivos.

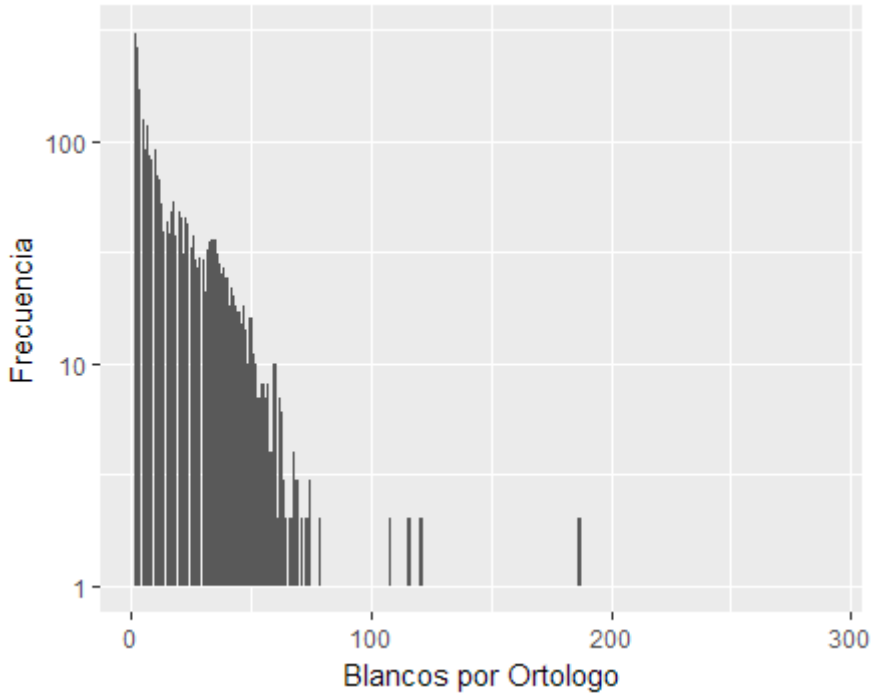


FIGURA 5.2: Distribución de cantidad de proteínas por grupo ortólogo.

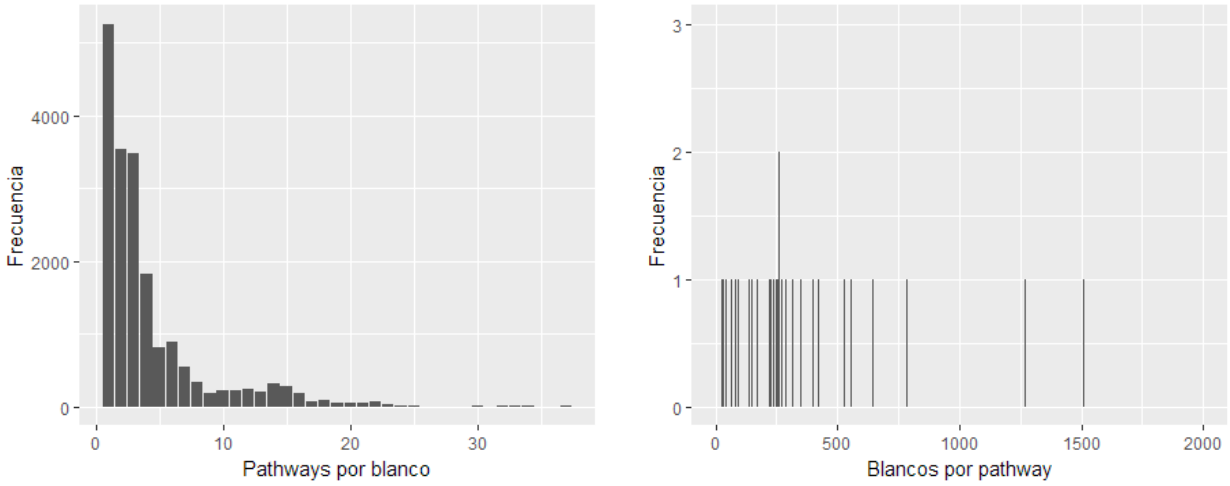


FIGURA 5.3: Izquierda histograma de la cantidad de vías s a los que están afiliados las proteínas.Derecha histograma cantidad de proteínas afiliadas por vías.

La cantidad de pathways por blancos (figura 5.3) es esperable ya que en general las proteínas participan de unas pocas vías metabólicas conectadas. La distribución de proteínas por pathway es bastante amplia y relativamente uniforme, lo cual refleja la variedad y los distintos grados de complejidad y jerarquías de rutas metabólicas.

5.2.Relevancia de anotaciones

5.2.1.Drogabilidad y anotaciones

Como mencionamos anteriormente, un objetivo de nuestro trabajo es priorizar proteínas a través de un algoritmo de primeros vecinos. Para esto proyectaremos las relaciones explicitadas en la red bipartita targets-anotaciones para introducir relaciones de similaridad entre targets. De esta manera los targets quedarán vinculados en la medida en que compartan características y anotaciones que consideramos relevantes, posiblemente relacionadas con la *drogabilidad* de los mismos.Como ya mencionamos, las características que tenemos en cuenta están relacionadas con conceptos tales como poseer un origen evolutivo común (similitud de secuencia) o compartir building blocks estructurales (i.e. dominios pfam).

Es importante destacar que en realidad podemos anticipar que no todas las categorías serán igual de relevantes en términos de actuar como proxy de

drogabilidad. Por ejemplo, la **figura 5.4** muestra que diferentes dominios pfam presentan diferentes fracciones de proteínas drogables anotadas.

En la **figura 5.4** se marcan con nombres aquellas anotaciones con mayor cantidad de proteínas drogables. Dentro de los dominios con más proteínas asociadas reportadas como *drogables* encontramos que aparecen por ejemplo: kinasas (*Pkinase,Pkinase_tyr*),canales ionico (*Ion_trans*), receptores g-acoplados (*7tpm_1*) y receptores nucleares (*Anf_receptor*).

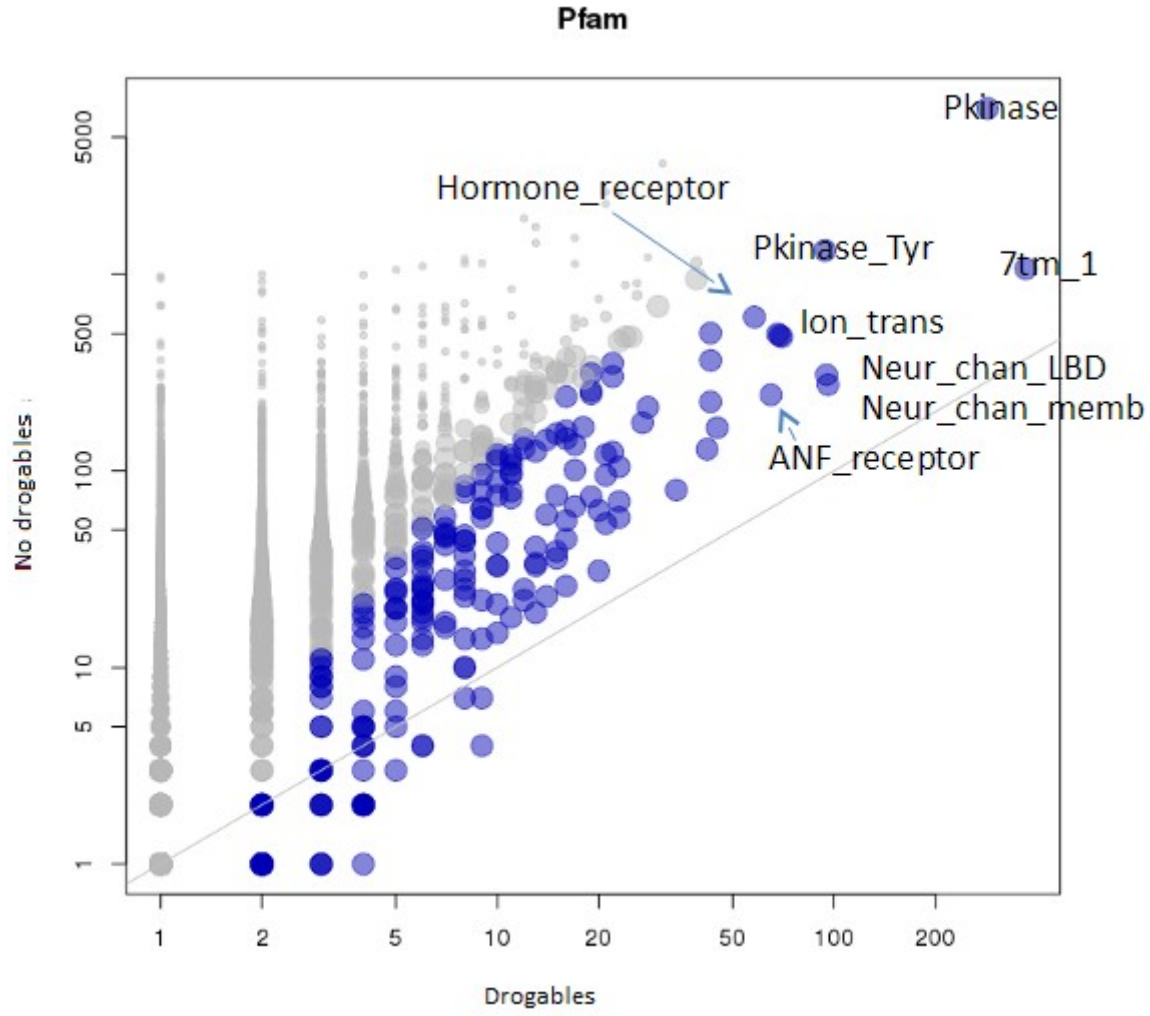


FIGURA 5.4: Cantidad de proteínas drogables y no drogables por PFAM.El tamaño de cada círculo corresponde a la fracción de proteínas drogables con respecto a la cantidad de no drogables.Se señalan los dominios PFAM con mayor cantidad de proteínas drogables.

Las kinasas son enzimas que participan en reacciones de fosforilación que típicamente se vincula con la activación de moléculas en vías de señalización

intracelular. Los receptores acoplados a proteína-G y nucleares, intervienen en la recepción de mensajes transmembrana. Los canales iónicos regulan el paso de iones y pequeñas moléculas a través de las membranas celulares.

Estos 4 tipos de proteínas participan o bien de cadenas de señalización intracelular o pueden inhibir el paso o alterar la sensibilidad a factores externos a la célula y por estas razones son el tipo de proteínas más buscadas a la hora de comenzar el desarrollo de una nueva droga ya que incrementan las tasas de éxito de los mismos [32]. Nuestro análisis pone de manifiesto que esta tendencia se refleja en los datos.

5.2.2.Drogabilidad y p-valores

Cómo se ha visto anteriormente no todas las anotaciones están asociadas a la misma cantidad de proteínas drogables. Es entonces posible pensar en asignar a cada anotación un índice que hable sobre la relevancia de saber que una dada proteína está afiliada a una dada categoría, en relación a la condición de drogabilidad.

Para ello utilizamos una medida estadística relacionada con el Test de Fisher, que permite cuantificar el grado de asociación entre ser drogable, o no serlo, y estar anotado, o no, a una dada categoría. La importancia de cada categoría en particular queda entonces dada por la tabla de contingencia de 2 x 2 que aglomera la información de cuántas proteínas VP están (o no) anotadas a él, y cuántas de ellas son (o no) blancos de droga.

La red contiene un total de N proteínas de las cuales B son blancos de alguna droga . 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.

La probabilidad de que por azar el nodo VB tenga exactamente ese número de blancos anotados sigue una distribución hipergeométrica que se calcula de forma exacta mediante la ecuación

$$P_{V_B(V'_p)} = \frac{\binom{\#V_B}{V'_p}\binom{N-\#V_B}{\#B-V'_p}}{\binom{N}{\#B}} \qquad \textbf{(5.1)}$$

La importancia de una anotación queda asociada entonces al p-valor que la tabla de contingencia tiene para dicha distribución. p-valores más pequeños indican que es

más improbable que la distribución de proteínas drogables de una anotación haya sido generada al azar.En la figura 5.4 en azul se muestran aquellas anotaciones que resultan significativas, es decir con el 20 percentil de p-valor más bajo.

En trabajos anteriores [33] se decidió definir a los fines del correcto funcionamiento del sistema de recomendación y para incorporar lo p-valores . La forma en la que se establece la importancia de cada anotación j en este trabajo está dada por el “Relevance score”.

$$R_{score}(j) = \begin{cases} 1 & si \quad p_j \leq q_{0.2} \\ [\frac{-log_{10}(p_j)}{máx_j(-log_{10}(p_j))}]^{\alpha} & si \quad p_j > q_{0.2} \end{cases} \qquad \textbf{(5.2)}$$

donde el p_j es el pvalor test de Fisher y $q_{0.2}$ denota el 20-percentil inferior de la distribución de p-valores.) Como se ve la estructura de la función es un $log(p)$, normalizado para que la función esté acotada entre 0 y 1. Donde el valor 1 de relevancia corresponde al 20% de las anotaciones con el menor p-valor.

Para cada tipo de anotación hay una distribución diferente de p-valores, observando la **tabla 5.1** se puede ver que la categoría con más anotaciones relevantes es la de dominios PFAM. Además es la que tiene el promedio de p-valor más pequeño y la anotación de menor p-valor con lo cual está será la categoría de mayor importancia a la hora de priorizar.

	Número de significativas	Fracción de significativas
Pfam	335	0,148
Ortología	347	0,124
Pathway	18	0,124

TABLA 5.1: Características de las anotaciones por capas.

Es importante distinguir que el definir la relevancia a través del p-valor es muy diferente a la caracterización más simple de fracción de proteínas drogables.En particular esa diferencia es notable en anotaciones con grados altos pero una fracción de proteínas drogables menor que 0,5. Si tomamos una anotación con 1000 proteínas

y 50 drogables el p-valor es de 0,0130, mientras que para una anotación más específica pero con mayor fracción de proteínas drogables, como es el caso de una anotación con 50 proteínas y 5 drogables se reporta un p-valor de 0.0330. Este efecto queda evidenciado en la **figura 5.4** donde se nota que a valores mayores del p-valor la frecuencia normalizada de pares de alta fracción de proteínas drogables por proteína anotada disminuye.

5.2.3.Entropía de anotaciones por especie

En la sección anterior introdujimos una manera de cuantificar que tan relevante puede considerarse a priori una categoría para el proceso de sugerir nuevas proteínas drogables. Esto se hizo en base a la evidencia ya existente, que en su mayoría corresponde a datos obtenidos en organismos modelo. Nuestro foco está puesto en identificar nuevos targets moleculares de especies asociadas a enfermedades desatendidas. Estas en general involucran proteomas poco estudiados, y resulta necesario saber en qué medida anotaciones compartidas nos sirven para “transferir” información de targets drogables entre especies. Para esto nos propusimos analizar la promiscuidad inter-especie de cada anotación, cuantificándola mediante la entropía definida según

$$S = -\sum_{i=1}^{k} p_i log(p_i), \qquad \textbf{(5.3)}$$

donde p_i es la proporción de proteínas de la especie i asociadas en la anotación de interés.

En general la entropía resulta nula si la anotación presenta proteínas de una sola especie y es máxima si las proteínas afiliadas a la misma se distribuye homogéneamente entre especies. En el caso de una anotación que presente asociaciones con proteínas de k especies en la misma proporción (con frecuencia $1/k$) resulta $S_{max} = -\sum_{i=1}^{k}(1/k)log(1/k)$ y entonces $S = -log(1/k) = log(k)$. Por tanto la entropía máxima varía según la cantidad de especies a las que un PFAM está afiliado. Esto nos permite definir una cantidad que sólo indica conque uniformidad se reparten las proteínas de una anotación (sin importar el número de especies, que denotamos entropía normalizada $S_{normalizada} = \frac{S}{S_{max}}$.

En las siguientes figuras se presentan gráficos de barras para la distribuciòn de entropía normalizada por anotación y la cantidad de especies que anota. También se incluye el gráfico de entropía en función de las especies alcanzadas. Se usan ejes horizontales logarítmicos, para hacer evidente la recta ajustada y cómo se compara esta con el caso de entropía máxima.

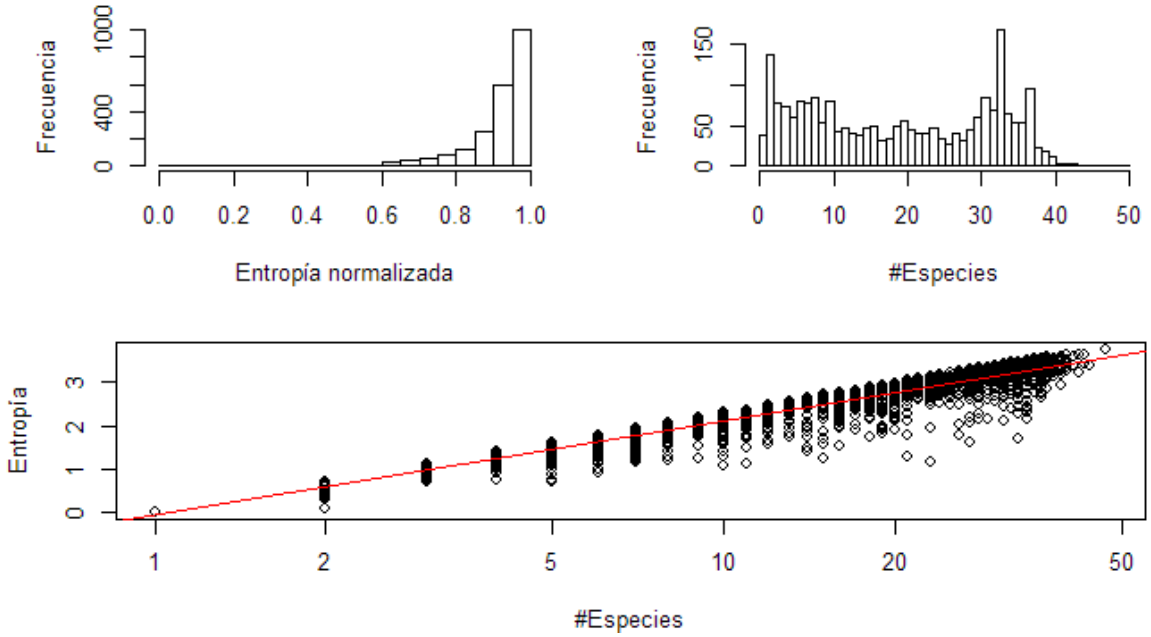


FIGURA 5.5:. Panel superior:Gráfico de barras de anotaciones PFAM con una entropía y un número de afiliaciones a especies dadas.Panel inferior entropía vs número de especies con afiliaciones. La recta obtenida tiene gradiente 0.94 $\pm$ 0,01.

