



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

Redes complejas aplicadas a la caracterización, integración y priorización de datos bioquímicos

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

Lic. Bruno Kaufman

Director de tesis: Dr. Ariel Chernomoretz
Consejero de estudios: Dr. Pablo Balenzuela

Lugar de trabajo: Laboratorio de biología de sistemas integrativa. Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA), Fundación Instituto Leloir (FIL)

Defendida en Ciudad Universitaria de Buenos Aires, 15 de noviembre de 2024

Resumen

En el campo de la biología computacional se utilizan cantidades enormes de información genética, química, y farmacológica para responder preguntas de interés en estos campos. En esta tesis se investigan conceptos en redes complejas que habilitan un análisis estructurado de estos datos.

Como ejemplo de aplicación genómica, primero se exploran redes resultantes de un análisis mecanico-estadístico de sitios de splicing 5'. A través de este análisis se infieren conclusiones acerca del funcionamiento interno del splicing, fenómeno mediante el cual se procesa la información genética almacenada en la molécula de ADN, y se identifican regiones que serán tenidas en cuenta para la construcción de proteínas (exones), en contraste a las secciones que serán descartadas (intrones).

Luego se emplean las redes para la caracterización y consolidación de datos de interacción entre fármacos y proteínas, además de sus características químicas y funcionales. Se establece un procedimiento para la utilización en conjunto de datos de similaridad química entre moléculas de fármacos, datos de interacciones conocidas entre fármacos y proteínas, y relaciones entre proteínas basadas en la presencia de dominios funcionales (PFAM) y linaje evolutivo en común (ortología).

Finalmente se utiliza un enfoque de difusión en red, generalizando ecuaciones de difusión a espacios no-euclidianos, para el problema del aprendizaje semi-supervisado sobre estos datos, y así proponer nuevos blancos terapéuticos para fármacos conocidos. Se explora el rol del operador Laplaciano en red, responsable del traspaso del concepto de difusión desde la física a las redes complejas. A través de estos análisis se establece la importancia del formalismo de redes complejas en el uso de datos diversos, en especial cuando se los busca utilizar en conjunto.

Palabras clave: redes complejas, bioinformática, priorización, difusión, proteínas, fármacos

Abstract

Complex networks applied to the characterization, integration and prioritization of biochemical data

The field of computational biology makes use of enormous quantities of genetic, chemical and pharmacological information in order to find answers to questions of interest in these fields. This thesis will investigate concepts in complex network theory relevant to a structured analysis of said data.

As an example of genomic application, we first explore networks resulting from a statistical mechanics-based analysis of 5' splice sites. Through this analysis, conclusions are inferred regarding the inner workings of the splicing process, by which the genetic information stored in DNA is processed, and sections relevant to the construction of proteins (exons) are identified and separated from those to be ignored (introns).

Afterwards, a network framework is used in order to characterize and consolidate data regarding the interaction between drugs and proteins, as well as their chemical and functional qualities. A procedure is established for the simultaneous use of this information, including chemical similarity data between drug molecules, known interaction data between drugs and proteins, and protein-protein relationships based on the presence of functional domains (PFAM) and evolutionary lineages in common (orthology).

Finally, a network diffusion approach is used, generalizing diffusion equations to non-Euclidean space, in order to perform semi-supervised learning on this data, and thus propose new therapeutic targets for known drugs. The thesis hence explores the role of the network Laplace operator, responsible for the use of the physical concept of diffusion within a complex networks framework. Through these analyses, this thesis establishes the importance of complex networks theory in employing diverse data, especially when the goal is to make use of them in tandem.

Keywords: complex networks, bioinformatics, prioritization, diffusion, proteins, drugs

Índice general

Resumen	II
Abstract	III
Agradecimientos	VII
Introducción	1
Fundamentos de bioquímica	2
Las proteínas	3
Dogma central de la biología molecular	4
Fundamentos de teoría de redes	8
Objetivos	10
Estadística en sitios de splicing	13
Metodología	15
Características de los datos	18
Características del ajuste	19
Convergencia y bases de atracción	21
Señales conservadas	23
Divergencia de las señales de interacción según diferenciación de especies	26
Modelos consolidados	30
Discusión	31
Conclusión	34
Consolidación de datos bioquímicos heterogéneos	36
Metodología	37
Medidas de similaridad estructural química	37
Consistencia entre medidas de similaridad química	41
Interacciones droga-target	44
Similaridad entre proteínas	46
Interacción específica ligando-proteína	47
Consistencia entre interacciones bioquímicas y similaridades químicas	50

Drogabilidad de grupos PFAM	51
Entropía inter-especie y flujo de información	61
Discusión	63
Conclusión	65
Priorización de fármacos	67
Metodología	68
Continuidad y circulación en un espacio continuo	69
Operador gradiente en grafos	70
Operador divergencia en grafos	72
Matriz Laplaciana en grafo	73
Propiedades del Laplaciano en red	75
Difusión en red	77
Difusión sobre redes sintéticas	79
Priorización de asociaciones droga-PFAM	82
Red híbrida para datos droga-target	87
Discusión	93
Conclusión	95
Conclusiones	96

Agradecimientos

El trayecto de este doctorado, y de esta tesis, se construyó no solo con estudios y persistencia, sino también con el apoyo de quienes me rodearon.