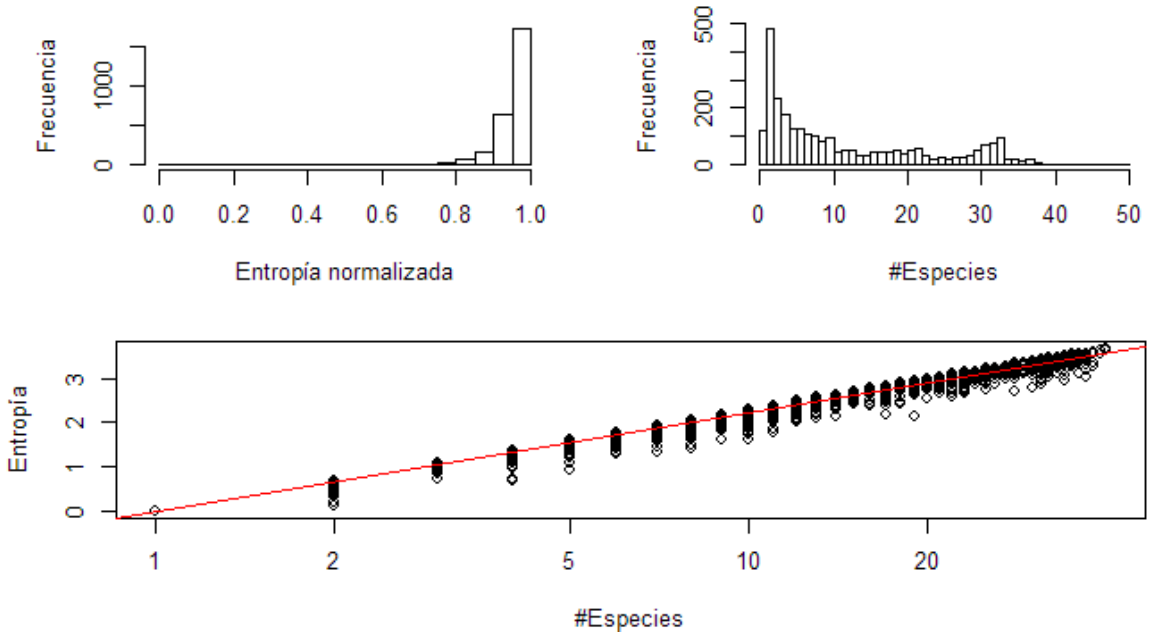


FIGURA 5.6: Panel superior: Gráfico de barras de grupos ortólogos con una entropía y un número de afiliaciones a especies dadas.Panel inferior entropía vs número de especies con afiliaciones. La pendiente del ajuste hecho es 0.970 ± 0.002 .

En el caso de anotaciones de clase PFAM (**figura 5.5**) y ortología(**figura 5.6**) se aprecia que las distribuciones de entropía normalizada son semejantes , más abultadas para valores altos en la segunda (la misma caracterización se puede hacer para la cantidad de especies). Al ajustar la entropía en función del logaritmo del número de especies por anotación, ambas categorías tienen pendientes cercanas a 1 , y por tanto se nota que la distribución de entropías es cercana al valor máximo independientemente del número de especies, por lo que cumplen un rol importante para la conectividad y procesos de difusión en la red de información entre especies. Para el caso de los pathways (**figura 5.7**) o caminos metabólicos la correlación entre números de especie y entropía es similar a de los otros tipos de anotaciones pero menor, cómo se puede observar por la dispersión de los puntos en el panel inferior de la figura 5.7 .

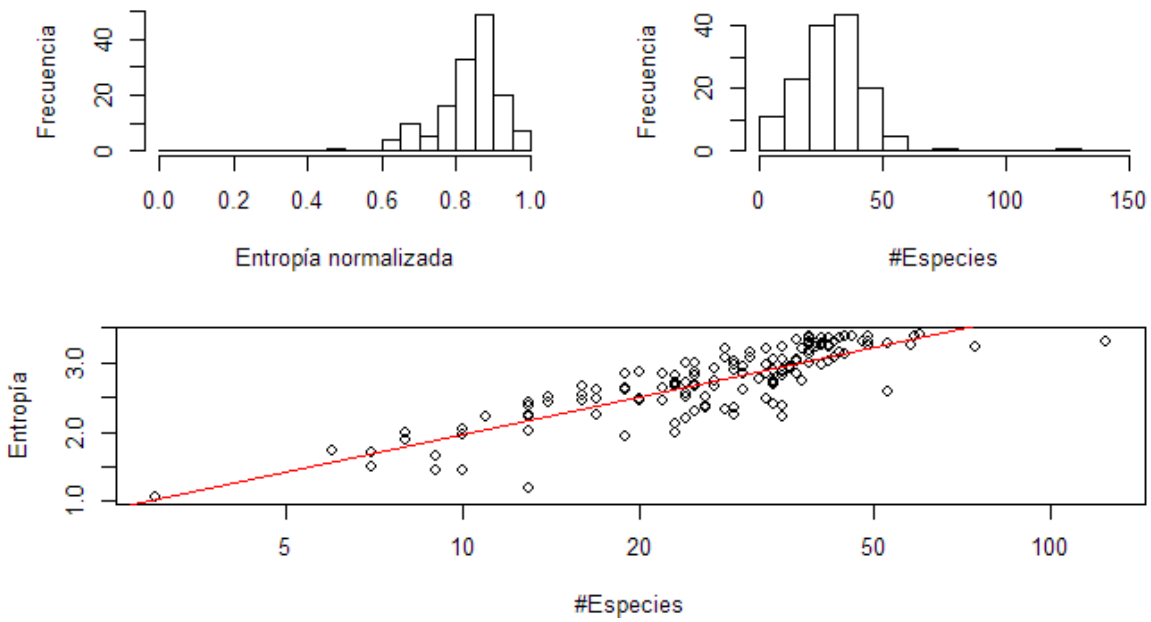


FIGURA 5.7: Panel superior:Gráfico de barras de Vías con una entropía y un número de afiliaciones a especies dadas.Panel inferior entropía vs número de especies con afiliaciones

5.2.4.P-valores y entropía

Una vez definidas nuestras dos cantidades de interés, una primera que caracteriza la sobrerrepresentación de proteínas drogables en cada anotación(p-valor de test fisher o Rscore) y otra asociada la capacidad de la anotación para transferir información entre especies (entropía S),surgen preguntas sobre las relaciones entre ellas ya que afectarán nuestros métodos de priorización. La figura 5.8 reporta las distribuciones obtenidas de Rscore para cortes dados de entropías inter-especies para el caso de anotaciones PFAM.

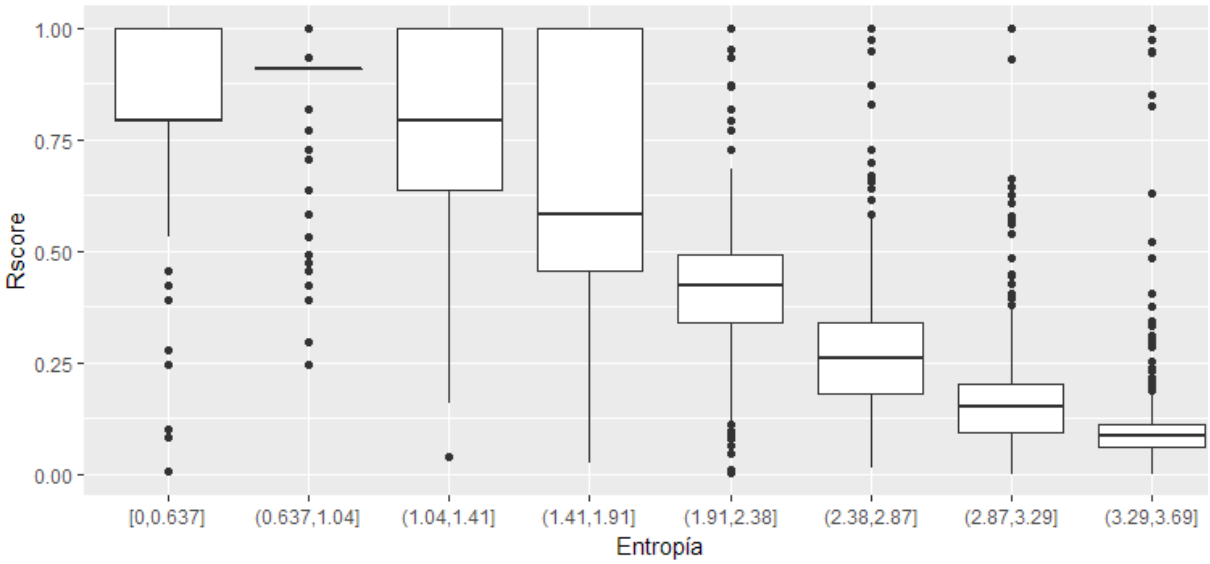


FIGURA 5.8: Gráfico de entropía versus Rscore para anotaciones de ortología (para PFAM se obtiene un gráfico similar). El Rscore mide la importancia de las anotaciones para definir la drogabilidad de una proteína, mientras que la entropía mide con qué facilidad esta transmite información entre especies. En promedio esto muestra que existe un trade-off entre la capacidad de difundir información y la relevancia biológica. En particular serán de interés los outliers por exceso [ara la entropía mayor.

Se observa que en general par anotaciones con mayor entropía (i.e. conectoras inter-especies) se obtienen scores de relevancia de menor valor. Sin embargo resulta interesante analizar cuáles son las anotaciones particulares que presentan alta relevancia y a la vez alta entropìa inter-especie. Estas anotaciones (tabla 5.5) serán aquellas que permitirán difundir por la red información del carácter de 'drogabilidad' de una especie a otra.

Los dominios de interés que aparecen en la tabla 5.5 están relacionados con enzimas que o sintetizan o sirven para reparar DNA. Por ejemplo *DHFR_1* se encuentra en todas las especies y tienen un rol central en la síntesis de precursores de ácidos nucleicos, mientras que *Thymidylate synthase* participa en los primeros estadios de la biosíntesis de ADN, y es seleccionado como blanco para controlar el crecimiento y desarrollo de varios tipos de cáncer, por quimioterapia [34]. Ribonucleotide reductase (RNR), es una enzima que cataliza la formación de desoxiribonucleotidos cuya función se conserva en todos los seres vivos, regula la tasa de síntesis de ADN, y ha sido usada como blanco para prevenir la replicación del herpes simplex virus (HSV) [35] . Como acabamos de mostrar, todos estos son procesos esenciales en numerosas

de especies. Además, por el hecho de ser esenciales son en general elegidos como blanco de droga.

PFAM	#Proteínas	Fracción drogables	#Especies (#NTD)	Rscore
PFAM186 ***DHFR_1***	61	0.262	47 (21)	1
PFAM303 *Thymidylat_synt*	53	0.188	40(22)	1
PFAM317 *Ribonuc_red_lgN*	43	0.139	36(21)	1
PFAM2852 *Pyr_redox_dim*	170	0.088	40(24)	1
PFAM2867 *Ribonuc_red_lgN*	45	0.133	37(22)	1

TABLA 5.2. Características de las anotaciones PFAM con máximo Rscore y entropía.NTD hace referencia a las especies que generan enfermedades tropicales desatendidas.

El primero de los grupos está asociado a dominios *Ribonuc_red_lgN*, *Ribonuc_red_lgC*, *ATP-cone* que aparecen generalmente en proteínas ribonucleotidas. El segundo está asociado a *tRNA-synt_2d* , *Phe_tRNA-synt_N* están vinculadas a catalizar el adjuntamiento de un aminoácido a su tARN (ARN transmisor), para la formación de nuevas proteínas en los ribosomas. El tercer grupo en la tabla corresponde al de mayor Rscore, y está asociado a los dominios PFAM *DHFR_1* ,*Thymidylat_synt* que son de interés por derecho propio (**tabla 5.2**). Estos dominios se conservan a lo largo de la mayoría de los seres vivos, y están relacionados a la síntesis de ácidos nucleicos. El cuarto grupo tiene como dominio principal al Prenyltrans , que está asociado a la creación de percursores para el colesterol, hormonas esteroides y vitamina D en vertebrados. Para el primer y el tercer grupo ortólogo, en la **tabla 5.3**, encontramos coincidencia con lo referido anteriormente para los dominios PFAM de la **tabla 5.2**.

	#Proteínas	Fracción drogables	#especies(#NTD)	Rscore

ORT10318	45	0.133	38(22)	0.82
ORT10500	40	0.100	40(24)	0.51
ORT10927	49	0.204	49(22)	1
ORT11497	40	0.150	36(22)	0.94

TABLA 5.3. Características de las anotaciones ortología con máximo Rscore y entropía.NTD hace referencia a las especies que generan enfermedades tropicales desatendidas.

5.3.Análisis de semejanzas de capas

Muchos espacios conviven en la red que nos hablan de similaridades químicas de drogas, de estructura y secuencia de proteínas ,y de bioactividades. En las secciones anteriores nos hemos referido por un lado a la capa de drogas y por otro lado a la capa de proteínas en conjunto con la de anotaciones. Aquí dilucidaremos qué tan coherentes es la información que la capa de drogas y anotaciones dan sobre la capa de proteínas. Nos valemos de que el esquema de interrelaciones entre drogas, targets y anotaciones permite inferir diferentes nociones de similaridad entre las entidades que componen nuestra red. Por ejemplo podemos establecer un criterio de similaridad entre proteínas en función de la similaridad de las drogas que las tienen como targets, o alternativamente, en función de las anotaciones PFAM que comparten o los grupos de ortología a los cuales pertenecen.

Para analizar la coherencia de similitudes entre proteínas inducidas por anotaciones pfam y de ortología, consideraremos la partición de proteínas en grupos a partir de su pertenencia a grupos de ortología y los comparamos con clusters detectados a través de distintos algoritmos, en la red de proteínas trivialmente proyectada por anotaciones PFAM y pathways alternativamente. Los valores obtenidos para las comparaciones se detallan en la tabla siguiente. Se utilizan las medidas para comparar particiones, de NMI y Rand según son definidas en la sección A.3.2.

	NMI	Rand
fast_greedy	0,80	0,97
infomap	0,83	0,98
louvain	0,81	0,97

TABLA 5.4. Valores obtenido para proyección PFAM con distintos métodos de clustering

	NMI	Rand
fast_greedy	0,36	0,71
infomap	0,27	0,44
louvain	0,37	0,74

TABLA 5.5.Valores obtenido para proyección Pathways con distintos métodos de clustering

En todos los casos encontramos una relación estrecha entre la clasificación de ortología y los dominios PFAM que resulta mucho mayor que entre la primera y las vías metábolicas. Esto ocurre sin importar el método de clustering o el índice de comparación, por lo que podemos decir que refleja una propiedad intrínseca de la red de relaciones entre anotaciones y proteínas. Así mismo esto nos permite establecer que los dominios PFAM y los grupos de ortología en general aportan información complementaria pero coherente ya que a pesar de representar características diferentes de las proteínas, no presentan informacion contradictoria. Si la división en comunidades difiriese mucho entre los diferentes dominios de conocimiento considerados, se tendría una situación en la cual típicamente la información de similitud/diferencia entre proteínas inducida por uno y otro dominio de conocimiento diferiría, lo que al integrar la información en una priorización resultaría poco provechoso.

Una vez descripta la similaridad inducida por conjuntos de proteínas entre los distintos tipos de anotaciones, es interesante evaluar en qué medida las semejanzas entre proteínas, inferidas por anotaciones, coinciden con las similaridades que se pueden inferir a partir de sus bioactividades.

En la **figura 5.9** mostramos la probabilidad de que un par de proteínas drogables compartan una droga como función de la cantidad de dominios PFAM compartidos. Se puede observar que a mayor cantidad de anotaciones PFAM compartidas, la probabilidad de tener una droga en común (línea celeste) o tener drogas que pertenecen a algún cluster común (línea roja) aumenta hasta 3 anotaciones y luego oscila entorno a 0.7. Como es de esperar la línea roja está relacionada a probabilidades más altas de que un par de proteína comparta una droga, y presenta

menos variación para más de 4 anotaciones.Dado que para 0 anotaciones en común la probabilidad de que se comparta alguna droga es de $1,05 \times 10^{-6}$, y en vista de lo anterior, queda establecido que la cantidad de dominios PFAM compartidos genera mayor similitud entre drogas que son activas en las proteínas.

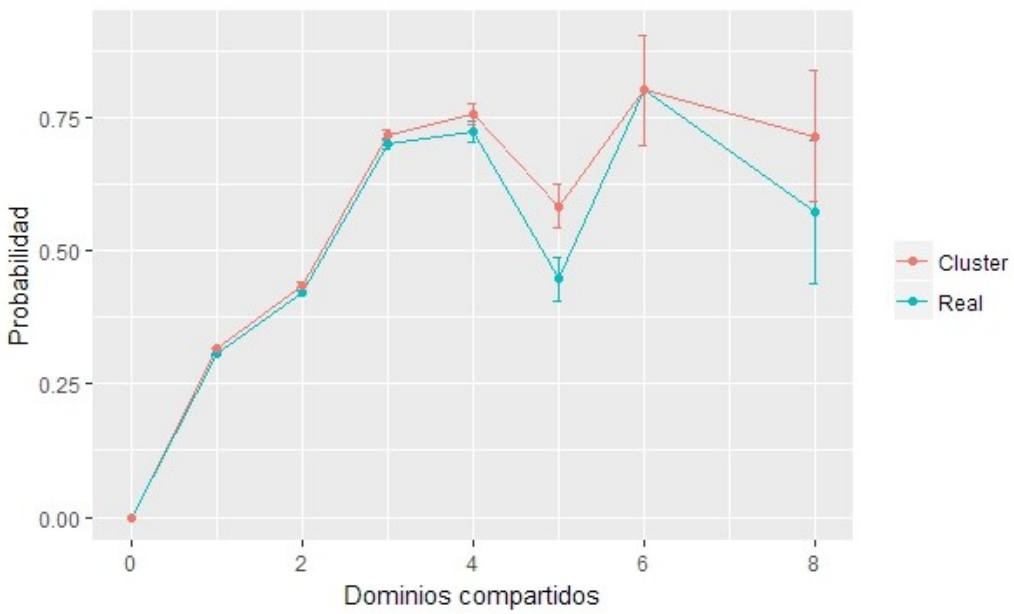


FIGURA 5.9: Cantidad de dominios PFAM compartidos por proteína drogables versus la probabilidad de que el par comparta una droga. En azul tomamos el hecho de compartir una droga de manera “fuerte”, es decir que haya una bioactividad desde la droga hasta ambas proteínas en el par. En rojo es la condición relajada de tener alguna droga en el mismo cluster S o T.

5.4.Conclusiones

En este capítulo detectamos categorías que presentan preferencia para anotar proteínas con bioactividades reportadas. Vimos que esto es un reflejo de sesgos que efectivamente existen en el área de desarrollo y búsqueda de nuevos fármacos

Muchas categorías tienen un rol muy importante en la interconectividad entre proteínas de diferentes especies, y esto lo cuantificamos mediante la entropía de la distribución de proteínas por especie. Esta medida resulta muy relevante para entender cómo puede ser transferido el conocimiento entre especies.

Además analizamos la coherencia entre los distintos tipos de anotaciones a través de similaridades a primeros vecinos en la red. Establecimos que la información que la capa de anotaciones ortólogas impone sobre la capa de proteínas es semejante a la de PFAM. Más aún establecimos la existencia de un vínculo entre drogas comunes, entre

proteínas y la cantidad de anotaciones PFAM compartidas. Se puede observar una clara relación creciente entre ambas magnitudes y que aproximadamente el 60% de proteínas que comparten 3 o más dominios en su arquitectura son alcanzadas por al menos una droga en común.