Del laboratorio agradezco primero a Ariel Chernomoretz. No solo por incluirme en este equipo, sino por dar (incontables veces) el ejemplo como investigador en biología de sistemas, ejecutando con sumo criterio el balance delicado entre las soluciones creativas a los problemas complejos, y su viabilidad y practicidad de implementación.

Agradezco a Maximiliano S. Beckel por ser el primero en explicarme el dogma central, entre tantos conceptos biológicos fundamentales al desarrollo de este trabajo, todos narrados con una capacidad didáctica sin paralelo. A Andrés Rabinovich le agradezco haber cumplido el mismo rol en estadística: es un gran especialista, y existen pocos expertos con su rigor y diligencia. Además incluyo en este agradecimiento a Ariel Bernardino y Tomás Vega Waichman por haber formado parte de esta comunidad, así como a todos sus y nuestros antecesores en el laboratorio, sobre cuyo trabajo se construye y sigue construyendo el presente.

Además se agradece al laboratorio de bioinformática estructural, por también formar parte de esta comunidad tan alentadora (cualidad fundamental para poder realizar un doctorado), y a la Fundación Instituto Leloir por proporcionar la infraestructura necesaria como lugar de trabajo.

Fuera del laboratorio, mis agradecimientos podrían alargar este texto por dos páginas más. A cada persona de la multitud de quien tomé prestada la voluntad, espero algún día poder devolver el favor. Hasta aquel momento, en son de vale daré las gracias, con esta tesis de testigo.

Capítulo 1

Introducción

Monjes, nosotros quienes observamos el entero y no solo las partes, sepan que también somos sistemas de interdependencia, de emociones, sensaciones, pensamientos y consciencia, todos interconectados.

Investigando de esta manera, nos percatamos de que no hay “yo” ni “mío” en ninguna parte individual, de la misma forma en que el sonido no le pertenece a ninguna parte individual del laúd.

—*Samyutta Nikaya*, del *Tripitaka*

Los campos de la biología, bioquímica y farmacología contemplan cantidades de datos cada vez más masivas. Al existir firmemente dentro del dominio del “big data”, las necesidades de análisis son satisfechas por metodologías de estadística y aprendizaje automatizado, y este área de la interdisciplina se denomina bioinformática. La evolución de este campo es constante y acelerada, incluyendo desarrollos en aprendizaje profundo, simulaciones físicas o motivadas en la física, y aprendizaje automatizado semi-supervisado.

En este trabajo consideraremos varios problemas en el análisis de datos masivos bioinformáticos, todos con un elemento en común: la interacción entre pares como unidad mínima de descripción. Este es el núcleo del enfoque de las redes complejas, que desde estas interacciones simples logra construir estructuras intrincadas que delatan un orden emergente a diversas escalas. Este marco teórico nos provee un abanico enorme de herramientas que utilizaremos para darle forma a estos datos y examinar los conceptos que rigen la dinámica de sus sistemas correspondientes.

Existe una gran variedad de formas que pueden tomar estos conceptos y las metodologías asociadas. Pueden mostrarnos relaciones entre pequeños grupos de elementos de una red, comunidades más grandes a distintas granularidades, o incluso características globales de la red completa: ejemplos de las varias escalas con las cuales se puede analizar una red. Pueden utilizarse para el aprendizaje automatizado sobre redes com-

plejas, aprovechando la información embebida en la topología de red. Pueden ofrecer una base de comparación entre nociones de similaridad, inspeccionando cómo se relaciona la similaridad entre dos nodos de una red, con la similaridad inferida para las mismas entidades en el contexto de una red diferente. Este último sería el caso de un dato complementario en dos distintas áreas de conocimiento.

Bajo algunos de estos abordajes encaramos el estudio de diversas problemáticas de interés biomédico donde son relevantes múltiples escalas de organización, que van desde el nivel molecular hasta la escala del organismo completo. Antes de comenzar con estas aplicaciones, debemos primero sentar las bases biológicas de las ideas y los datos que exploraremos, y luego dar una introducción a teoría de redes, en lo relevante a este trabajo. Esto haremos en las próximas secciones, para luego resumir nuestros objetivos y finalmente avanzar hacia los tópicos principales de la tesis.

1.1. Fundamentos de bioquímica

La célula es la unidad básica estructural de todo organismo viviente. Toda célula se encuentra limitada por una membrana, y un interior que consiste de un medio llamado citoplasma. Dentro del citoplasma se sitúan las organelas que desempeñan los diversos roles necesarios para vivir, además de macromoléculas individuales como las proteínas, distintos metabolitos, e información genética en forma de ácido desoxiribonucleico (ADN) y ácido ribonucleico (ARN mensajero, de transferencia, etc.).

Existen dos tipos de célula en la naturaleza: las llamadas procariotas y las eucariotas. La mayor diferencia entre ellas es que las primeras carecen de núcleo celular, mientras que los segundos poseen dicha estructura como se puede observar en la figura 1.1. Esta separación se ve acompañada de una serie de cualidades: las células eucariotas son de mayor tamaño, tienen organelas específicas para realizar funciones especializadas, y típicamente poseen genomas más grandes y complejos. Este último punto será de especial importancia en el capítulo 2, donde se analiza la información genética en diversas especies eucariotas, incluyendo animales, plantas y hongos.