Capítulo 6

Red proyectada en anotaciones

En las secciones anteriores se estudiaron y caracterizaron la capa de drogas, de blancos y de anotaciones. En esta sección se avanzará en el análisis de la interconectividad entre las mismas. En particular, nos interesará dilucidar si vinculos inducidos por targets drogables reflejan estructura no trivial en el espacio de anotaciones biológicas.

Para esto haremos uso del concepto de proyección de una red bipartita en una red mono-partita introducido en la sección 2.4. Mediante este procedimiento estableceremos asociaciones entre anotaciones (PFAM o de ortología) inducidas por la existencia de proteínas en común que posean pares de anotaciones. Así, el hecho de que dos anotaciones estén fuertemente conectadas en la red proyectada implicará que las anotaciones tienen un alto nivel de solapamiento en cuanto a las proteínas que cada una tiene asociada. Màs aún, si para esta estimación sólo consderáramos proteínas drogables, enlaces fuertes entre categorías PFAM o de ortólogos, podrìan sugerir que las mismas están asociadas a funciones biológicas o a estructuras que aparecen en moléculas que han sido repetidamente usadas como blancos. En general las redes que resulten de ambos tipos de proyecciones (i.e. considerando todas las proteínas o sólo las drogables) tendrán propiedades estructurales diferentes, como ilustran los paneles de la **figura 6.1**, donde se consignan las componentes gigantes de la red PFAM proyectados por proteínas y por proteínas alcanzadas por drogas respectivamente. Haremos una caracterización somera de dichas diferencias en las siguientes secciones de éste capítulo.

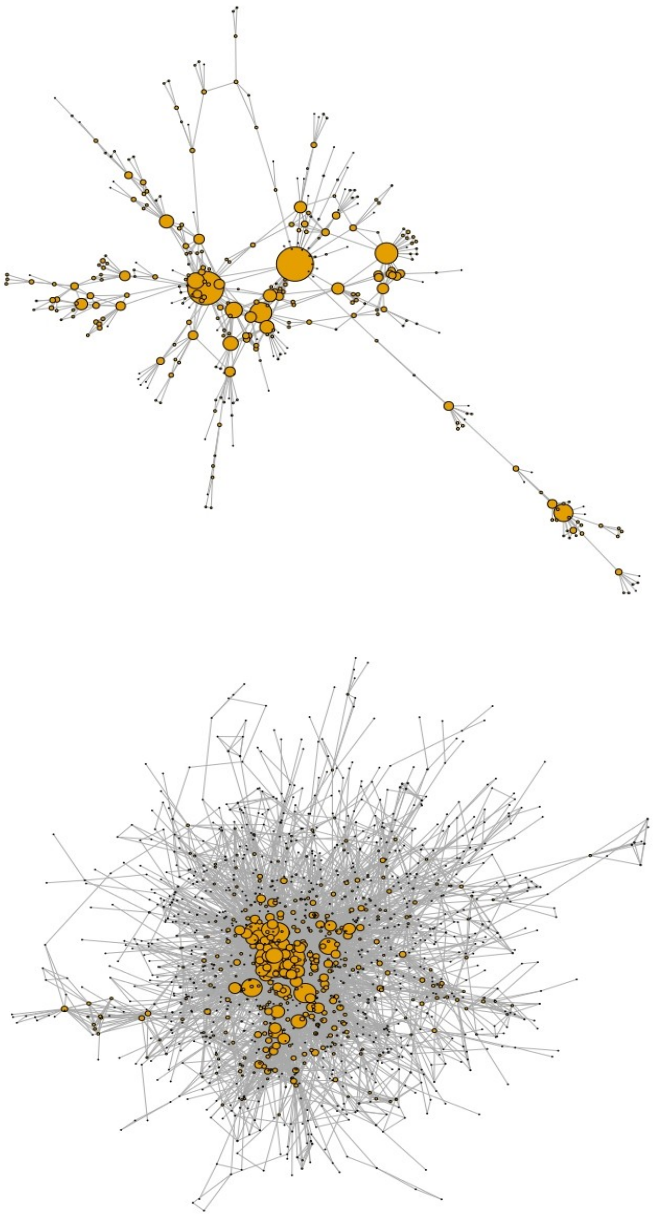


FIGURA 6.1:.En el panel superior se encuentra la componente gigante (CG) proyección del grafo de afiliación targets drogables a PFAMs, el tamaño de los nodos está asociado a la cantidad de proteínas drogables anotadas. Es visible la estructura de abanico relacionada con el coeficiente de asortatividad negativo. En el panel inferior la misma proyección pero teniendo en cuenta todas las proteínas, el tamaño de los nodos está asociado a la cantidad de proteínas anotadas.

6.1. Métodos de proyección

Es posible considerar diferentes metodologías para examinar estructura relevante en la red de anotaciones, relacionados con diferentes alternativas de proyección con las cuales obtener una red mono-partita de anotaciones (ver seccion 2.4). Nosotros consideraremos las tres metodologías de proyección introducidos en el capítulo 2: proyección trivial, proyección de difusión de probabilidad, Probs [24] y proyección por validación estadística, StatVal [25].

Bajo la proyección trivial, el peso del enlace que unirá a dos anotaciones es igual al número de proteínas en común que las mismas posean. Como vimos antes, eq (2.8), se calcula según

$$w_{i,j} = \sum_{l=1}^{m} a_{il} a_{jl} \qquad \textbf{(6.1)}$$

Según la metodología ProbS, la matriz de adyacencia de la red de anotaciones estará dada por la expresión de la eq (2.11).

$$w_{i,j} = \frac{1}{k_{x_j}} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k_{y_l}} \qquad \textbf{(6.2)}$$

Ahora, la contribución de una proteína en común a un par dado de anotaciones se pondera (i.e. penaliza) por la conectividad de la misma,k_i. Al mismo tiempo la matriz de adyacencia deja de ser simètrica, ya que el peso del enlace dependerá de la conectividad de la anotación de 'llegada' k_j. Es importante remarcar que estas dos proyecciones presentan las mismas características de conexiones subyacentes en el sentido de que lo que varía es el peso de los enlaces y la direccionalidad. Por cada enlace en la red naive hay dos enlaces con distinta dirección y distinto peso en Probs.

Finalmente en la proyección de validación estadística (StatVal), los enlaces quedan determinados por la existencia de proteínas comunes, de un dado grado, en una proporción significativamente superior a la esperada por azar. Esto significa, por ejemplo cuando se toman anotaciones PFAMs, que los enlaces entre dos anotaciones de ese tipo se definen analizando sistemáticamente, para cada grupo de proteínas que presenten un número dado de dominios PFAM, si existe un número inesperadamente alto (en el sentido estadístico) de proteínas comunes. Para ganar intuición sobre este procedimiento, la **figura 6.2** muestra la cantidad de enlaces validados entre anotaciones considerando proyecciones armadas con proteínas de un dado grado. Es posible observar que la mayor parte de las asociaciones provienen de la validaciòn de

enlaces inducidos por proteínas de grado menor que 4. La diferencia observada entre los valores absolutos del número de enlaces validados entre los obtenidos considerando todas las proteínas o solo las drogables surge del hecho que las proteínas drogables son sólo el 6% del total de proteínas , lo que resulta en una penalización mayor al hacer la selección de aristas con la distribución hipergeométrica.

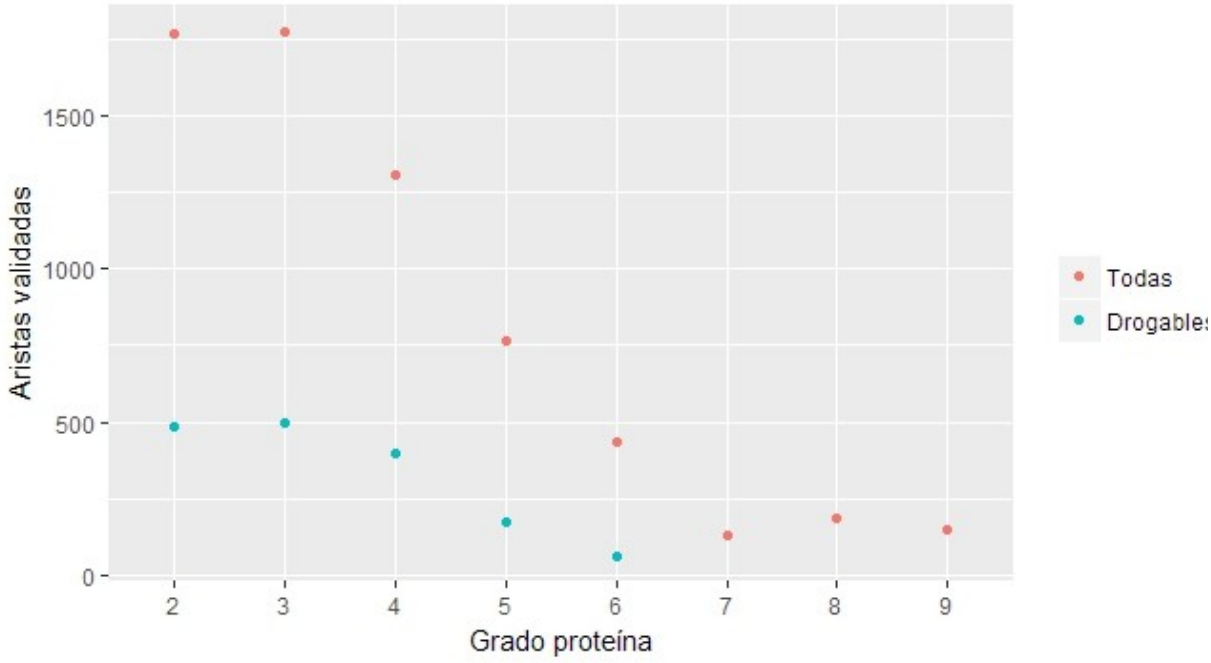


FIGURA 6.2:.Gráficos para la cantidad de enlaces validados por grado de anotación usando un umbral de p-valor 0.01 . Como la proyección de validación estadística depende del subgrafo inducido por todas las proteínas de un cierto grado, se ve que para bajos grados ambas proyecciones difieren en más cantidad de enlaces que para altos grados.

6.2.Redes obtenidas

En la **TABLA 6.1.** se muestran datos de interés de la estructura obtenida para cada una de las proyecciones realizadas. Es interesante notar que por las características con las que se obtienen los enlaces validados en el método de validación estadística se reduce notablemente la cantidad de nodos en la red proyectada, ya que las anotaciones para las cuales no se encontró ningún enlace relevante por este método son extraídos de la red. En todos los casos ocurre que al considerar todas las proteínas para proyectar se agranda la componente principal obtenida al proyectar únicamente sobre proteínas drogables. En otras palabras el efecto de tener en cuenta proteínas no drogables solo crea enlaces intra-componente

(i.e. dentro de estructuras ya conectadas) o entre componentes y la componente gigante CG, ver A.1.2 .

Proyección	proyectado con proteínas	#componentes	componente gigante	#nodos	#enlaces / stregth total
Naive / Probs	todos	729	1340	2252	4653/606.28
	drogables	1272	335	2252	1732/704.87
StatVal	todos	208	868	1422	2296
	drogables	141	12	287	366

TABLA 6.1. Características de los grafos obtenidos a través de proyecciones.

La **tabla 6.1** refleja el tamaño y escala de las estructuras obtenidas con los diferentes métodos de proyección. Vemos que en validación estadística se presenta como una metodología más restictiva, en el sentido que la cantidad de enlaces es sustancialmente menor. En particular para la proyección de solo drogables esta involucra 1/10 de los nodos y 1/7 de los enlaces de los obtenidos en las otras proyecciones.

Además. al tomar sólo proteínas drogables la cantidad de enlaces decrece en Naive y ProbS en un 63% mientras que en StatVal lo hace en un 85%. Debido a la pérdida de enlaces, se puede observar el decrecimiento del tamaño de la componente gigante para todos los métodos de proyección (en particular, para la proyección StatVal, deja nodos sin enlaces, que eliminamos de la red). En general la red proyectada con drogables resulta en una distribución mucho mas fragmentada, con una componente gigante que compromete sólo un 15% de la masa de la red (5% en el caso de StatVal), en comparación del 60% obtenido en proyecciones realizadas considerando todos los targets.

6.3 Medidas topológicas de interés

6.3.1.Transitividad

La transitividad o coeficiente de clustering, como se puede leer en la sección A.2.3, hace referencia a la probabilidad con que dos nodos que poseen un vecino común son a su vez vecinos entre ellos. En nuestro caso, por ejemplo, responde a la

pregunta: si dos PFAM presentan, cada uno, alguna proteína común con un tercer dominio PFAM, qué probabilidad hay de que anoten juntas a una misma proteína?. Evidentemente este número puede no ser el mismo dependiendo de si estemos considerando exclusivamente proteínas drogables o si analizamos el conjunto completo de targets disponibles. Cualquiera sea el caso es posible testear la significancia estadìstica de los valores de transitividad encontrados estableciendo una comparación con lo obtenido considerando una red al azar, en la cual se mantiene el grado de cada nodo de la red proyectada, pero se selecciona al azar qué nodos están vinculados entre sí por una arista.

Presentamos en la **tabla 6.2** los valores de transitividad global de la red de proteínas proyectada considerando la proyección trivial (resultados similares se obtienen para las otras metodologìas de proyección). La primer y segunda columna de la tabla muestra el valor obtenido sin considerar, o considerando, el peso de los enlaces respectivamente.

	Global(azar)	Barrat(azar)
drogables	$0,45(0,06 \pm 0.02)$	$0,819(0,200 \pm 0,032)$
proteínas	$0,20(0,0 \pm 0,01)$	$0,619(0,0910 \pm 0,033)$
drogables CG	$0,28(0,14 \pm 0,01)$	$0,753(0,068 \pm 0,034)$
proteínas CG	$0,20(0,18 \pm 0,01)$	$0,586(0,037 \pm 0,021)$

Tabla 6.2.Coeficiente de clustering de los grafos obtenidos a través de proyección trivial y su comparación con la red al azar.

Como primer comentario es posible notar a partir de la **tabla 6.2** que en general las componentes gigantes exhiben coeficientes de clustering menores que los de la red completa. Esto se explica debido a que la red completa incluye, además de la componente gigante, numerosas estructuras tipo clique muy pequeñas (de menos de 4 elementos). Estas componentes pequeñas, densamente conectadas internamente, aumentan el valor medio de transitividad reportado en el caso de la red completa.

Así mismo vemos que considerando proteínas drogables se obtienen proyecciones con coeficientes mayores que los obtenidos en redes control (i.e. recableadas al azar con idéntica distribución de grado) y mayor transitividad que cuando se usan todas las proteínas (con todas las proteín la transitividad global resulta 0.206 mientras que en la originada a través de targets drogables se reporta 0.459). Esto sugiere que los PFAMs relacionados con targets drogables tienden a formar conjuntos más interconectados localmente. Es decir, encontramos que en las redes de solo proteínas drogables los enlaces inducidos por co-anotación de proteínas drogables generan estructuras de entornos locales más interconectados para anotaciones pfam. Esto puede reflejar un efecto cooperativo de los dominios en el sentido que proteínas que incluyan a tales dominios de manera conjunta tengan más chances de ser drogables.

6.3.2.Centralidad

En estudios de redes el concepto de centralidad està asociado a una medida de la importancia de un nodo con respecto a sus vecinos y no está definido de manera unívoca sino que depende del algoritmo aplicado. Las medidas de centralidad de un nodo más simples de formular son el grado y el strength (ver apéndice). En nuestro caso, los dominios PFAM de mayor strength en todas las redes resultaron ser PFAM69 y PFAM1. PFAM1 representa receptores GPCR tipo *Rhipdosina*. PFAM 69 representa el dominio kinasa. Estos dominios se relacionan con características estructurales usualmente utilizadas para guiar la búsqueda y diseño de nuevos fármacos [32]. Más aún, para las proyecciones de sólo drogables, encontramos que el strength correlaciona positivamente con el Rscore (ver **figura 6.3**). Por ejemplo, considerando la relación strength vs Rscore obtuvimos valores de correlación 0.29 y 0.28 para proyecciones obtenidas con proteínas drogables utilizando la metodología trivial y la de validación estadística respectivamente Ambas resultan significativas al compararlas con las obtenidas a partir de redes recableadas al azar: $(0,07 \pm 0.03)$ y (0.0 ± 0.1) respectivamente.

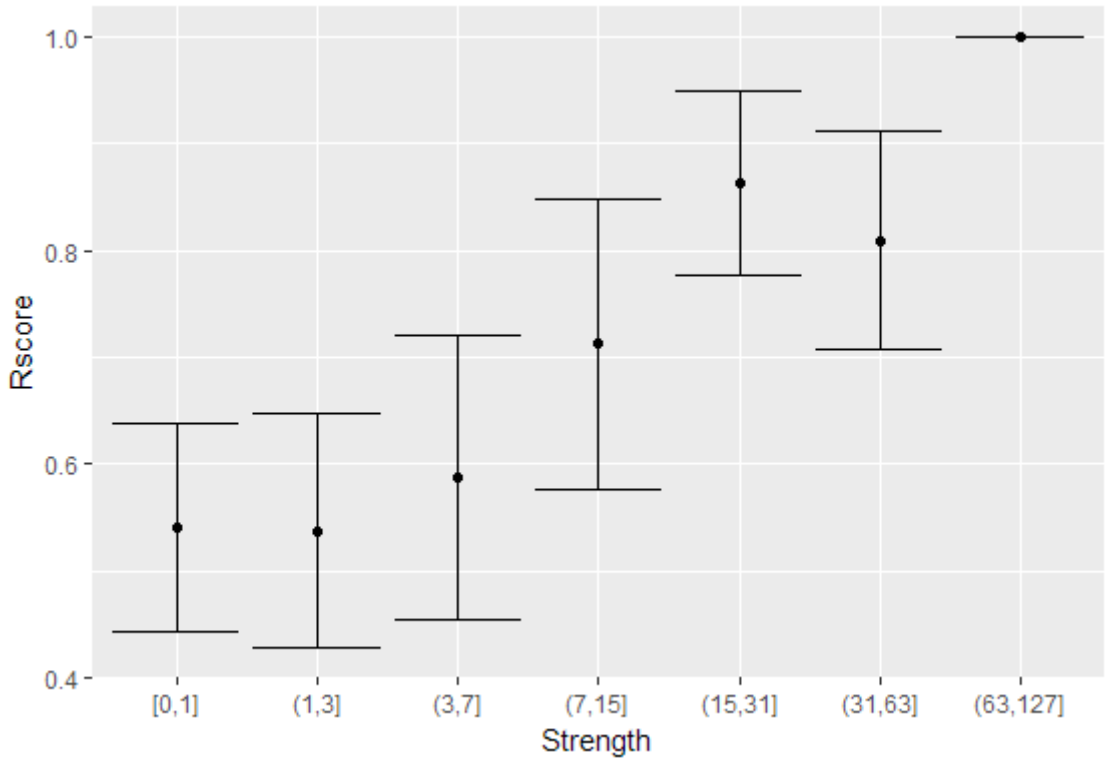


FIGURA 6.3: .Relación entre Strength y Rscore promedio para conjuntos de vértices tomados por intervalos logarítmicos en Strength.Estos corresponden a la proyección trivial de drogables.

Adicionalmente, diversas nociones de centralidad nos pueden proveer información distinta y complementaria acerca de la red. Las medidas que se analizaron alternativamente son las centralidades de Kleinberg (hub y autoridad), y eigen centralidad (definidos en A.2.4 del apèndice) en el contexto de redes de anotaciones PFAM.