Tanto el núcleo como los ribosomas adyacentes, visibles en la figura 1.1, desempeñan roles fundamentales en el uso de la información genética y la fabricación de nuevo material celular. Este proceso será discutido en detalle en la sección 1.1.2.

En la visualización de la célula se puede notar una variedad de organelas, unidades especializadas en diversas tareas. Tanto ellas como una buena parte de los metabolitos celulares están compuestas de un bloque básico de construcción llamado *proteína*. Ellas son manufacturadas por la maquinaria interna de la célula, y cumplen todos los roles imaginables, tanto metabólicos, o estructurales, conformando membranas (ver la figura 1.4), organelas u otros elementos. Su importancia y versatilidad es enorme, y en la sección siguiente explicaremos algunos puntos básicos de su comportamiento.

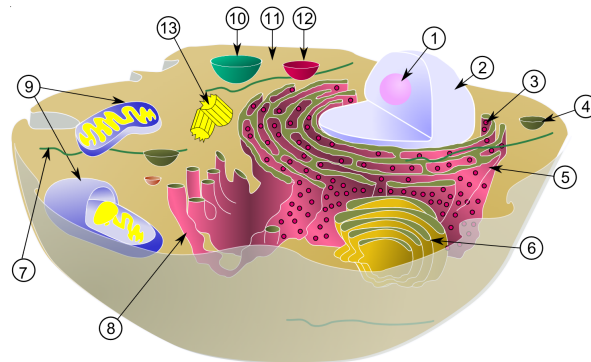


Figura 1.1: Diagrama de una célula animal, de los tres reinos más grandes de los eucariotas. Se nota un orden más estructurado que el de la célula procariota. Entre las diferencias más destacables se incluye el núcleo (2), donde está contenida la información genética, el nucléolo (1), donde se transcribe el ARN ribosomal, y el ribosoma (3) donde se traduce esa información genética en las proteínas que conforman la maquinaria celular. Entre otras organelas se puede notar la mitocondria (9), con su propia información genética. Esto solo es posible debido al tamaño mucho mayor de los eucariotas respecto a los procariotas. Imagen utilizada bajo dominio público.

1.1.1. Las proteínas

Tanto la estructura de las células como la mayoría de su maquinaria interna se compone de proteínas, macromoléculas complejas que a través de su forma y composición cobran funcionalidad. Se componen de una secuencia de *aminoácidos*, vinculados entre sí por *enlaces peptídicos* como vemos en la figura 1.2. Con cientos o miles de aminoácidos encadenados, y a través de las fuerzas fisicoquímicas entre ellos, se construye una estructura tridimensional compleja: la proteína.

Una proteína puede interactuar con otras proteínas o moléculas pequeñas mediante fuerzas fisicoquímicas (ver figura 1.3). En el segundo caso, a las moléculas pequeñas se las conoce como *ligandos*. Estas interacciones son el soporte molecular sobre el que se montan las estructuras de escala celular. En la figura 1.4 vemos una sección de la membrana plasmática de una célula eucariota, compuesta mayoritariamente por una bicapa de *fosfolípidos*, moléculas que se organizan en dicha estructura por la polaridad distinta entre sus dos extremos. Embebidas en esta membrana se pueden reconocer proteínas desempeñando distintos roles.

Como mencionamos, las proteínas pueden interactuar con pequeñas moléculas llamadas ligandos, como vemos en la figura 1.3. El funcionamiento habitual de muchas proteínas depende de esta interacción. Esto se puede usar para intencionalmente habilitar moléculas específicas que interactúen con sitios protéicos de interés. Más aún, pueden existir compuestos químicos que monopolicen estos sitios de forma duradera, con efectos fisiológicos notables a escala celular. De hecho, cuando esto se aprovecha con intención farmacológica, esta molécula se denomina una *droga* o *fármaco*, y la proteína identificada es una *diana terapéutica*. La búsqueda de drogas y dianas terapéuticas es

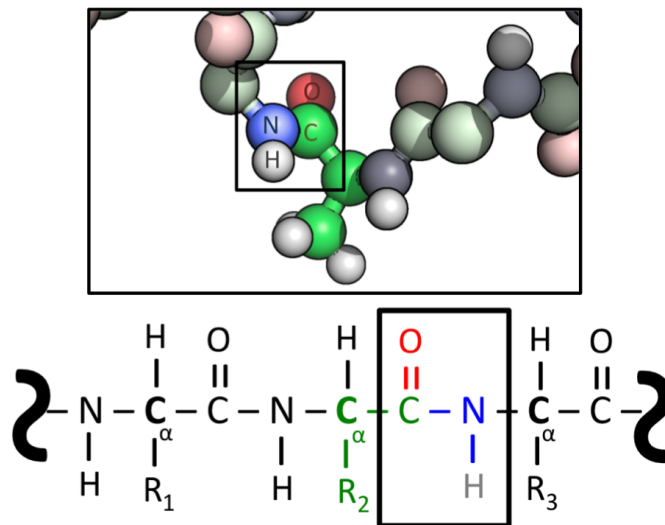


Figura 1.2: Esquema de una sección de una proteína. El enlace peptídico rodeado por un rectángulo negro vincula al aminoácido correspondiente al grupo R_2 con el R_3 : ambos estos aminoácidos tendrán propiedades fisicoquímicas distintas, y en conjunto por las fuerzas involucradas formarán la estructura tridimensional vista arriba. Esta cadena continúa por cientos o miles de aminoácidos, formando una macromolécula compleja en tres dimensiones: la proteína. Imagen utilizada bajo licencia CC BY-SA 3.0, agradeciendo la contribución original[1].

un trabajo enorme, y abarcaremos una pequeña parte de él en este trabajo de tesis.