En las **tablas 6.5** y **6.6** reportamos dominios PFAM de interés, identificados como aquellos que presentaban altos valores de centralidad y Rscore~ 1. En la primer tabla se incluyen resultados obtenidos para proyecciones ProbS, mientras que en la segunda para proyecciones trivial y estadística. Para estos últimos se encontró que los valores máximos obtenidos correspondían a los mismos nodos, independientemente de la medida de centralidad elegida.

	Hub	Autoridad	Eigenvector
Todos	PFAM69 * PFAM433* PFAM169 *	PFAM659* PFAM12474* PFAM11629 *	PFAM69* PFAM433* PFAM659 *

Bioactivos	PFAM69 * PFAM433 * PFAM130 *	PFAM659 * PFAM12474 * PFAM11629 *	PFAM69* PFAM433* PFAM130 *

TABLA 6.5 Valores de centralidad obtenidos con la proyección Probs. Con * se marcan los nodos que son de mayor interés ya que presentan valores de centralidad y de Rscore cercanos a 1.PFAM 69 kinasa PFAM659 *Polo kinasa***,** PFAM12474 *TGF beta receptor 2* ,PFAM11629 *C terminal SARAH*.Y PFAM 69 PFAM 433 kinasas PFAM 2931 neurotransmisor de canal ionPFAM 2932 ligand binding domain.

Tipo de proyección	Todas las centralidades
Trivial Bioactivos CG	PFAM69* PFAM433* PFAM130*
Trivial Bioactivos	PFAM2931* PFAM2932* PFAM413*
Trivial todas / CG	PFAM271 PFAM270 PFAM176
StatVal Bioactivas/CG	PFAM520* PFAM2815 PFAM1365
StatVal todas CG	PFAM1* PFAM10* PFAM1017
StatVal todas	PFAM1* PFAM10* PFAM1012

TABLA 6.6: Tabla para los valores de centralidad obtenidos con la proyección StatVal y trivial. Con * se marcan los nodos que son de mayor interés ya que presentan valores de centralidad y de Rscore cercanos a 1. PFAM 69 PFAM 433 kinasas, PFAM 2931 neurotransmisor de canal ion PFAM 2932 ligand binding domain, PFAM413 C1 domain, PFAM 413 *Peptidase_M10,* PFAM 1 *7 transmembrane receptor* , PFAM 10 *Helix-loop-helix DNA-binding* domain y PFAM520 *Ion_trans.*

En general se puede apreciar que para la proyección trivial con todas sus proteínas no aparecen casos claros de anotaciones que presenten simultáneamente alta centralidad y Rscore. Sin embargo, para otros casos las anotaciones de centralidad mayor tienen un alto Rscore y se observa que algunos de estos dominios pertenecen los cuatro grandes tipos, vistos en la sección 5.2.1, que suelen buscarse como blancos. A saber están las kinasas: PFAM 69 kinasa PFAM659 *Polo kinasa***,***PFAM 433 Protein kinase C***. Los** receptores acoplados a proteinas g PFAM12474 *TGF beta receptor 2,* y los canales-ionicos *canal ion PFAM2931 Neur_chan_LBD, PFAM* ***520*** *Ion_trans.* De esta manera vemos cómo la estructura de interconexiones en nuestros datos refleja el sesgo de la industria en la búsqueda de fármacos hacia proteínas de señalización (kinasas), canales, proteínas receptores, etc.

Finalmente, resulta interesante mencionar que las anotaciones de máxima centralidad reportadas en las **tablas 6.5** y **6.6** son distintas a las reportadas en el capítulo 5 y que presentaban máxima entropía y alto Rscore. Esto se debe a que la centralidad da una noción de cómo se relacionan unas anotaciones con otras a través de proteínas compartidas, mientras que la entropía mide la capacidad de las anotaciones para transferir información entre especies, en la medida que valores altos de la entropía introducida en el capítulo anterior, permite identificar anotaciones asociadas a una característica estructural común a proteinas de diferentes especies.

6.3.3. Asortatividad

Para entender la estructura de la red a escala global y detectar relaciones no triviales entre anotaciones, es a veces de utilidad poder medir si nodos parecidos tienden a enlazarse con nodos parecidos. Esto se hace calculando la asortatividad de la red (ver apéndice). En nuestro caso nos interesa saber si anotaciones con muchas proteínas drogables tienden a vincularse en la red proyectada con otras anotaciones que posean una cantidad parecida de proteínas drogables (o valores similares de Rscore). Para esto se midió la asortatividad de Rscore y la de grado propio en las diferentes proyecciones. El grado propio hace referencia, en el caso de proyecciones sobre proteínas bioactivas, a la cantidad de proteínas bioactivas que anota el dominio PFAM, y en el caso de proyecciones sobre todas las proteínas, a todas las proteínas que anota dicho dominio.

	Grado propio (Azar)	Rscore (Azar)
PFAM unidos por drogables CG	$-0,111(-0,094 \pm 0,0211)$	$0,350(0,259 \pm 0,025)$
PFAM unidos por proteínas CG	$-0,064(-0,059 \pm 0,030)$	$0,302(0,288 \pm 0,045)$

TABLA 6.7: Asortatividades obtenidas para proyecciones probs/trivial. Las asortatividades son indistintas por lo observado antes de que entre una y otra proyección solo cambia el peso de las aristas y la dirección.

Se nota en general de la **Tabla 6.7** (y se constata para el resto de las proyecciones) que las proyecciones que tienen en cuenta sólo el conjunto de proteínas drogables son las que presentan mayor disasortatividad de grado. Se ve que esto está

reflejado en la **figura 6.1** , donde se discierne la estructura tipo abanico característica de redes con esta propiedad. Esto implica que dominios con muchas anotaciones (i.e. de gran tamaño en la figura) tienden a co-aparecer en proteínas drogables junto a otros dominios menos comunes. Sin embargo no podemos decir que la asortatividad sea significativamente diferente a la obtenida en una red recableada aleatoriamente, que mantenga la distribución de grado original. Por esta razón no podemos descartar que los valores obtenidos surjan como consecuencia de dicha distribución y no provengan de correlaciones no triviales de dos cuerpos.

Por otro lado, para la proyección teniendo en cuenta proteínas drogables según se observa en la **tabla 6.7** (y se constata para el resto de las proyecciones) la asortatividad del Rscore es apreciablemente grande, comparando con el re-cableado al azar manteniendo la distribución de grado y con lo obtenido en la red de proyectada con todas las proteínas. Esto implica al igual que, lo que ocurre con la relación Rscore-transitividad, que las proteínas drogables tienden a nuclearse alrededor de PFAMs de alto relevance score, y además que dominios de bajo Rscore tienden a compartir proteínas drogables con otros de bajo Rscore.

6.4.Priorización de drogas desde dominios fuertemente conexos

Como hemos visto en la sección anterior, la conectividad entre anotaciones inducida por targets drogables permite poner de manifiesto caracterìsticas estructurales interesantes que presentan proteínas que son blanco de drogas. En esta sección nos interesará investigar si existe alguna relación entre alguna de las estructuras detectadas y determinadas drogas específicas. Para llevar esto adelante consideramos la red proyectada, considerando sólo proteínas drogables, por validación estadística restricta (ver sección 2.4.2).

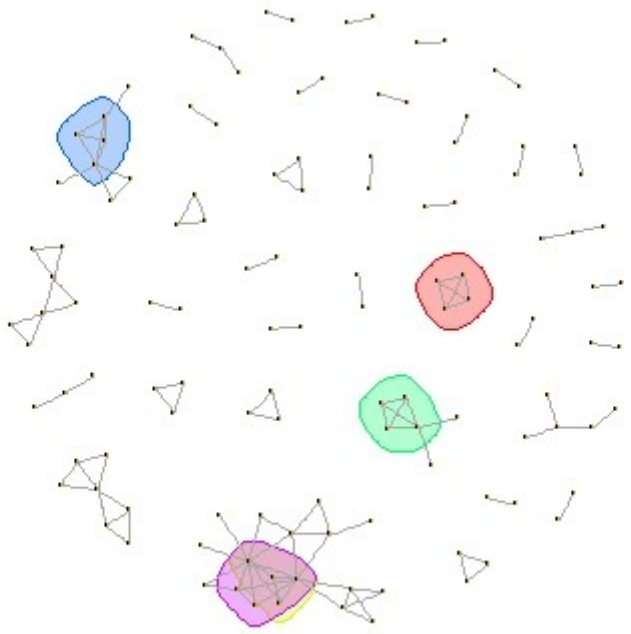


FIGURA 6.4: Grafo proyectado con validación estadística teniendo en cuenta sólo proteínas drogables y con la corrección de cálculo sólo sobre componentes.

La figura 6.4 muestra un esquema de la red de anotaciones PFAM obtenida, donde se identificaron 4-cliques (es decir, componentes completamente conexas de 4 elementos). Estas estructuras representan dominios PFAM fuertemente conectados gracias a la existencia de proteínas drogables en común, y lo que nos preguntamos era si era posible identificar la existencia de algún grupo específico de drogas responsable de las bioactividades involucradas.

Para investigar eso, a cada uno de estos 4-cliques de anotaciones PFAM se los consideró, por vez, como semillas con puntuación unitaria para un proceso de priorización utilizando una metodología de votación hacia la capa de proteínas. Luego se propagó la puntuación de cada proteína haciendo VS de nuevo hacia la capa de drogas. De esta manera asignamos un puntaje a cada droga según

$$W = PAV \tag{6.3}$$

donde A es, como antes, la matriz de adyacencia proteína-anotación, P es la matriz de adyacencia droga-proteína, y V es el vector de las semillas, es decir tiene 1s en los PFAM del 4-clique que se analiaza y 0's en el resto.

En la **tabla 6.9** consignamos un resumen de los resultados de este procedimiento. En la segunda columna se consignan los PFAM pertenecientes al clique desde el que se prioriza. En la tercer columna se muestra el conjunto de drogas que obtuvieron lo 5 valores de priorización más alto. En la cuarta se incluye una breve descripción del espectro común de acción de estas. En la quintacolumna se menciona a qué enfermedades y que acciones tenían las 5 drogas mas puntuadas (que además tenían como blanco a una proteína con los 4 dominios). En las columnas 5 y 6 se encuentran los resultados del ejercicio de validación de las drogas que se obtienen, que serán explicados a continuación.

Para validar la significancia del ranking que permite identificar a drogas relevantes se consideraron dos alternativas. En primera instancia se consideró generar rankings mediante un modelo nulo que incluía priorizar drogas según ecuación 6.3 a partir de 4 anotaciones PFAM elegidas al azar. Repitiendo este procedimiento 1000 veces se confeccionaron rankings de drogas, consignando el porcentaje de veces que aparecían las mismas 5 drogas top-rankeadas. El otro método que usamos se basa en cuantificar el “poder de rankeo" de los scores producidos, es decir la inhomogeneidad con que se distribuye puntuación entre las drogas a partir de un dado conjunto de semillas. Para que el ranking y por tanto la selección de los elementos más puntuados sea buena, debería haber una diferencia clara entre los niveles de score más alto y el resto de los elementos, es decir que se diferencie del modelo nulo. El modelo nulo propuesto para analizar validez estadística supone que la puntuación x normalizada generada por el clique PFAM se asignan para cada elemento (en este caso droga) rankeado con una distribución uniforme. Esto da un valor de probabilidad $\alpha = (1-x)^{k-1}$ de que exista una puntuación mayor a x cuando el score se distribuye uniformemente entre *k* elementos rankeados [36]. En nuestro caso *k* es la cardinalidad del conjunto de drogas que reciben un score no nulo a partir de un conjunto de semillas dado. Valores bajos de α serán indicio de scores inusualmente altos respecto a este control. Como medida de confianza en el ranking que las produjo, en la tabla 6.9 se muestra el máximo valor α de entre las drogas mejor rankeadas seleccionadas.

Los resultados se muestran en la **tabla 6.9**. El 4-clique 3 y el 4-clique 4 resultaron esencialmente indistinguibles al tener en cuenta las 5 drogas más puntuadas.Para cada caso las drogas seleccionadas muestran espectros de acción comunes, aunque para ninguno de los casos las más puntuadas pertenecían al mismo cluster Tanimoto o subestructura. La información así derivada no está trivialmente contenida en la información de similitud química y en general se nota que las drogas así halladas

tienen mecanismos de acción comunes, según fueron relevados en las bases de datos PubChem y Chembl.

	Anotaciones	Drogas	Acción conjunta de las drogas	Probabilidad con semillas al azar	Validación alfa(máxima)
1	PFAM69 (***Pkinase***) PFAM433(***Pkinase_C***) PFAM168(***C2***) PFAM130(***C1_1***)	D 34576* D 453394 D 453399* D 32183* D 40110	Inhibidores de kinasa, relacionado con diabetes y candidatos para tratamientos de cancer	2,9%	$2,707042 \times 10^{-09}$
2	PFAM520(***Ion_trans***) PFAM11933 (***Na_trans_cytopl***) PFAM612 (***IQ***) PFAM6512(***Na_trans_assoc***)	D 34576 * D 453391 D 453394 D 370532 D 363932 D 511441	Drogas relacionadas con vasodilatacion y epilepsia.Principalmente bloqueadores de calcio.	0,6%	$5,561837 \times 10^{-13}$
3	PFAM41 (***fn3***) PFAM7714 (***Pkinase_Tyr***) PFAM1030 ***Recep_L_domain*** PFAM757 (***Furin-like***)	D 3571 D 30211 D 116479 D 154477 D 36752 D 62971 D 35548 D 38936	Drogas asociadas a distintos tipos de cáncer	2,4%	$9,796857 \times 10^{-03}$
4	PFAM41 (***fn3***) PFAM7714 (***Pkinase_Tyr***) PFAM536 (***SAM_1***) PFAM1404(***Ephrin_lbd***)	D 34576* D 453391 D 453394 D 453399 D 511441	Drogas asociadas a distintos tipos de cáncer	1,3%	$1,562485 \times 10^{-06}$
5	PFAM13855 (***LRR_8***) PFAM1 (***7tm_1***) PFAM1462 (***LRRNT***) PFAM12369(***GnHR_trans***)	D 2174 * D 757* D 7724* D 4112 D 3571 D 33147	Drogas psicotrópicas	2,6%	$1,153909 \times 10^{-07}$

TABLA 6.9. Cliques usado como semillas, con el tipo de drogas obtenidas al hacer VS en la capa de drogas. Se notan los usos comunes de las drogas obtenidas y se caracteriza la calidad de la validación usando una comparación con priorización

desde 4 anotaciones al azar y el método visto en [36]. Números más pequeños de las características de la validación, implica que las drogas obtenidas no pueden haber sido obtenidas de semillas al azar. D43576 staurosporine, D 453399 Vandetanib, D 32183 canertinib ,D 2174 Closapine ,D757 Mianserin, D 7724 Risperidone.

6.5. Predicción de drogas a partir de arquitectura de dominios

De los 40 enlaces de la red de anotaciones inferida a partir de targets drogables (**figura 6.2**), sólo dos no aparecen cuando se consideran todas las proteínas en la proyección por validación estadística. Quisimos investigar entonces si los mismos codificaban algún tipo de información relacionada con la drogabilidad de las proteínas que le dieron lugar.

Para ello, para cada uno de los dos enlaces, identificamos, mediante el procedimiento de propagación antes descripto, drogas que podrían estar asociadas a con los dominios PFAM respectivos. Luego identificamos targets que presentaran una arquitectura de dominios dada por el par de PFAM involucrados con la idea de analizar la posibilidad de que los mismos sean drogables y que pudiera existir un vínculo entre los mismos y las drogas arriba identificadas.

Dominios	Drogas	α validacion	Red al azar
“PFAM780" "PFAM621" *CNH RhoGEF*	D 549548 (**metabisulfito de potasio**) D 17322(adenosine a 2 a3 receptor modulator antagonist) D 581648 (latrunculi a)	2.163809×10^{-10}	2%
"PFAM51" "PFAM24 *Kringle PAN 1*	D 173221 D 172257 D 176011	1.153909×10^{-08}	3%

TABLA 6.10 .Pares usados como semillas,que aparecen en la proyección de proteinas pero no de proteínas drogables,utilizadas para hacer VS en la capa de drogas. Se notan los usos comunes de las drogas obtenidas y se caracteriza la calidad de la validación usando una comparación con priorización desde 4 anotaciones al azar y el método visto en [36]. Números más pequeños de las características de la validación, implica que las drogas obtenidas no pueden haber sido obtenidas de semillas al azar.

Al realizar el procedimiento de priorización descripto se obtienen los resultados de la **tabla 6.10.** Se observa que las drogas que han sido validadas muestran una probabilidad muy baja de haber aparecido en priorizaciones con semillas al azar y además un α mucho menor a 0,05. Para el primer enlace que involucra a los dominios *CNH RhoGEF*, las drogas obtenidas están relacionadas con enfermedades del corazón como *adenosina a2 a3* receptor modulador antagonista y *latrinculin a* , pero en menor medida ha sido estudiado y resulta de interés el efecto del metabisulfito de potasio , que es un conservante ampliamente utilizado.

En el caso de los dominios *Kringle PAN 1* las drogas validadas están relacionadas a factores de coagulación y plasminogenesis. Interesantemente se encontró en bibliografía que una de las 7 proteínas que presentan ambos dominios en su arquitectura y no reportaba bioactividades en nuestra red, es en realidad target de la droga de TDR D 173221 , priorizada por los PFAM del enlace correspondiente [37].

6.6.Conclusión

Hemos proyectado a la capa de anotaciones PFAM dando lugar a 6 grafos distintos. Hemos comprobado que las proyecciones usando sólo proteínas bioactivas eran más transitivas que las de todas las proteínas, lo cual permite decir que los dominios se asocian para permitir actividad sobre las proteínas. Llegamos a conclusiones similares al estudiar la asortatividad de Rscore. Establecimos que los dominios más centrales no coinciden con los de mayor entropía, lo cual puede tener un efecto negativo a la hora de priorizar.

Además reconocimos conjuntos de 4 dominios totalmente conexos, 4-*cliques,* para la proyección StatVal de drogas bioactivas. Hicimos un ejercicio de priorización desde los cliques a la capa de drogas, y constatamos que al hacer esto se obtienen drogas con un espectro de acción similar. Por otro lado encontramos dos enlaces con dominios que aparecen en la proyección de proteinas específicamente drogables pero no al considerar el conjunto completo de proteinas. Se comprobó en los dos caso que actuaban sobre mecanismos comunes, aunque para los resultados del primero no hay estudios que soporten la acción de una de las tres drogas seleccionadas en el mecanismo. Finalmente se estableció que una de las proteínas que tenias dos de los dominios seleccionados en su estructura tenía una bioactividad desde una de las drogas seleccionadas,que no se encontraba en la red.

Capítulo 7

Crecimiento de la red y priorización

Existe un aspecto aún no explorado en nuestros datos relacionado con la dimensión temporal. El análisis de la evolución cronológica de este tipo de redes quimio-genómicas resulta interesante debido a que puede servir para entender características y modos de funcionamiento tanto de la industria farmacéutica como de la dinámica de proyectos de investigación biomédicos [38].