1.1.2. Dogma central de la biología molecular

Como vimos, las proteínas son fundamentales para el funcionamiento de la célula, componiendo casi toda su maquinaria. Las mismas son manufacturadas por la misma célula, utilizando la información genética como un instructivo para su fabricación.

Este proceso de manufactura se puede describir, como primera aproximación, por las relaciones “generales” contempladas por el llamado *dogma central* de la biología molecular, ilustrado por la figura 1.5. Como primer paso, la información contenida en el ADN es transcrita a ARN mensajero en el núcleo de la célula. Esta transcripción no copia el genoma completo, sino una región específica asociada al gen que codifica a la proteína de interés.

Luego de la transcripción se procesa el transcripto inmaduro quitando partes de la secuencia y consolidando el resto en una forma madura. Esta eventualmente será exportada del núcleo y utilizada por los ribosomas del citoplasma para traducir a una secuencia de aminoácidos, como la vista en la figura 1.2. Esta cadena de procesos da como resultado una estructura aminoacídica particular en tres dimensiones, que conforma una proteína potencialmente funcional. Esta proteína ocupará su rol dentro de la célula, cumpliendo una parte de sus funciones metabólicas y/o su estructura.

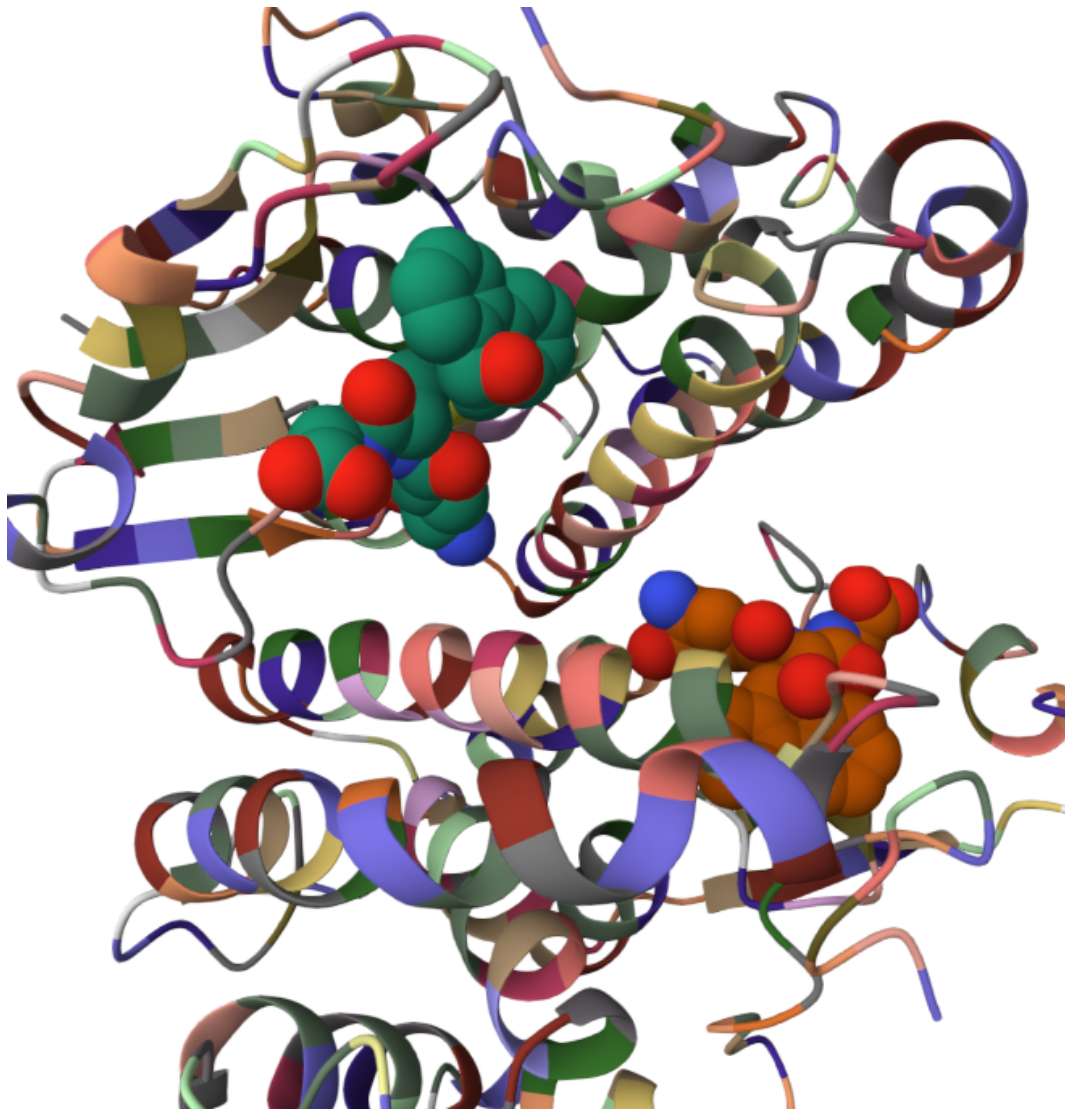


Figura 1.3: Estructura de una proteína (S-transferasa de glutationa) interactuando con ligandos por fuerzas fisicoquímicas. La proteína se ilustra con una estructura de bucles, donde cada color distinto representa un aminoácido distinto en la cadena, que por sus propiedades forman la estructura vista. Los ligandos (círculos de color representando átomos individuales de la molécula) interactúan con sitios específicos, también según las propiedades físicas de la proteína.

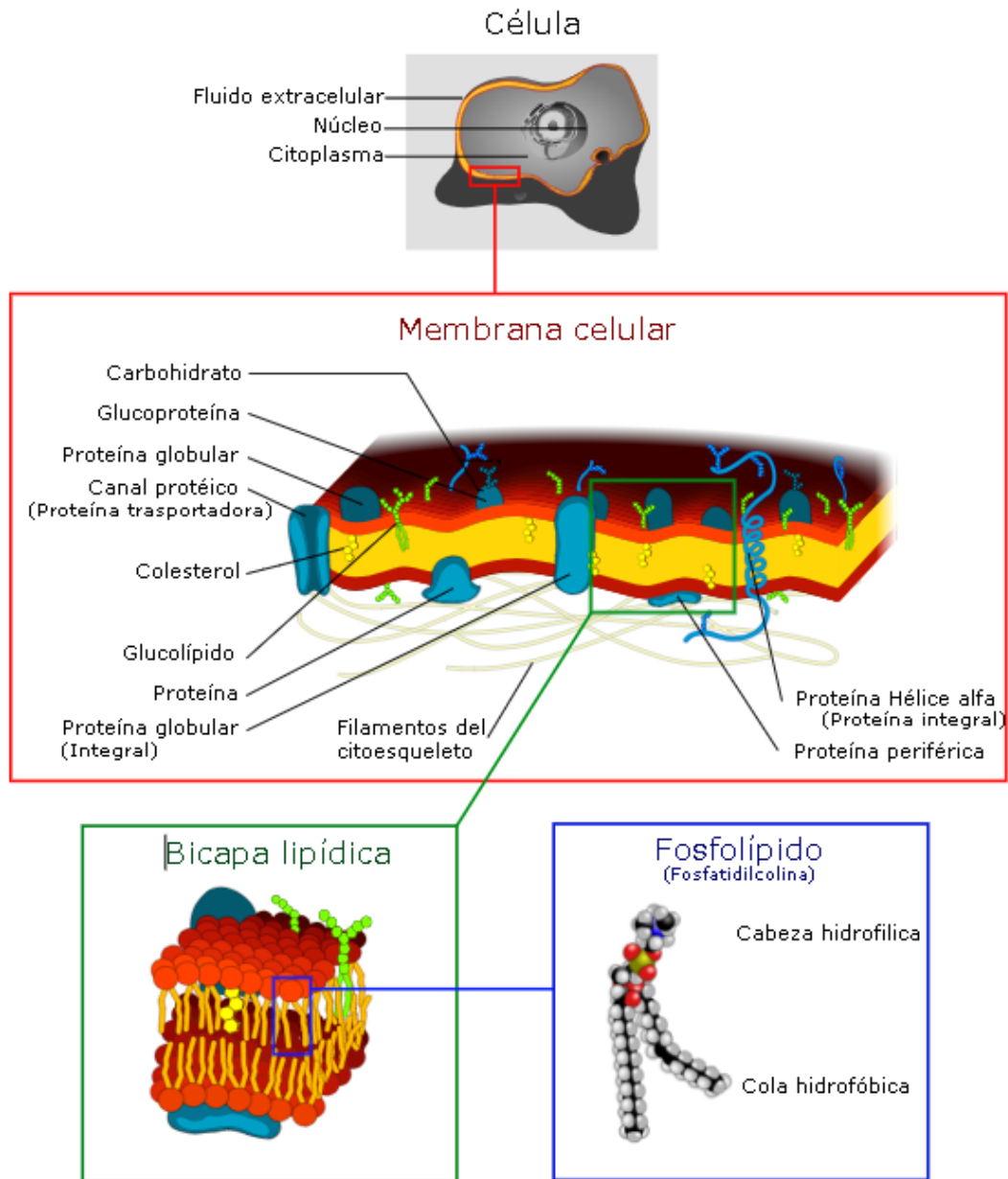


Figura 1.4: Membrana celular en un organismo eucariota, formada de una bicapa de fosfolípidos con proteínas incorporadas en su estructura para desempeñar distintas funciones. Se puede observar como, desde la escala molecular, se logran construir estructuras funcionales a escala celular. Imagen utilizada bajo licencia CC BY-SA 3.0, agradeciendo la contribución original[2].

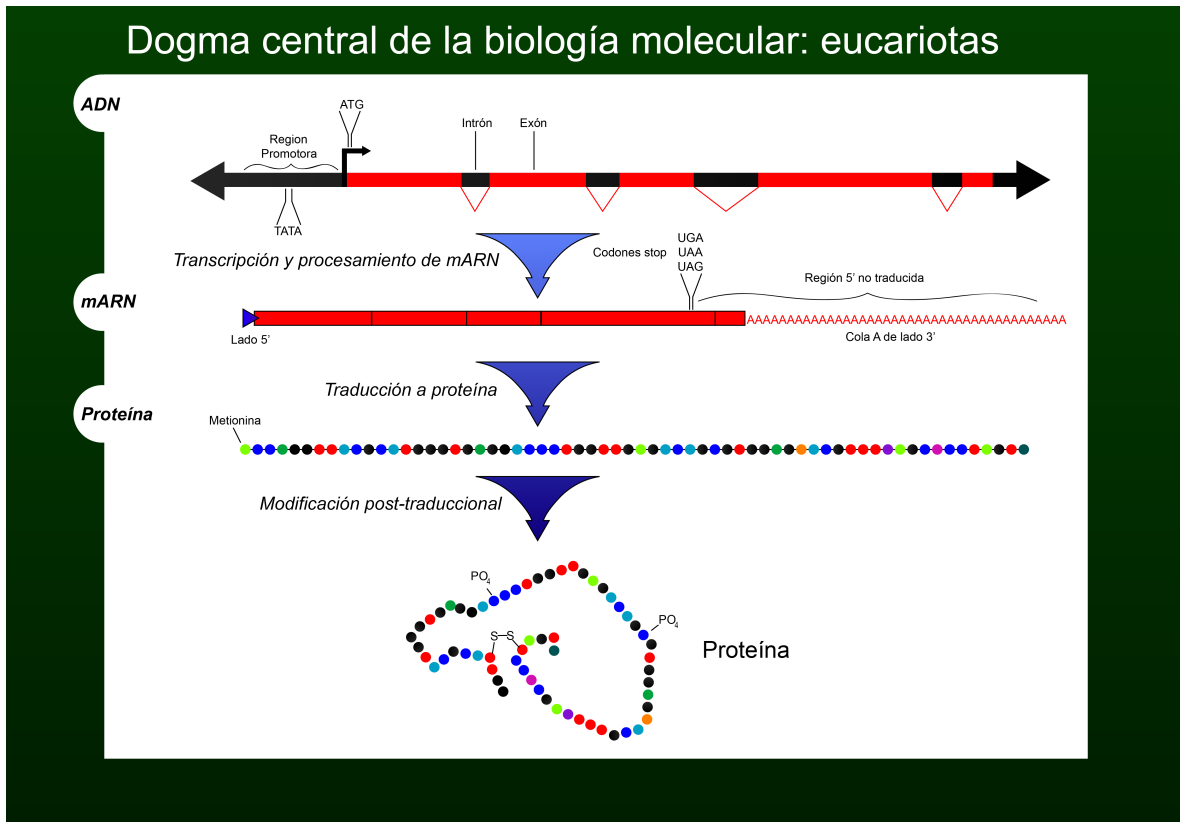


Figura 1.5: Ilustración esquemática de dos de las transferencias generales de información bajo el dogma central de la biología molecular: la transcripción de ADN a ARN mensajero, y la traducción de ARN mensajero a proteína. Imagen utilizada y modificada bajo licencia CC BY-SA 2.5, agradeciendo la contribución original[3].

La transcripción de ADN a mARN y el procesamiento del transcripto forman un paso complejo e importante en la maquinaria de fabricación de proteínas. En el centro de este proceso está el fenómeno del *splicing*, mediante el cual se extraen las regiones no codificantes de la información genética (intrones, ver figura 1.5). Este paso de maduración del ARN es llevado a cabo por una maquinaria protéica denominada *spliceosoma*, y constituye un paso fundamental para la identificación de la información genética que dará lugar a la fabricación de una proteína o a distintos procesos regulatorios. En el capítulo 2 exploraremos algunos aspectos de este fenómeno, relacionado con el reconocimiento de los sitios de quita de intrones, junto a nuestro análisis estadístico de sus patrones subyacentes.

Tanto para este análisis como para los capítulos 3 y 4 necesitaremos una base de teoría de redes. En la sección 1.2 cubriremos este requerimiento.