En nuestro caso usaremos la dimensión temporal para analizar particularidades del crecimiento de nuestra red, como por ejemplo: cómo y con qué frecuencia son incorporadas nuevas drogas a la red a medida que pasa el tiempo o qué relación guardan las nuevas bioactividades reportadas con las que las preceden. Entender estos mecanismos nos servirá para contextualizar los procedimientos de priorización in-silico que implementaremos para *hacer crecer a la red* a través de sugerencias de nuevas posibles bioactividades entre drogas conocidas y posibles nuevos targets. En consonancia con la visión general de este trabajo nos dedicaremos principalmente a buscar blancos en organismos asociados a enfermedades tropicales desatendidas.

7.1. Análisis de las conexiones droga-proteína

7.1.1 Análisis atemporal de las conexiones droga-proteína

Como primer paso, relegamos la descripción cronológica y comenzaremos nuestro análisis considerando la totalidad de las bioactividades reportadas en nuestra red que involucran proteínas alcanzadas por una única droga. Denominaremos a las mismas 'targets únicos' y al conjunto de drogas que las alcanzan 'drogas únicas' (ver figura 7.1). Así mismo, dentro del grupo de 'targets únicos' reconocemos un subconjunto de las mismos alcanzados por drogas de grado k=1, que denominaremos 'targets hiper-únicos' (se trata de proteínas alcanzadas por una única droga que no posee otros targets). A su vez llamaremos 'drogas hiper-únicas' a aquellas 'drogas únicas' mono-dirigidas, es decir de grado k=1.

Como ya se ha mencionado en el capítulo 3 en la red hay un total de 177506 drogas bioactivas y 6051 proteínas drogables. De ellas hay en total 386 'targets únicos' alcanzados por un set de 262 'drogas únicas'. Sólo 1 de cada 3 (86 de 262) drogas dirigidas a targets únicos tienen a su vez un único target (ver figura 7.2) Esto

sugiere que sólo esta fracción de drogas sea posiblemente cualitativamente novedosa y mono-dirigida.

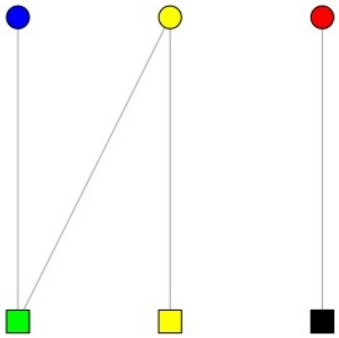


FIGURA 7.1.Esquema de un conjunto de nodos droga y proteína. Los círculos representan proteínas y los cuadrados representan drogas. En azul se muestra un target único, recibe un solo enlace .La droga (verde) que recibe el enlace es por tanto una “droga única” es promiscua, ya que tiene dos enlaces. En rojo se marca un target “hiper-unico” este tiene asociado una única droga, que a su vez sólo tiene un enlace .En negro se marca la droga que es hiper-única.

En el mismo sentido, en la **figura 7.2** mostramos la fracción de targets-únicos alcanzados por drogas-únicas con más de un target asociado. Vemos que en el 80% de los casos los 'targets únicos' son alcanzados por drogas que tambíen presentan bioactividades hacia targets que no son únicos. Esto podría estar reflejando la importancia del reposicionamiento de drogas como estrategia de expansión de targets drogables.

.

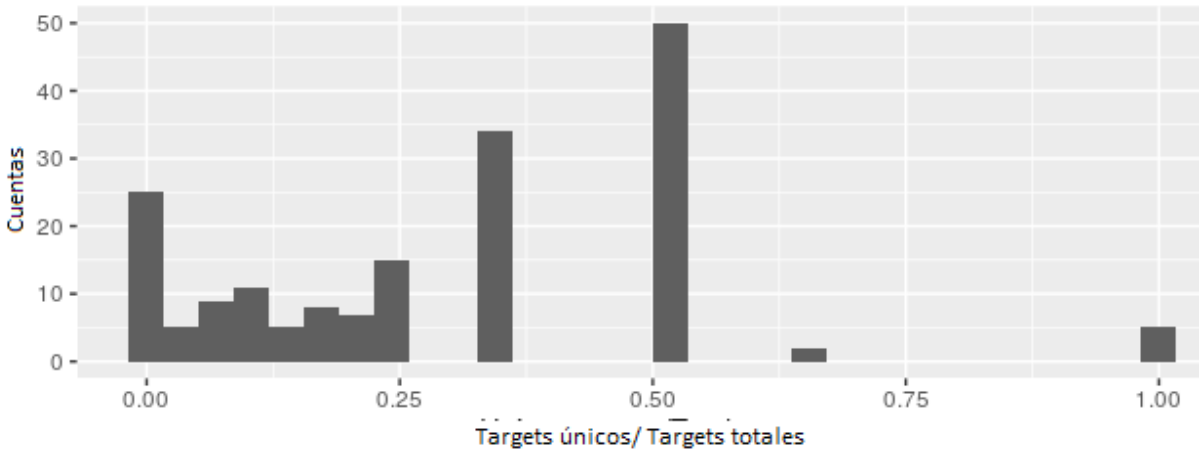


FIGURA 7.2: Proporción de targets únicos sobre el total de targets para una dada droga que apunta al menos a un target “único”. Es decir un target que tiene solamente asociado una droga.

Como mostramos en la **figura 7.3**, es interesante reparar en que más del 80% de las 'drogas únicas' tienen como target al menos una proteína de un organismo modelo. Esto apoya la idea de la importancia del reposicionamiento de drogas como estrategia de desarrollo de fármacos. Para organismos no-modelo, las bioactividades asociadas a estas drogas surgirían en general a partir de pruebas previas en otros organismo.

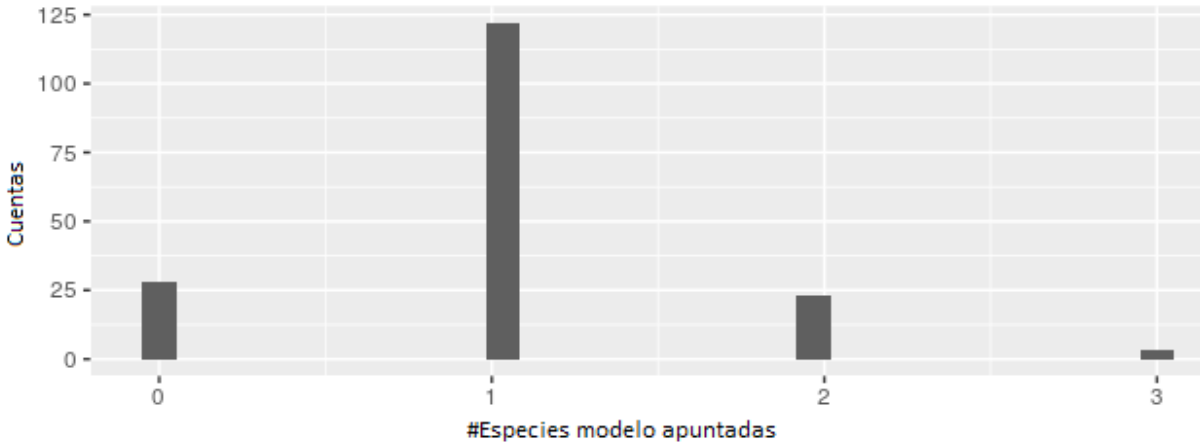


Figura 7.3.Número de organismos modelos distintos apuntados por drogas únicas.

7.1.2 Evolución temporal

Los resultados de la sección anterior parecen sugerir que gran parte de las nuevas bioactividades que involucran a targets aún no explorados surgen a partir de drogas ya conocidas. Sin embargo, para poder avanzar en la caracterización de esta tendencia de exploración, es necesario evaluar cómo se establecen vínculos de bioactividad a nuevas proteínas a lo largo del tiempo.

Las drogas en la red TDR fueron extraídas de 3 bases de datos DrugBank, Chembl, y Pubchem. Para poder establecer las fechas en que se publicó cada bioactividad, hubo que partir estas bases de datos, minar la información y unificarla con la red. Se logró hacer esto para el 97% de las drogas con bio-actividad conocida de la red . Es decir, sólo para 14926 drogas de un total de 494294 no se pudo consignar una fecha de publicación.

En la **figura 7.4**, donde se reportan las fechas de primera publicación por droga, se puede observar la tasa de crecimiento del universo de drogas exploradas en investigaciones biomédicas. La adquisición de datos de la versión de TDR utilizada para armar nuestra red se produjo mayoritariamente hasta el año 2009, lo que explica la disminución observada como un artefacto técnico.

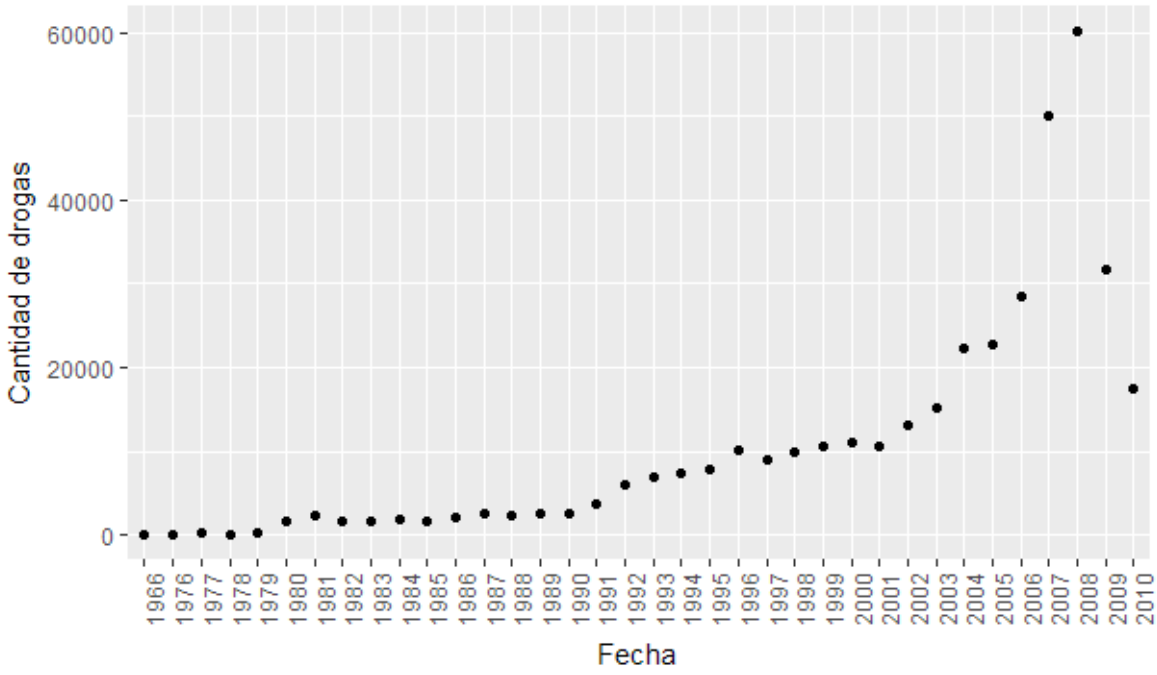


FIGURA 7.4. Gráfico con la cantidad de drogas nuevas por año. Es decir aquellas para las que se encontró su primera bioactividad publicada en aquel año.

Con los datos temporales recabados, es posible estudiar patrones que describan el modo de crecimiento de la red, es decir de qué manera surgen nuevas bioactividades en el contexto de las ya reportadas. Para ello, recabamos la información del momento en que cada bioactividad fue reportada, y siguiendo conceptos introducidos por Yildrim y colaboradores [37], clasificamos a las drogas de nuestra red en dos categorías: *saltadoras* (*hoppers*) o *reptantes* (*crawlers*) (ver **figura 7.5**). Las primeras incluyen drogas para las cuales su primer bioactividad reportada no involucra targets ya alcanzados por alguna droga anteriormente introducida en la red. Las segundas, por el contrario, sí involucran conocimiento de alguna manera ya embebido en la red porque refieren a nuevas bioactividades que vinculan: (a) targets nodvedosos (aún no alcanzados por ninguna droga) con drogas ya incorporadas a la red, y (b) drogas nuevas con targets que ya reportan bioactividades asociadas (paneles superior-derecho e inferior-derecho de la **figura 7.5**, respectivamente)

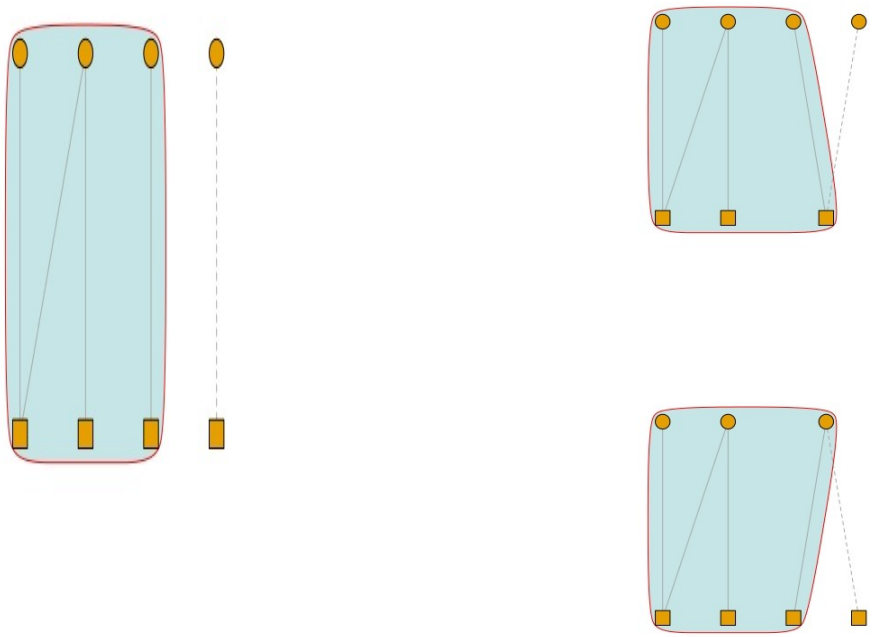


FIGURA 7.5. Modos de crecimiento de la red droga(cuadrados) proteínas (círculos) con la red original marcada en un globo celeste. En el panel izquierdo se muestra el crecimiento por drogas *saltarinas* (hopping), se agrega una droga a la red de tal manera que es bioactiva a una droga novedosa (nueva no unida previamente al resto de la red), y forman una entidad disconexa del resto del grafo. En los paneles derechos se ilustran mecanismos de *crawling*. En la parte superior del panel derecho se obseva el crecimiento a través del agregado de una proteína novedosa, mientras que en el panel inferior el crecimiento es vía el agregado de una droga que tiene al menos una bioactividad a proteínas viejas (i.e. ya unidas al resto de la red).

En el panel superior de la **figura 7.6** se observa la evolución del número de drogas de cada categoría en función del tiempo. Es posible apreciar que la mayor parte de las incorporaciones a la red involucra drogas del tipo *reptante*. Las *saltadoras* por otra parte, no sólo representan una fracción pequeña del total de incorporaciones, sino que su tasa de incorporación se mantuvo relativamente constante durante el período analizado. En el panel inferior de la misma figura, incluimos información sobre el crecimiento de las bioactividades reportadas año a año, diferenciando contribuciones que apuntan hacia targets aún no alcanzados por ninguna droga.

A partir de los datos se puede ver que se reportaron 6339 drogas *saltadoras (panel izquierdo* **figura 7.5***)* para el período 1970–2010, lo que implica que sólo el 10% de todas las drogas aparecieron por primera vez asociadas a targets novedosos. A su vez, los blancos de tales drogas constituyen el 56% de todas los targets nuevos (ver panel izquierdo), mientras que el otro 44% corresponde a crawling (ver **figura 7.5** panel izquierdo inferior) introducidas durante todo el periodo en estudio. El 74% de la bioactividades agregadas corresponden a targets viejos, por ende, en general la mayoría de las drogas se agregan por reptación sobre la red aumentando mayoritariamente enlaces a targets viejos.

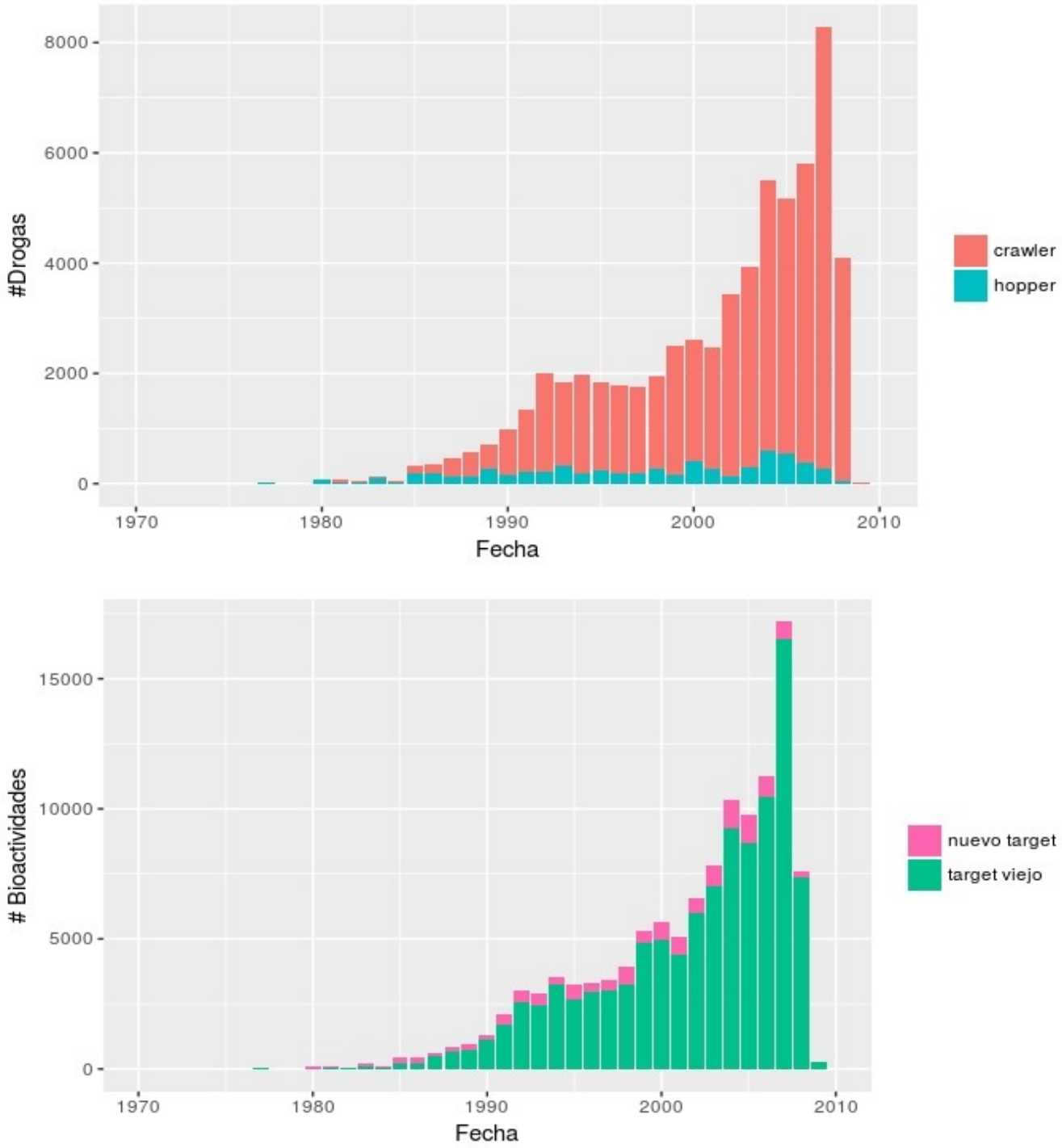


FIGURA 7.6. Gráficos obtenidos teniendo en cuenta todas las drogas y proteínas de la red TDR fechadas. En el panel superior se muestran la cantidad de drogas nuevas por año y se hace la distinción si son hoppers (comparten una bioactividad en la fecha de inclusión con una droga previa) y crawlers (que si comparten targets con una droga anterior). En el panel inferior se muestran la cantidad de bioactividades nuevas por año, y se hace la distinción entre bioactividades a targets, con bioactividades anteriores o no.