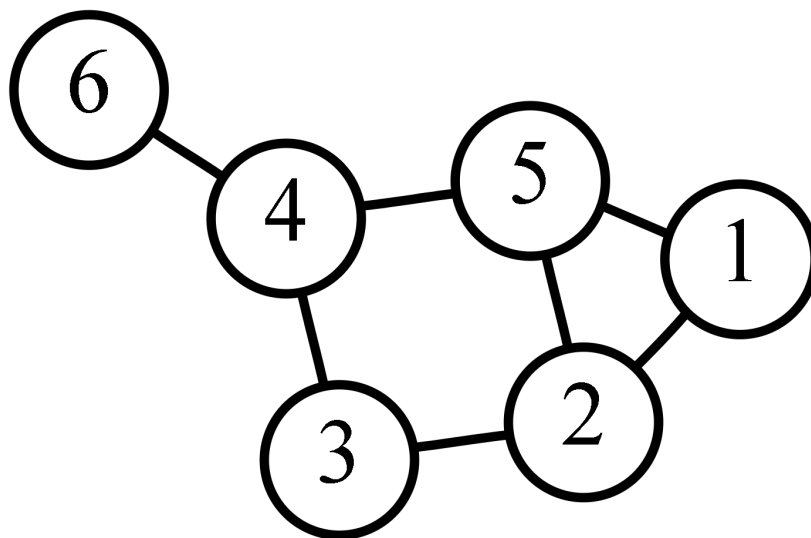


Figura 1.6: Un ejemplo de un grafo simple no dirigido. Tiene seis nodos, con sus números correspondientes, y entre ellos forman un componente totalmente conexo. Imagen utilizada bajo dominio público.

1.2. Fundamentos de teoría de redes

Las redes complejas se pueden entender desde el marco de teoría de grafos. Un *grafo*, término que usaremos de forma intercambiable con *red* es un par ordenado $G(V, E)$ donde V es el conjunto de *vértices* o *nodos*, y E es el conjunto de *aristas* o *vínculos* que conectan a estos vértices entre sí.

Como podemos ver en la figura 1.6, tan solo vinculando nodos de a pares surge una estructura más amplia de interconectividad. Si uno imagina a un caminante sobre los nodos, cruzando de uno a otro mediante los vínculos, entonces existe una gran variedad de recorridos para tomar, y entre todos ellos forman la topología de la red. Cuando el caminante podría lograr recorrer la red entera desde cualquier otro punto, como en el caso de la figura 1.6, se dice que la misma es *conexa*. En el caso contrario, la red se separa en componentes distintos. El más grande entre ellos se denominará el *componente gigante*.

Lo anterior aplica cuando los vínculos entre nodos son recíprocos. Sin embargo, en muchas situaciones, un nodo puede influir sobre otro con una relación no recíproca. Este es el caso de un vínculo dirigido, y da lugar a una red dirigida como la vista en la figura 1.7. Es posible, en este caso, que el caminante pueda recorrer la red al provenir de una dirección, pero no logre hacerlo a la vuelta. Se tiene en este caso dos criterios de conectividad: bajo el criterio *fuerte*, que considera la dirección de los vínculos, la red se considera desconexa, mientras que bajo el criterio *débil*, se considera conexa.

Cuando la red es dirigida, existen casos donde el caminante comience en un nodo y haga un recorrido donde regresa al mismo sitio. Estos caminos se denominan *ciclos*,

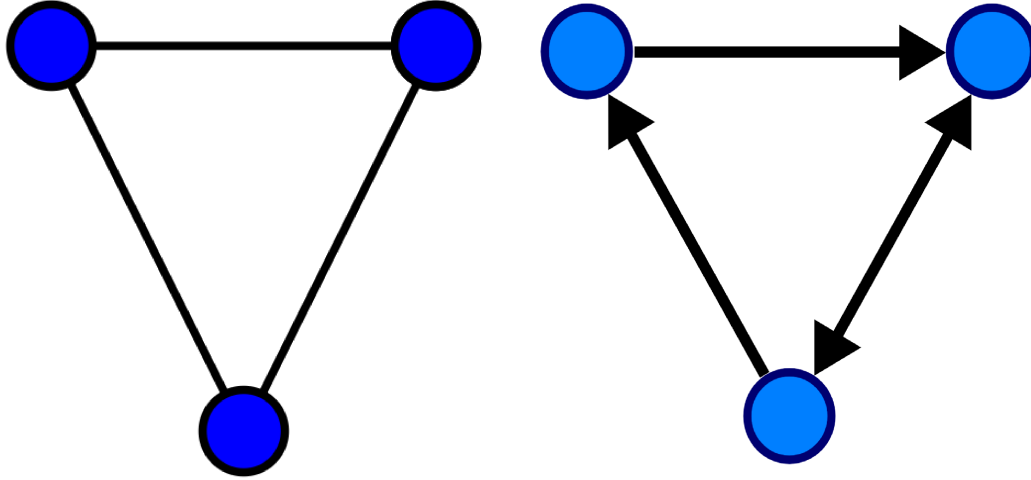


Figura 1.7: (Izquierda) grafo no dirigido, comparado con (derecha) grafo dirigido. En el caso de que los datos se adecúen mejor a una red dirigida, aumentará la complejidad del problema, y también la información para extraer. Imagen utilizada bajo dominio público.

y cuando no existe ninguno, la red se denomina *acíclica*. Es fácil razonar que, si una red es acíclica, de comenzar en un nodo no podrá regresar a él, por lo que la red es desconexa bajo el criterio dirigido.