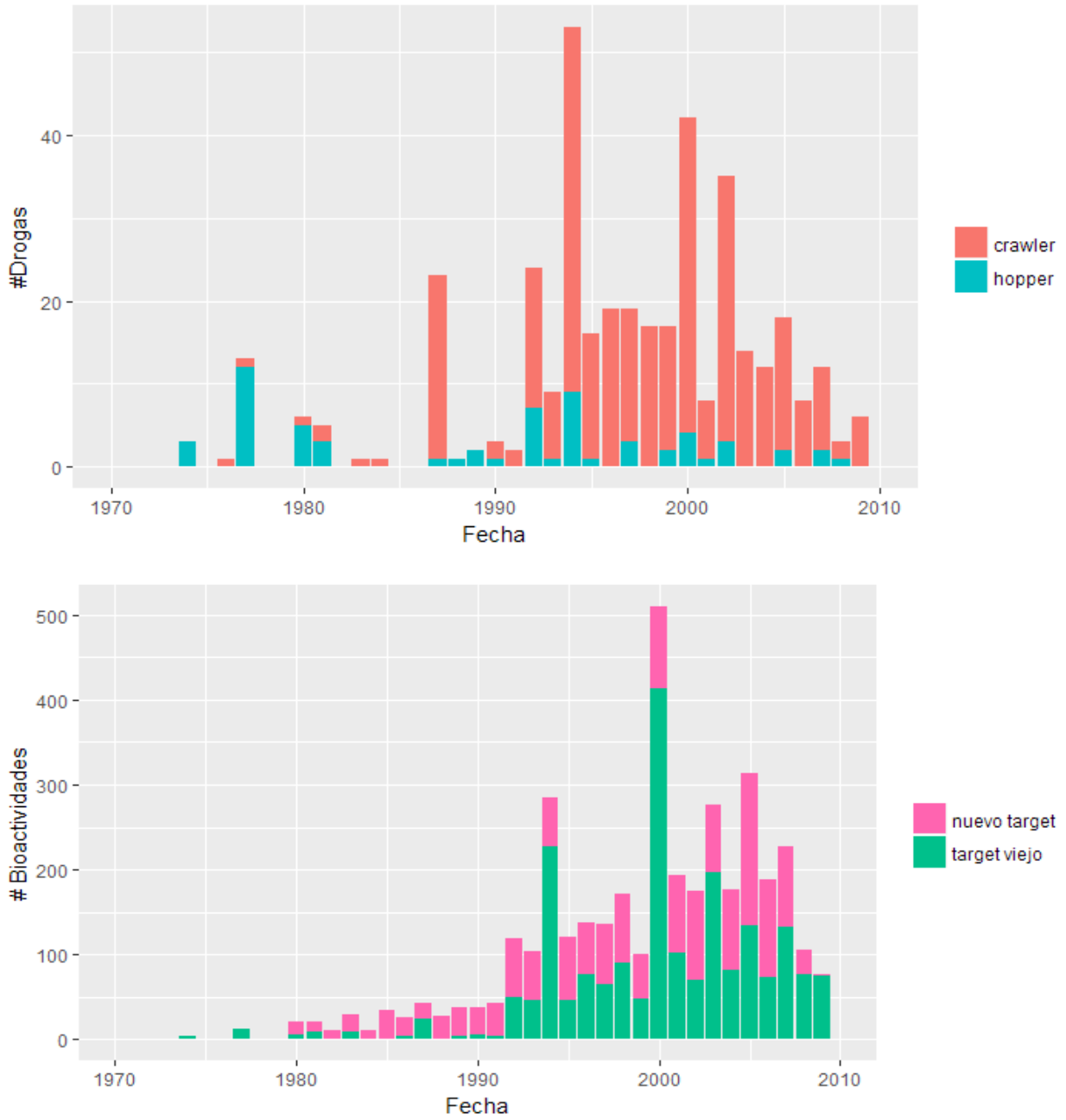


FIGURA 7.7. Gráficos obtenidos teniendo en cuenta todas las drogas aprobadas para el uso en humanos y proteínas de la red TDR fechadas.

Finalmente, la **figura 7.7** muestra la evolución temporal del número de drogas y de bioactividades aprobadas por la FDA para uso en humanos. Lo observado permite sugerir un escenario para el desarrollo de fármacos asociados a TDR. Por un lado, la abundancia de drogas tipo 'crawlers' sugiere que al momento de introducir la bioactividad en la especie asociada a TDR , la droga ya presentaba bioactividad a un target de otra especie. Por otro, la fracción de targets nuevos respecto a los conocidos

es mayor teniendo en cuenta el grafo de solo estos organismos, en particular el 0,269 para el caso de sólo bioactividades aprobadas en humanos, contra el 0,097 de la red original. Todo esto es compatible con una estrategia de reposicionamiento de drogas para alcanzar targets novedosos en patógenos TDR.

7.2. Priorización

A partir de lo visto en la sección anterior, el modo típico de crecimiento de la red quimio genómica involucra una dinámica donde las nuevas asociaciones entre drogas y targets no son del todo novedosas, sino que se establecen, por ejemplo, involucrando targets que ya presentan bioactividades reportadas con otras moléculas. Esto es compatible con la idea de que, en la práctica, la industria avanza utilizando una estrategia conservadora, tipo *crawling*, desde el conocimiento ya adquirido.

En el mismo sentido es posible entonces pensar en elaborar una estrategia *in-silico* que a partir del conocimiento de bioactividades conocidas prediga nuevas asociaciones entre targets y drogas. Esto es justamente lo que planeamos hacer en este capítulo presentando metodologías de priorización sobre nuestra red. Consideraremos para ello el esquema de priorización introducido en la sección 2.3.1 , para llevar adelante tareas de priorización de especies completas. Es decir vamos a retirar toda la evidencia de bioactividad relacionada con una especie de interés y, mediante técnicas computacionales, puntuaremos a cada target de esa especie con un score de drogabilidad obtenido utilizando el resto de la info embebida en la red. Utilizaremos las métricas de desempeño que presentamos en 2.3.2 como las estrategias de evaluación de nuestro procedimiento predictivo. Finalmente trataremos de contextualizar los resultados obtenidos a la luz de propiedades estructurales y de flujo de información de nuestra red.

Para llevar adelante nuestro objetivo, consideramos la red de proteínas obtenida mediante la proyección ProbS a partir de la red bipartita targets-anotaciones e identificamos como semillas a targets drogables de especies diferentes a la que interesa analizar. Para hacer la proyección tuvimos en cuenta la información sobre la relevancia *a-priori* de cada anotación capturada en los respectivos valores de Rscore (5.2). Para ello definimos a la matriz de conectividad proteína-proteína de la red proyectada,W, según:

$$W = \hat{A}R\hat{A}^t . \qquad (7.1)$$

donde A es la matriz de adyacencia y R es la matriz diagonal de elementos r_{ll}, es decir el Rscore de la anotación l.

Como estrategia de priorización utilizamos un esquema de votación o, como se conoce en la jerga, KNN (K Nearest Neighbours) con K=1, sobre la red de proteínas proyectada con ProbS- El método de validación que consideramos fue evaluar el AUC0.1 en la curva ROC, que compara la tasas FPR y TPR.

7.2.1. Validación de los métodos de priorización y vinculación target-species

El protocolo de análisis presentado fue aplicado para llevar adelante la priorización de genoma completo de 5 especies diferentes: 3 especies de organismos modelo y otras 2 especies de patógenos. Los patógenos considerados fueron: *tripanosoma Cruzi* (causante del mal de chagas) y *Plasmodium falciparum* (causante de la malaria). *Homo Sapiens*, Mus musculus (ratón), y *Saccharomyces Cerevisiae* (levadura) fueron los modelos considerados. *Mus Musculus* es usado principalmente para pruebas de laboratorio antes de habilitar el uso de una droga en humanos. Mientras que *Saccharomyces Cerevisiae* es un organismo que se ha estudiado de manera exhaustiva debido a que es un organismo de fàcil manejo para el cual existen numerosas técnicas experimentales desarrolladas.

En la **tabla 7.3** se presenta el resultado de la validación de las predicciones obtenidas en cada caso

	tcr	pfa	mmu	hsa	sce
AUC 0.1	0.851	0.705	0.745	0.641	0.495

TABLA 7.3.Resultados de AUC 0.1 para la validación de organismo completo de *Homo sapiens* (hsa),*Mus musculus*(mmu), *Plasmodium falsiparium* (pfa), *Saccharomyces cerevisiae*(sce), y *Trypanosoma crucis* (tcr).

Como vemos la priorización en los organismos patógenos con la metodología propuesta resultó muy satisfactoria. También se obtuvo una buena validación para

ratón. Sin embargo vemos que para humano y levadura los resultados no fueron del mismo tipo. En particular, se obtuvo para esta última especie valores de AUC.01 cercanos a 0,5, que como vimos en la sección 2.3.2 , corresponde a realizar una predicción aleatoria sobre la condición de drogabilidad. En la siguiente sub-sección tratamos de comprender porque se da esta diferencia.

7.2.2 Estructura, flujo en la red y priorización

Con el fin de entender el score alcanzado en nuestro ejercicio de recomendación recordemos que removemos toda la información de una dada especie sp, por lo que nos valemos del conocimiento embebido en la red respecto a las bioactividades del resto. Por esta razón vale la pena estudiar como es el patrón de interconexión y el flujo de información entre diferentes especies inducido por el proceso de priorización. Para eso introducimos las nociones de: participación, *P,* y de extra-especificidad,*O*:

$$P(i) = 1 - \sum_j \left(\frac{S^{sp}_{i,j}}{\sum_j S^{sp}_{i,j}}\right)^2 \quad \textbf{(7.2)}$$

$$O(i) = \frac{\sum_{j|j\neq sp} S^{sp}_{i,j}}{\sum_j S^{sp}_{i,j}} \quad \textbf{(7.3)}$$

con, S^{sp} es la matriz de puntuación para la validación en la especie *sp,* cuya componente *ij* corresponde a la puntuación para una proteína $i$en sp generada por la proteínas de la especie *j*.

Vemos entonces que mientras que la participación està asociada a la diversidad de especies desde la que una dada proteína recibe puntuación, la extra-especificidad cuantifica la fracción de la puntuación total que proviene de contribuciones desde otras especies.

En nuestro caso, el score total $\Sigma(i)$ que recibe una proteína de una especie dada resulta: $\Sigma(i) = \sum_{j\neq sp} S^{sp}_{i,j}$.

Para entender el resultado del ejercicio de validación que se reporta en la **tabla 7.3**, las **figuras 7.7** , **7.8** y **7.9** muestran la relación entre participación y extra-especifidad para cada una de las proteinas drogables de las distintas especies consideradas. Cada símbolo representa una proteína drogable, el tamaño es proporcional al score recibido y su posición en el plano refleja el flujo de score recibido a partir del conexionado interespecie. Se puede demostrar que la forma parabólica interior corresponde al caso en que hay dos especies que aportan score en parte

iguales a la proteína. En este caso se obtiene que la participación en función de la extra-especificidad resulta $P = -2(O^2 - O)$.

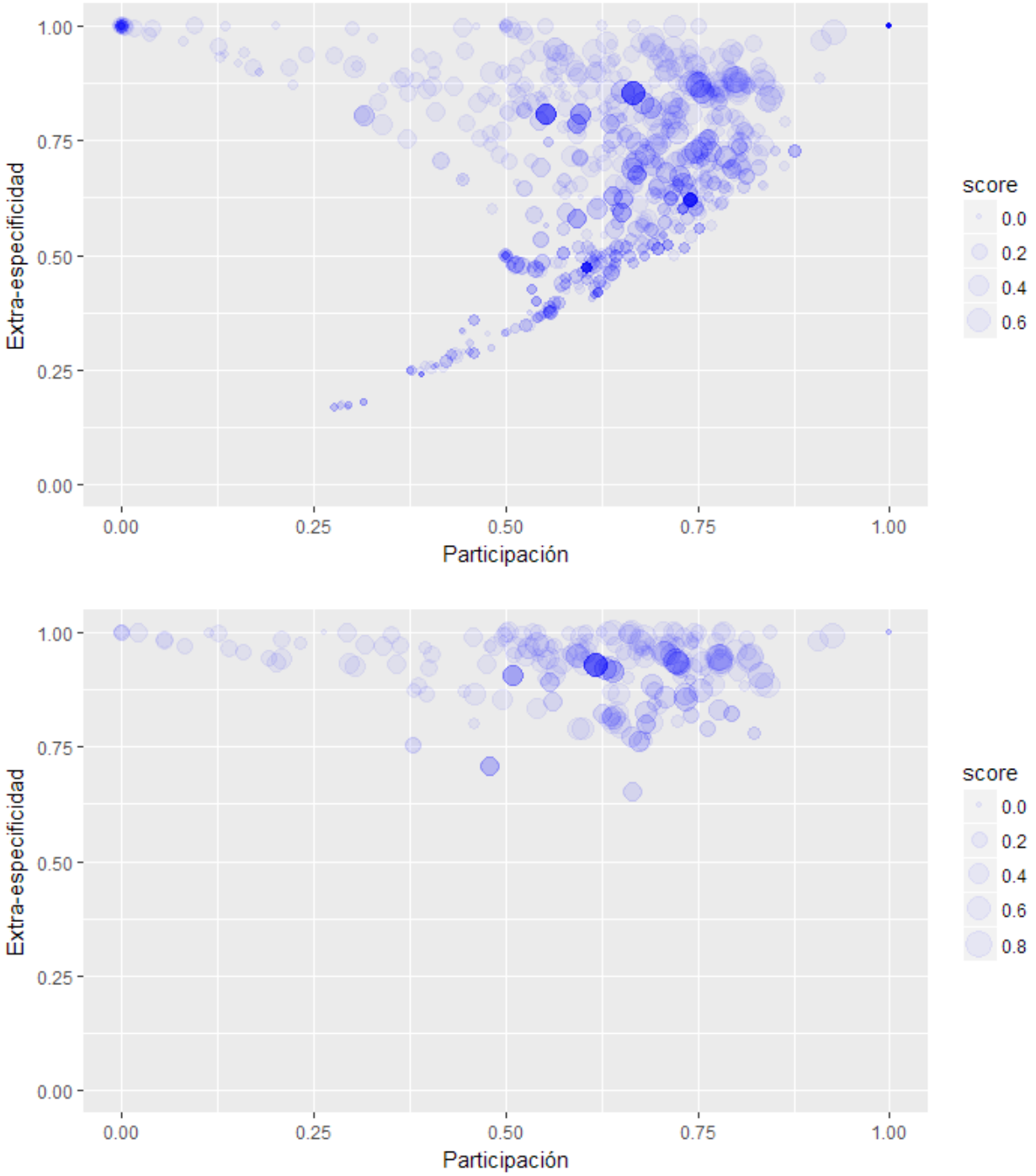


FIGURA 7.7. Gráfico con los resultados de scoring para proteínas drogables y parámetros de borde para *Homo Sapiens* (panel superior) *Mus musculus (panel inferior).* La extra-especificidad hace referencia a cuanta puntuación proviene de otra especie y la participación a cómo se encuentra esta distribuida

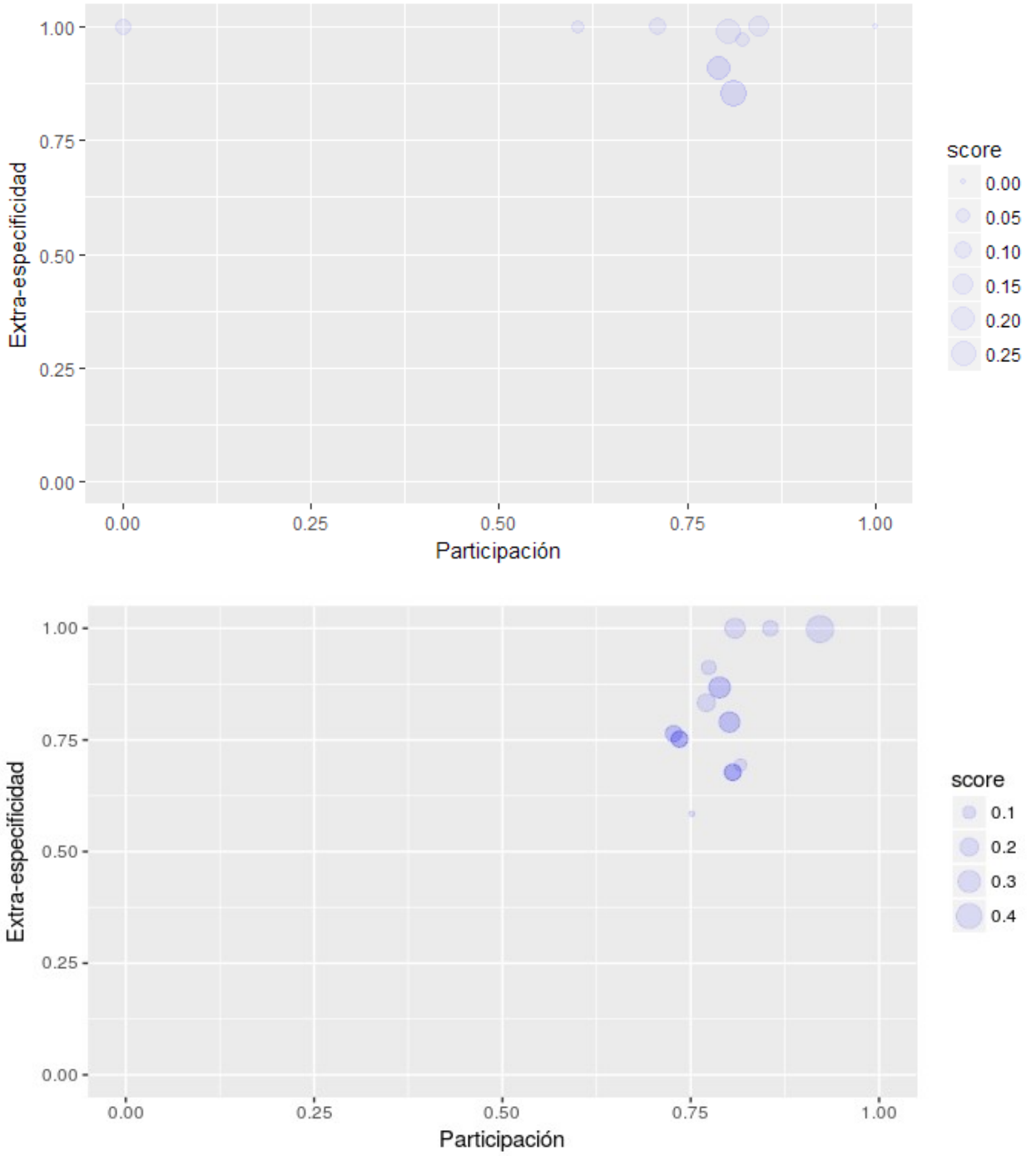


FIGURA 7.8. Gráfico con los resultados de scoring para proteínas drogables y parámetros de borde para *Plasmodium falsiparium (panel superior)* y *Trypanosoma cruzi* (panel superior).La extra-especificidad hace referencia a cuanta puntuación proviene de otra especie y la participación a cómo se encuentra esta distribuida.

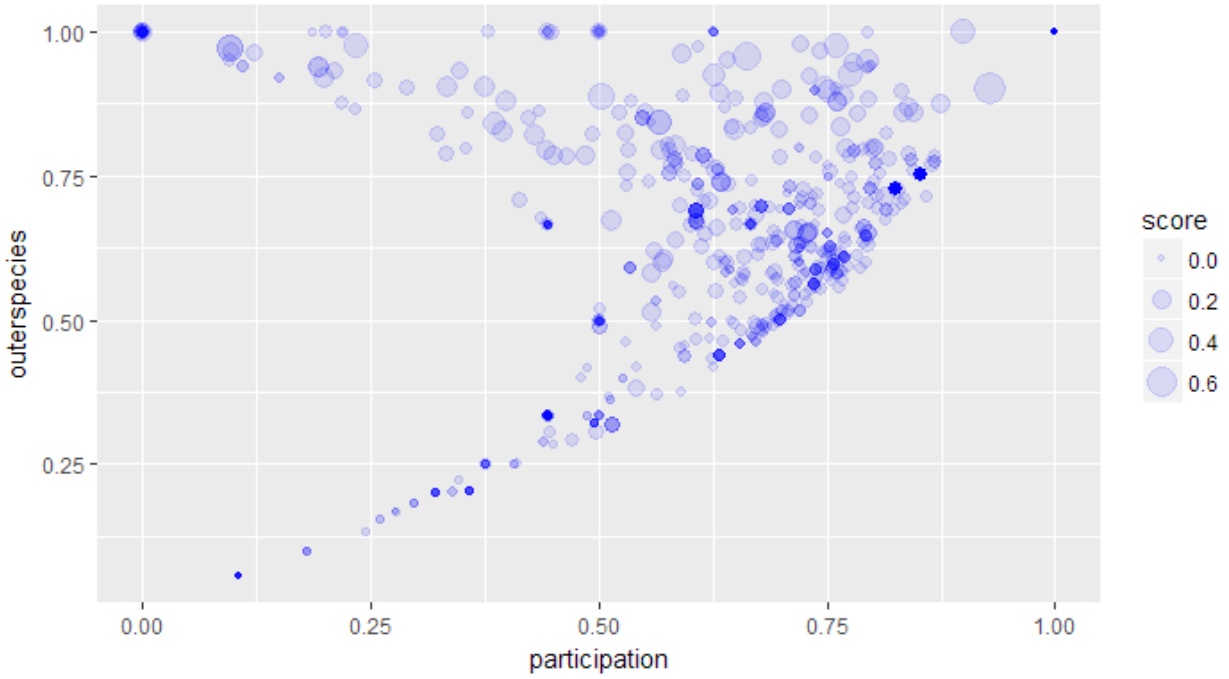


FIGURA 7.9. Gráfico con los resultados de scoring para proteínas drogables y parámetros de borde para *Saccharomyces cerevisiae*.

En las **figuras 7.7**, **7.8** se observa que para las especies que presentaron altos valores de AUC01 (mmus, pfa, tcr), los valores de extra-especificidad de sus proteinas drogables se mantienen por encima de 0,6, lo que favorece la obtención de altos valores de score en el proceso de priorización de genoma completo. Más aún, casi la totalidad de las proteinas de los patógenos analizados, presentan así mismo altos valores de participación que da cuenta de que el score recibido proviene de flujo de información proveniente de diversas especies.

La situación para humano y levadura, es muy diferente. En ambos casos se reportan bajos niveles de AUC0.1 en el proceso de validación y esto se debe, como vemos en la el panel superior de la **figura 7.7** y en la **figura 7.9**, a que existe una fracción considerable de sus targets drogables que presentan valores de extra-especificidad menor a 0.5. Esto indica que son proteínas "difíciles de alcanzar" desde otras especies en la priorización , por lo que no reciben buen score. Más aún es posible notar que en este caso varias proteínas drogables se encuentran en el caso límite de sólo dos especies que aportan puntajes.

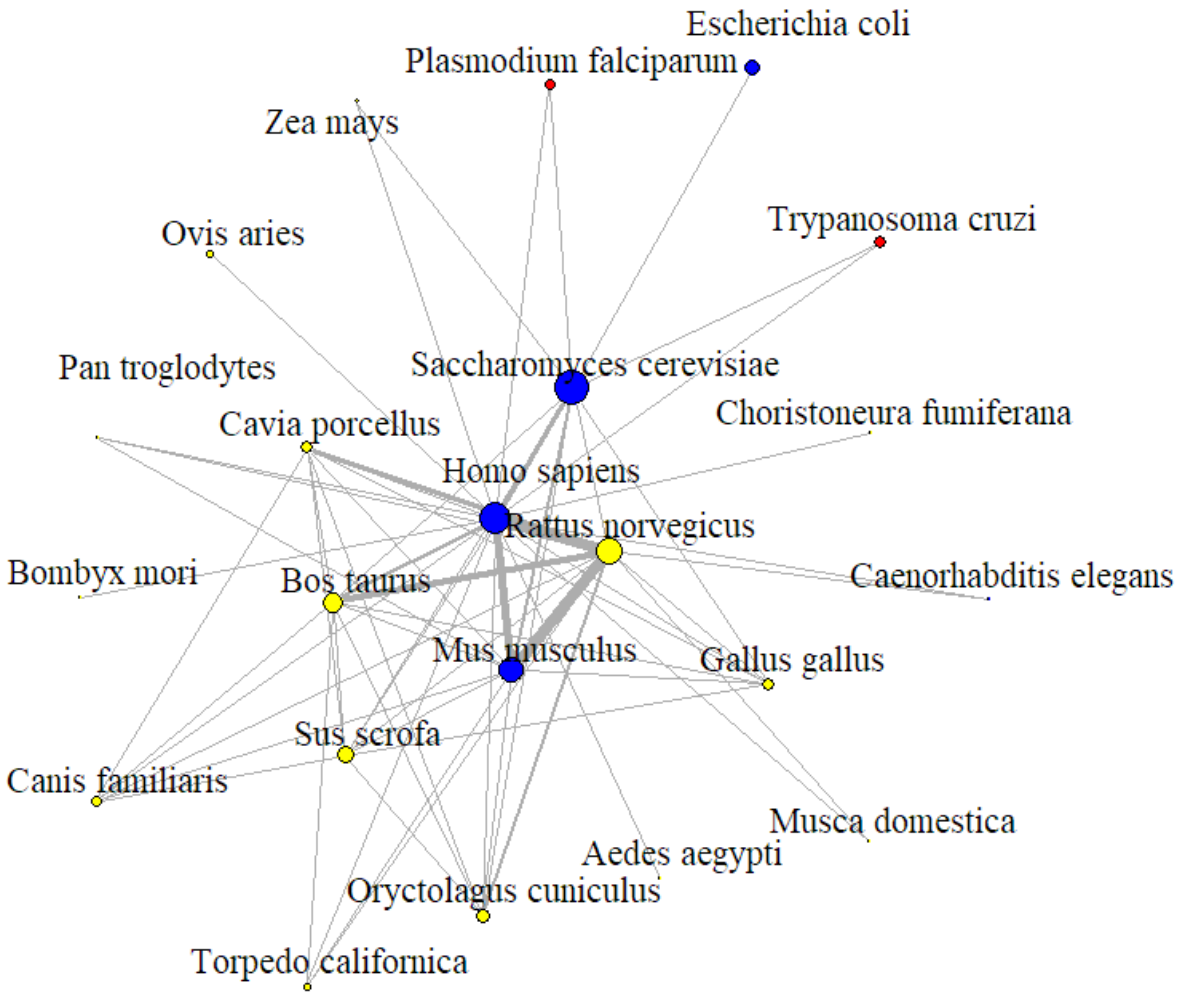


FIGURA 7.10. Gráfico de conexitud mediada por anotaciones entre las especies con proteínas drogables y con un peso mayor a 50. El ancho de cada arista está dado por el logaritmo de la cantidad de proteínas y anotaciones que comparten las especies en sus extremos normalizado por la cantidad de proteínas drogables de dichas especies. En azul se muestran los organismos modelo, y en rojo los organismos patógenos relacionados a NTD. El tamaño de los nodos está dado por el logaritmo del número de proteínas drogables en cada especie.

Es interesante notar que el rango de valores de score de priorización recibido por targets de distintas especies es diferente (ver las escalas de **figura 7.7-7.9**) Por ejemplo las especies pfa y tcr tienen valores dentro del rango 0-0.4 mientras que humano y mmus dentro de 0-0.8. Esto también es un reflejo de la estructura de interconectividades de la red. Para poder entenderlo mostramos en la **figura 7.10** la red de especies, con el ancho de cada arista dado por el logaritmo de la cantidad de caminos mediados por proteínas drogables y anotaciones entre las especies, normalizado por la cantidad de proteínas drogables de las especies. Este esquema

sirve por lo tanto para ilustrar el patrón de interconectividad entre especies puesta en juego en los procesos de priorización. Es posible reconocer que existe un core de especies fuertemente interconectadas que incluye a organismo modelo como *Homo sapiens, Mus musculus, Rattus norvegicus, Bos taurus, Saccharomyces cerevisiae* y *Cavias porcellus*. Y entonces se entiende que el proceso de difusión de información alcance con diferente intensidad targets de diferentes especies, y en particular, en menor medida a a tcr y pfa, que están en la periferia de esta red.

7.2.3 Drogabilidad y flujo

En la sección anterior vimos cómo la estructura de conexiones entre targets drogables de diferentes especies repercute en el proceso de priorización de genoma completo. Cabe ahora preguntarse, ¿las características de conexionado que observamos para el conjunto de proteínas drogables que una especie presenta en la actualidad, es particular en algún sentido? o ¿representa lo que cabría esperar para cualquier otro conjunto de proteínas de dicha especie?

En otras palabras ¿están las proteínas drogables en una situación de privilegio que las hace más fáciles de alcanzar por puntuación externa ó por el contrario son más difíciles de alcanzar que el resto del proteoma de dicha especie? Para responder a esta pregunta para cada especie analizada realizamos 1000 selecciones de targets al azar, con la restricción que tengan la “misma” cantidad de anotaciones de cada tipo que el set original. A los targets seleccionados así los consideramos drogables para esa iteración (semillas a los usos de la priorización). Luego registrabamos el scoring recibido, así como la externalidad promedio y la participación promedio recibido por el set de proteínas semillas considerado en cada iteración. Utilizando las medias y dispersiones de estos observables estimamos finalmente parámetros estandarizados Z para la externalidad y participación media del conjunto de proteínas drogables en nuestro dataset original.

$$Z_x = \frac{<x> - x_{real}}{DS(x)}. \qquad \textbf{(7.4)}$$

donde x es una variable aleatoria (en nuestro caso los resultados de cada selección de semillas, promediados sobre todas las proteínas) $< x >$ y $DS(x)$ es el valor promedio y desviación estándar sobre todas las realizaciones, y x_{real} es el valor obtenido al calcular x con las semillas de la red real.

La **tabla 7.4** muestra que los z encontrados- para especies modelo son mucho mayores que para especies patógenas. Esto daría en principio una explicación de

porque todos los métodos de scoring dan mejor para patógenos: los targets drogables que se pretende recuperar en el ejercicio de priorización de especie completa tienden a aparecer en la interfase con otras especies, posiblemente debido a que la identificación de las bioactividades respectivas tuvo lugar vía alguna forma de *crawling*, a partir de conocimiento disponible desde otras especies.

	Z participación	Z externalidad
Homo Sapiens	37,75	-48,8
Mus musculus	11,24	-26,36
Plasmodium Falsiparium	-0.208	1.58
Sacharomyces cerevisae	-146,98	-142,93
Trypanosoma cruzi	-2.58	0.54

TABLA 7.4. Datos obtenidos usando la ecuación 7.4 al seleccionar proteínas al azar y tratarlas como si fueran drogables en el proceso de validación antes visto.

Al analizar con más cuidado la **tabla 7.4** Se ve que existe el siguiente patrón, valores de Z_participacion negativos y valores de Z_externalidad positivos, indican que la distribución real tiene más conexiones al exterior que si se tomaran semillas al azar. en las **figuras 7.7-11** esto coincidiría con puntos en la región izquierda superior.

7.3 Conclusión

Analizando la estructura de los enlaces de bioactividad hemos establecido que en la gran mayoría de los casos que las drogas que actuen sobre proteínas únicas tengan actividades sobre proteínas en otros organismos. Más aún al tener en cuenta la fecha de primer publicación de las bioactividades hemos establecido, que los enlaces “nuevos” en la red droga-proteína aparecen principalmente por “crawling”, es decir usan información ya existente en la red.

Además utilizamos un método de priorización a primeros vecinos, como propusimos a principios de este trabajo, y lo hemos validado en 5 especies distintas al tratar de hacer una recuperación de proteoma completo. Las validaciones fueron notablemente mejores para especies patógenas que para el resto de las utilizadas. Finalmente propusimos dos cantidades que evalúan de qué manera llega la puntuación a las proteínas y lo utilizamos para explicar el desenvolvimiento de las priorizaciones.

Capítulo 8

Conclusiones

En el capítulo 3 de este trabajo se introdujo la red TDR y se explicó de qué manera se redujo su tamaño al tomar clusters para dos tipos de similitud: Tanimoto y subestructura,a partir del análisis de estructuras identitarias . Estos permiten realizar una descripción en términos de estructuras 'coarse grained' y reducir así el tamaño efectivo de la red en ambos tipos de métricas de similaridad En el capítulo 3 se analizó la coherencia entre estos dos tipos de clusters analizando los valores de persistencia y comparando la distribución de peso. La persistencia indicó que la partición generada por clusters subestructura, es mucho más fuerte que la de tanimoto, en el sentido de que la persistencia es mucho mayor al pasar de una partición de la capa en clusters S a T que al revés. Finalmente establecimos un criterio de comparación y calibración entre el valor numérico de la similitud por subestructura y Tanimoto. La escala de calibración entre ambas medidas permite apreciar que para un mismo valor de similaridad la relación de subestructura se corresponde con una mayor coincidencia de blancos compartidos.

Luego estudiamos las capas de proteínas y anotaciones. Detectamos categorías que presentan preferencia para anotar proteínas con bioactividades reportadas, a través del uso de la estadística de fisher y la definición del Rscore. Vimos que esto que es un reflejo de sesgos que efectivamente existen en el área de desarrollo y búsqueda de nuevos fármacos. Muchas categorías tienen un rol muy importante en la interconectividad entre proteínas de diferentes especies, lo cual asociamos con la entropía de la distribución de proteínas por especie. Medida que es muy relevante para transmitir información por la red sobre conocimiento adquirido con sesgo hacia ciertas especies. Además analizamos la coherencia entre los distintos tipos de anotaciones a través de similaridades a primeros vecinos en la red. Establecimos que la información de la capa de anotaciones ortólogas es semejante a la de PFAM . Más aún establecimos la existencia de un vínculo entre drogas comunes entre proteínas y la cantidad de anotaciones PFAM compartidas. Se puede observar una clara relación creciente entre ambas magnitudes y que aproximadamente el 70% de proteínas que comparten 3 o más dominios en su arquitectura son alcanzadas por al menos una droga en común.

En el capítulo proyectamos a la capa de anotaciones PFAM dando lugar a 6 grafos distintos, por 3 métodos de proyección distintos y por la selección de todas las proteínas o sólo las drogables. Hemos comprobado que las proyecciones usando sólo proteínas bioactivas eran más transitivas que las de todas las proteínas, lo cual permite decir que los dominios se asocian para permitir actividad sobre las proteínas. LLegamos a conclusiones similares al estudiar la asortatividad de Rscore. Establecimos que los dominios más centrales no coinciden con los de mayor entropía, lo cual puede tener un efecto negativo a la hora de priorizar. Además reconocimos conjuntos de 4 dominios totalmente conexos, *cliques,* para la proyección StatVal de drogas bioactivas. Hicimos un ejercicio de priorización desde los cliques a la capa de drogas, y constatamos que al hacer esto se obtienen drogas con un espectro de acción similar. Por otro lado encontramos dos enlaces con dominios que aparecen en la proyección con el total de las proteínas, pero no en la específica de bioactivas. Se comprobó en los dos caso que actuaban sobre mecanismos comunes, aunque para los resultados del primero, no hay estudios que soporten la acción de una de las tres drogas seleccionadas en el mecanismo. También se encontró que una de las drogas priorizadas tenía una bioactividad no incluída en la red sobre una proteína con la estructura de los dos dominios mencionados antes.

En el capítulo 7 establecimos que resulta importante en la gran mayoría de los casos que las drogas que actuen sobre proteínas únicas tengan actividades sobre proteínas en otros organismos. Más aún al tener en cuenta la fecha de primer publicación de las bioactividades hemos establecido, que los enlaces “nuevos” en la red droga-proteína aparecen principalmente por “crawling”, es decir usan información ya existente en la red.

Además utilizamos un método de priorización a primeros vecinos, como propusimos a principios de este trabajo, y lo hemos validado en 5 especies distintas al tratar de hacer una recuperación de proteoma completo. Las validaciones fueron notablemente mejores para especies patógenas que para el resto de las utilizadas. Finalmente propusimos dos cantidades que evalúan de qué manera llega la puntuación a las proteínas y lo utilizamos para explicar el desenvolvimiento de las priorizaciones.

Apéndice A

A.1. Subgrafos, Conexidad y Componente Gigante.

Dado un grafo $G(\mathfrak{N}, \mathfrak{E})$ es posible considerar un subconjunto de nodos y aristas del mismo para definir otro nuevo grafo G'. Diremos que $G'(\mathfrak{N}', \mathfrak{E}')$ es un subgrafo de $G(\mathfrak{N}, \mathfrak{E})$, si se verifica que $\mathfrak{N}' = n_1, n_2, n'_N \subseteq \mathfrak{N}$ y $\mathfrak{E}' = {e_1}', {e_2}', ..., {e_m}' \subseteq \mathfrak{E}$. Por otro lado, si $\mathfrak{E}'$ contiene todas las posibles aristas de G que unen a nodos del conjunto $\mathfrak{N}'$ se dice que G' es un subgrafo inducido o subgrafo completo y se denota simplemente por $G' = G(\mathfrak{N}')$.

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. Finalmente, 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 $G(\mathfrak{N}, \mathfrak{E})$, es una componente cuyo tamaño (cantidad de nodos) es del mismo orden que N [39].

A.1.1. 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 $G^W = G(\mathfrak{N}, \mathfrak{E}, \mathfrak{W})$ consta de un conjunto de N nodos $\mathfrak{N} = \{n_1, n_2, ...n_N\}$, un conjunto de m aristas $\mathfrak{E} = \{e_1, e_2, ..., e_m\}$ (cuyos elementos serán pares ordenados si el grafo es dirigido) y un conjunto de m pesos $W : E \to R, W = w_1, w_2, ...w_m$, cada uno de ellos asociado a la correspondiente arista en el conjunto E. Usualmente el conjunto W toma valores positivos, pero es importante destacar que tal condición no es estrictamente necesaria (ver ejemplos [40]).

A.1.2. Matriz de adyacencia y matriz de pesos

Una representación particularmente útil para grafos es mediante notación matricial. Dado un grafo no pesado $G(\mathfrak{N}, \mathfrak{E})$ de N nodos, su *matriz de adyacencia* $A \in NxN$ es una matriz cuadrada de elementos binarios$A = \{a_{ij} \mid i, j \in 1...N\}$, de forma que $a_{ij} = 1$ si existe la correspondiente arista $e_{i,j}$ y 0 en caso contrario. La matriz de adyacencia A será simétrica o asimétrica, según se trate de grafos no dirigidos o dirigidos respectivamente. Los elementos diagonales deben ser nulos en orden de satisfacer la ausencia de bucles que requiere la definición de grafo dada.Por otro lado, en caso de tratarse de grafos pesados $G_W = G(N, E, W)$, la correspondiente representación matricial suele referirse como matriz de pesos $W \in N \times N$. 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 W será simétrica sólo si el grafo es no dirigido.

A.2.Principales Observables Topológico

A.2.1 Distribución de grado, asortatividad y disasortatividad

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

$$k_i = a_{ij} \quad j \in Ne(i) \qquad \textbf{(A.1)}$$

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 A del grafo. En caso de grafos pesados $G(W) = G(N, E, W)$, con matriz de pesos dada por $W \in NXN$ 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:

$$s_i = w_{ij} j \in Nei(i) \quad \textbf{(A.2)}$$

con w_{ij}/inW**.**

Estas medidas permiten establecer la caracterización más simple posible para un grafo, que es su distribución de grado $P(K)$. 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 caracteriza por completo las propiedades estadísticas en redes no correlacionadas [40] , 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 se define un observable relacionado a esta probabilidad de mayor utilidad práctica, el grado medio de primeros vecinos. Para un nodo$i$ esta magnitud$k$

$$k_{nn,i} = \frac{1}{k_i} \sum_{j \in Nei(i)} k_j \quad \textbf{(A.3)}$$

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

$$k_{nn}(k) = \frac{1}{n} \sum_{k'} k' P(k'|k) \quad \textbf{(A.4)}$$

expresión que bajo la ausencia de correlaciones de grado, no depende de k.

A.2.2 Asortatividad

Similarmente a como se definieron pesos para las aristas, se pueden definir distintas características para los vértices dependiendo lo que los mismos representen. Por ejemplo, para una red social estas podrían ser: etnia, género,nivel socio económico . Dada esta caracterización existe la posibilidad, de que los vértices tiendan a unirse en mayor parte con vértices con valores de características similares , en tal caso se dice que el grafo es *asortativo*. Si por el contrario los vértices tienden a tener aristas con vértices de características muy disimiles el grafo presenta *disasortatividad*. Esto puede tener grandes efectos en la estructura del grafo, ya que al haber asortatividad fuerte, se pueden generar grupos altamente conectados con valores similares.

Se puede definir una cantidad e_{xy}, que es la fracción de aristas que unen nodos con valores de vértice x, y para alguna variable escalar de interés.Usando la matriz e_{xy} se define la medida de asortatividad. Primero se nota que e_{xy} satisface

$$\sum_{xy} e_{xy} = 1 \;\; \sum_{y} e_{xy} = a_x \;\; , \sum_{x} e_{xy} = b_y \qquad \textbf{(A.5)}$$

donde a_x, b_y son , respectivamente , la fracción de aristas que empiezan y terminan en vértices con valores x e y. (si el grafo es no dirigido y monopartito, $a_x = b_x$.) Entonces si no hay mezcla asortativa $e_{xy} = a_x b_y$.Si hay mezcla asortativa, se la puede obtener calculando la correlación de Pearson, por tanto:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}, \qquad \textbf{(A.6)}$$

donde σ_a y σ_b son las desviaciones estándar a_x y b_y. Este coeficiente de asortatividad se conoce como r de Newman, y sera el utilizado en el resto del trabajo. El valor de r asume valores en $-1 \leq r \leq 1$, con $r = 1$ indicando perfecta asortatividad y $r = -1$ indicando perfecta *disasortatividad*.

Similarmente a como se definió el coeficiente de asortatividad para una característica de los vértices general, se define la asortatividad de grado para un grafo como

$$r = \frac{\sum_{jk} jk(e_{jk} - a_j b_k)}{\sigma_a^2 \sigma_b^2} \qquad \textbf{(A.7)}$$

donde j y k son los valores del *grado en exceso* de los vértices y $a_k\ b_j$ son , respectivamente , la fracción de aristas que empiezan y terminan en vértices con valores k y j. El *grado en exceso*, es el número de todas las aristas salientes o entrantes en un vértice, descontando la arista por la que se llegó o se salió de dicho vértice. Las redes que presentan asortatividad de grado en el mundo real son las redes de coautoría, gente que es activa en la publicación de trabajos académicos tiende a trabajar con gente de las mismas tendencias, mientras que un ejemplo de disortatividad son las redes metabólicas [41].

A.2.3. Coeficiente de agrupamiento

Una medida topológica fundamental surge de cuantificar cuán conectado esta el entorno de un nodo se encuentra conectado. La medida más básica para este fin es el coeficiente de agrupamiento local ci 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

$$c_i = \frac{2}{k_i(k_i - 1)))} \sum_{j,m \in Neigh(i)} a_{ij} a_{jm} a_{mi} \qquad \textbf{(A.8)}$$

Equivalentemente, este observable topológico se puede interpretar como una relación entre el número de triángulos que conforman el nodo i, y el número total de posibles triángulos que podrían producirse entre éste y sus primeros vecinos $(k_i(k_i - 1)/2)$**.**

Esta medida da una idea intuitiva de cuán 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 través 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.Alternativamente el coeficiente de agrupamiento global es

$$C = \frac{\text{triadas cerradas}}{\text{total de triadas}}. \quad \textbf{(A.9)}$$

En el caso de redes pesadas, una generalización posible para el coeficientede agrupamiento de un nodo i, fue propuesta en [42]

$$c_i^w = \frac{1}{s_i(k_i-1)))} \sum_{j,m \in Neigh(i)} \frac{w_{ij}+w_{im}}{2} a_{ij} a_{jm} a_{mi} \quad \textbf{(A.10)}$$

es decir, que cada triángulo se contabiliza a menos de un factor de peso dado por el promedio de los arcos del mismo que incluyan al nodo i. Notar que esta definición se reduce al caso de redes no pesadas cuando los pesos son todos uniformes. Ademés, 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.

A.2.4. Centralidad

Una gran cantidad de trabajos han sido presentados en el estudio del concepto de centralidad que surge al hacer la pregunta “¿Qué vértices son importantes en la red ?”. Existen por tanto distintas definiciones de la centralidad en grafos. Cada definición usa una heurística distinta, y proviene de diferentes nociones de cual es un vértice importante en el grafo. En particular aquí consignaremos dos métodos a saber: el de *hub y authority score* de Kleinberg y la *centralidad de autovector.*

La centralidad de eigenvector, le otorga una puntuación de centralidad a un vértice proporcional a la de sus primeros vecinos. Los vértices de máxima puntuación

serán entonces aquellos que tengan vecinos de alta puntuación o que simplemente tengan muchos vecinos.Bonacich[43] demuestra que al suponer puntuaciones iniciales $\mathrm{x}(0)$ sobre la red y propagarlas iterativamente t veces según

$$\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0) \qquad \textbf{(A.11)}$$

para t suficientemente grande las puntuaciones tienden al valor de equilibrio

$$x_i = \sum_{ij} \kappa_1^{-1} A_{ij} xj \qquad \textbf{(A.12)}$$

con κ_1 el autovalor mayor de la matriz $\mathbf{A}$.

Las heurística detrás de las centralidades de Kleinberg[44] es la siguiente, un vértice será importante en el sentido de autoridad, si tiene aristas entrantes a muchos vértices importantes en el sentido de hub. Y un hub será importante si tiene muchas aristas salientes hacia vértices que son importantes como autoridades. Así el algoritmo para su cálculo queda dado por dos difusiones de puntuación con parametros α y β

$$x = \alpha \mathbf{A} y, \; y = \beta \mathbf{A}^{\mathbf{T}} x. \qquad \textbf{(A.13)}$$

El resultado final es independiente de la elección de coeficiente.

Lo importante de este segundo método para definir centralidad es que carece del problema del primero, en el que la puntuación de vértices fuera de componentes fuertemente conexos o solo con aristas salientes es nula. En particular un vértice con numerosas aristas salientes tendrá una puntuación de hub alta, y puede inducir una puntuación de autoridad alta en los vértices a los que apunta.

A.3 Estructura modular y particiones

A.3.1 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 [45]. 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 proteı´nas como la que presentaremos en el siguiente capítulo estos grupos modulares pueden representar y/o correlacionar con grupos funcionales o algún otro conjunto de interés.

En general, dado un grafo $G(N, E)$, una comunidad o módulo puede pensarse como un subgrafo 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. Més aún, la definición de módulo usualmente depende del sistema y la aplicación específica que se tenga en mente[45] .

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 este trabajo, 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 cuán buena es una partición dada. La función de calidad más ampliamente utilizada es la *modularidad* Q de Newman y Girvan [46]. 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 extraído 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(N, E) la modularidad Q queda definida según

$$Q = \frac{1}{2m} \sum_{j,m \in N} (A_{ij} - P_{ij}) \delta(C_i, C_j) \quad \textbf{(A.14)}$$

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 están conectados en el modelo nulo elegido. El módulo al cual pertenece el nodo i se denota mediante C_i y se tiene $\delta(C_i, C_j)$sólo si los nodos i, j pertenecen al mismo modulo (caso contrario $\delta(C_i, C_j) = 0$).

El modelo nulo más usual para el cálculo de modularidad es el *modelo configuracional* [47,48].

Este tipo de grafo aleatorio preserva el número total de aristas de cada nodo, y por consiguiente 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 $G(N, E)$ 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. Con esto, la forma más usual de calcular la modularidad de una partición resulta

$$Q = \frac{1}{2m} \sum_{j,m \in N} (A_{ij} - \frac{K_i k_j}{2m}) \delta(C_i, C_j) \quad \textbf{(A.15)}$$

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

$$Q = \frac{1}{2W} \sum_{j,m \in N} (A_{ij} - \frac{s_i s_j}{2m}) \delta(C_i, C_j) \quad \textbf{(A.16)}$$

esta es la forma más general que adoptaremos en esta tesis para calcular la modularidad.

A.3.2 Comparación entre particiones

De la misma manera que existen diversos cuantificadores para evaluar la calidad de una partición en cluster,s también existen distintos coeficientes que se pueden calcular para evaluar el parecido de dos particiones distintas en comunidades. En particular en este trabajo se utilizan dos tales medidas, información mutua normalizada (NMI) y Rand.

NMI es una medida que proviene de la teoría de la información, y parte de la noción de entropía de una partición C dado por

$$H(C) = -\sum_{i=1}^{k} P(i) log_2(P(i)) \quad \textbf{(A.17)}$$

con C_i las comunidad i de la partición c y $P(i)$ la probabilidad de encontrar un vértice en la comunidad C_i, a saber $\frac{|C_i|}{N}$.

Sea una segunda partición C' la información mutua de ambas particiones queda dada por

$$I(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i,j) log_2 \frac{P(i,j)}{P(i)P(j)} \quad \textbf{(A.18)}$$

con P(i,j) la probabilidad de encontrar un vértice en la comunidad i de la partición C y también la comunidad j de C'. Normalizando por la entropía queda definida entonces

$$NMI(C, C') = \frac{I(C,C')}{\sqrt{H(C)H(C')}} \quad \textbf{(A.19)}$$

y asume valores entre 0 y 1.

A diferencia de NMI que compara el contenido de nodos de cada partición, el índice de Rand compara los pares de nodos que se preservan en ambas particiones.

$$R = \frac{a+b}{\binom{N}{2}} \quad \textbf{(A.20)}$$

con a el número de pares de nodos que se encuentran juntos en clusters C y C' y b el número de pares de nodos que se encuentran separados.

A.3.3. 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 [46], 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 A.16. Por otro lado se consideró el algoritmo *Infomap [49]* que hace uso de criterios de optimización completamente diferentes, basados en teoría de información. En este algoritmo los módulos quedan definidosde 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 los niveles es necesario que la partición refleje los patrones de flujo dentro de la red, de manera que los diferentes módulos se corresponden 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 $P = P1, P2...Ps$ 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 $(P_i, P_j, i \neq j)$ y otra asociada a los movimientos que ocurren dentro de cada uno de los módulosP_i. El algoritmo infomap cuantifica este hecho mediante la función de costo descripta según

$$L(P) = q_{inter} H(Q) = \sum_i^s p_{intra}^i H(P^i) \qquad \textbf{(A.21)}$$

donde q_{inter}es la probabilidad de pasar de uno a otro módulo en la caminata, $H(Q)$es a entropía de movimientos entre módulos, p_{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 [49] .

Bibliografía

1. Molecular Biology of the Cell. 2002.

2. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery. Pharmacol Ther. 2013;138: 333–408.

3. Watts DJ. Networks, Dynamics, and the Small-World Phenomenon. Am J Sociol. 1999;105: 493–527.

4. White D, Johansen U. Network Analysis and Ethnographic Problems: Process Models of a Turkish Nomad Clan. Lexington Books; 2005.

5. Robertson SA, Renslo AR. Drug discovery for neglected tropical diseases at the Sandler Center. Future Med Chem. 2011;3: 1279–1288.

6. Parkkinen JA, Kaski S. Probabilistic drug connectivity mapping. BMC Bioinformatics. 2014;15: 113.

7. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol. 2013;9: 662.

8. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, et al. Drug target prediction and repositioning using an integrated network-based approach. PLoS One. 2013;8: e60618.

9. Cloete TT, North-West University (south Africa). Eflornithine Derivatives for Enhanced Oral Bioavailability in the Treatment of Human African Trypanosomiasis. 2009.

10. Burri C, Brun R. Eflornithine for the treatment of human African trypanosomiasis. Parasitol Res. 2003;90 Supp 1: S49–52.

11. Magarinos MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, et al. TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res. 2011;40: D1118–D1127.

12. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, et al. Thousands of chemical starting points for antimalarial lead identification. Nature. 2010;465: 305–310.

13. Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, Berriman M, et al. Identification of Attractive Drug Targets in Neglected-Disease Pathogens Using an In Silico Approach. PLoS Negl Trop Dis. 2010;4: e804.

14. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, et al. Chemical genetics of Plasmodium falciparum. Nature. 2010;465: 311–315.

15. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. PLoS Comput Biol. 2012;8: e1002503.

16. Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug–target interaction prediction through

domain-tuned network-based inference. Bioinformatics. 2013;29: 2004–2008.

17. Lü L, Linyuan L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, et al. Recommender systems. Phys Rep. 2012;519: 1–49.

18. Gillis J, Pavlidis P. The impact of multifunctional genes on “guilt by association” analysis. PLoS One. 2011;6: e17258.

19. Gillis J, Pavlidis P. The role of indirect connections in gene networks in predicting function. Bioinformatics. 2011;27: 1860–1866.

20. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making. 1989;9: 190–195.

21. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. 2000;407: 651–654.

22. Lambiotte R, Ausloos M. Collaborative Tagging as a Tripartite Network. Lecture Notes in Computer Science. 2006. pp. 1114–1117.

23. Hotho A, Jäschke R, Schmitz C, Stumme G. Information Retrieval in Folksonomies: Search and Ranking. Lecture Notes in Computer Science. 2006. pp. 411–426.

24. Zhou T, Ren J, Medo M, Zhang Y-C. Bipartite network projection and personal recommendation. Phys Rev E Stat Nonlin Soft Matter Phys. 2007;76: 046115.

25. Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN. Statistically Validated Networks in Bipartite Complex Systems. PLoS One. Public Library of Science; 2011;6: e17994.

26. Boccaletti S, Bianconi G, Criado R, del Genio CI, Gómez-Gardeñes J, Romance M, et al. The structure and dynamics of multilayer networks. Phys Rep. 2014;544: 1–122.

27. Rogers DJ, Tanimoto TT. A Computer Program for Classifying Plants. Science. 1960;132: 1115–1118.

28. Kruger FA, Rostom R, Overington JP. Mapping small molecule binding data to structural domains. BMC Bioinformatics. 2012;13 Suppl 17: S11.

29. Flower DR. On the Properties of Bit String-Based Measures of Chemical Similarity. J Chem Inf Comput Sci. 1998;38: 379–386.

30. Snarey M, Terrett NK, Willett P, Wilton DJ. Comparison of algorithms for dissimilarity-based compound selection. J Mol Graph Model. 1997;15: 372–385.

31. Dixon SL, Koehler RT. The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. J Med Chem. 1999;42: 2887–2900.

32. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24: i232–40.

33. Berenstein AJ, Magariños MP, Chernomoretz A, Agüero F. A Multilayer Network Approach for Guiding Drug Repositioning in Neglected Diseases. PLoS Negl Trop Dis. 2016;10: e0004300.

34. Peters GJ, Backus HHJ, Freemantle S, van Triest B, Codacci-Pisanelli G, van der

Wilt CL, et al. Induction of thymidylate synthase as a 5-fluorouracil resistance mechanism. Biochim Biophys Acta. 2002;1587: 194–205.

35. Filatov D, Ingemarson R, Gräslund A, Thelander L. The role of herpes simplex virus ribonucleotide reductase small subunit carboxyl terminus in subunit interaction and formation of iron-tyrosyl center structure. J Biol Chem. 1992;267: 15816–15822.

36. Serrano MA, Boguñá M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci U S A. 2009;106: 6483–6488.

37. Bruncko M, McClellan WJ, Wendt MD, Sauer DR, Geyer A, Dalton CR, et al. Naphthamidine urokinase plasminogen activator inhibitors with improved pharmacokinetic properties. Bioorg Med Chem Lett. 2005;15: 93–98.

38. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug—target network. Nat Biotechnol. 2007;25: 1119–1126.

39. Wasserman S, Faust K. Social Network Analysis in the Social and Behavioral Sciences. Social Network Analysis. pp. 3–27.

40. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D. Complex networks: Structure and dynamics. Phys Rep. 2006;424: 175–308.

41. Newman MEJ. The Structure and Function of Complex Networks. SIAM Rev. 2003;45: 167–256.

42. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences. 2004;101: 3747–3752.

43. Bonacich P. Power and Centrality: A Family of Measures. Am J Sociol. 1987;92: 1170–1182.

44. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM. 1999;46: 604–632.

45. Fortunato S. Community detection in graphs. Phys Rep. 2010;486: 75–174.

46. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004;69. doi:10.1103/physreve.69.026113

47. Molloy M, Reed B. The Size of the Giant Component of a Random Graph with a Given Degree Sequence. Comb Probab Comput. 1998;7: 295–305.

48. Molloy M, Reed B. A critical point for random graphs with a given degree sequence. Random Struct Algorithms. 1995;6: 161–180.

49. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences. 2008;105: 1118–1123.