Para expresar el grafo de forma matemática se recurre a la *matriz de adyacencia* $\overline{\overline{A}}$. El elemento A_{ij} representa que existe un vínculo directo que permite acceder al nodo i desde el j . En el caso más simple, estos índices toman un valor binario: uno o cero, indicando la presencia o ausencia de un arista. Sin embargo en casos más complejos, el vínculo puede tener un peso, indicando una mayor importancia, una mayor similaridad entre ambos nodos, o cualquier clase de significados atribuidos por el análisis que se desee realizar. En este caso se dice que es una red *pesada*.

Cada nodo puede ser caracterizado en base a su posición en la red, y su relación con sus vecinos y más allá. La más simple de estas se llama el *grado* k de un nodo. Este grado es el número de vecinos que posee. En el caso de un grafo dirigido, cada nodo tendrá dos grados distintos: un grado *interno* k_{in} y un grado *externo* k_{out} , midiendo respectivamente el número de vecinos que apuntan a él y el número de vecinos a quien apunta. El grado interno se puede pensar como la suma en filas de la matriz de adyacencia:

$$k_{in} = \sum_i^{nodos} A_{ij}$$

mientras que el externo se puede pensar como la suma en columna:

$$k_{out} = \sum_j^{nodos} A_{ij}$$

y, en el caso del grafo no dirigido, ambas son equivalentes al ser simétrica $\overline{\overline{A}}$. Cuando las entradas de A tienen peso, se generaliza el grado k_i a una fuerza s_i , la suma de los pesos que interactúan con un dado nodo (sea hacia afuera, hacia adentro, o de manera no dirigida).

Además de caracterizar propiedades individuales de cada nodo (escala microscópica), como su conectividad local a través del grado o su semejanza con vecinos por la transitividad, también es posible obtener información global (macroscópica) a partir de la estructura de interconectividades embebida en la red. Un ejemplo de esto es el diámetro: el camino más largo que existe, estableciendo una especie de longitud que tendrá la red de un extremo a la otra. Uno puede notar que, de ser desconexa la red, se tendrá que calcular un diámetro separado para cada componente distinto. Otras cantidades macroscópicas de interés pueden incluir el tamaño de la red (la cantidad de nodos), la cantidad de enlaces, y el número y tamaño de componentes conexas.

Finalmente existe una tercera escala para caracterizar la red: la mesoscópica. En base a cuan interconectados o separados están los distintos grupos de nodos se puede plantear la noción de que existen en estructuras, denominadas *comunidades*, definidas a partir de subconjuntos de nodos de mayor interconectividad interna. Aquí entra el dominio de la detección de comunidades, con sus cientos de metodologías posibles. Un ejemplo de una partición de comunidad se puede observar en la figura 1.8. Se cuenta con algoritmos de detección de comunidad más rápidos o más precisos, pero por lo general buscan, con ciertas limitaciones, optimizar la llamada *modularidad* de la partición de comunidades:

En capítulos posteriores veremos como, a partir de estas nociones simples, se construyen redes enormes y complejas. Con estas ideas básicas será suficiente para abarcar los conocimientos de esta tesis, explicando los detalles pertinentes y la metodología utilizada a cada paso.

1.3. Objetivos

El objetivo de este trabajo es utilizar la aproximación de redes complejas como herramienta para, desde las interacciones entre pares, estudiar diferentes sistemas biológicos de relevancia biomédica. A cada paso pondremos énfasis en la estructura y los conocimientos que emergen tan solo relacionando pares de elementos, sin que sea necesario hacer referencia explícita a sus características individuales, si las posee. Todo esto para, en cada capítulo, dejar un mensaje claro: el orden que interpretamos no le pertenece al elemento individual, sino a la interacción entre todos.

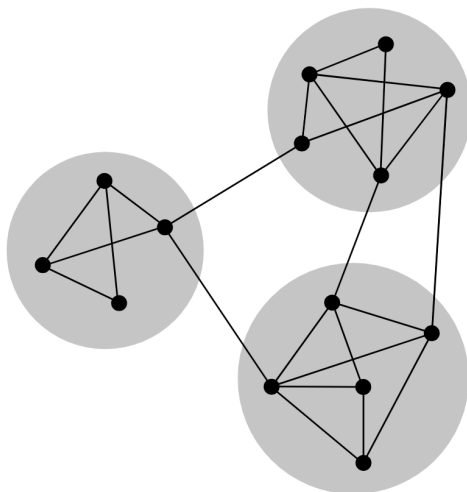


Figura 1.8: Una red partida con una estructura de tres comunidades. Se puede observar que la partición separa conjuntos que internamente se ven bien conectados, mientras que hacia las demás comunidades poseen escasas conexiones. Imagen utilizada bajo licencia CC BY-SA 3.0, agradeciendo la contribución original[4].

En el capítulo 2 estudiaremos patrones específicos de compensación en la composición de la secuencia utilizada por el spliceosoma para reconocer los sitios de splicing. Estas secuencias cortas, denominadas *donoras*, reflejan el balance entre la variabilidad nucleotídica natural y los requisitos básicos moleculares necesarios para asegurar la fidelidad del splicing. Es de interés caracterizar a este equilibrio mediante los patrones estadísticos presentes en distintas muestras de tales secuencias.

Empleamos un modelo de máxima entropía regularizado para minar interacciones de dos sitios no triviales en las secuencias donoras presentes en conjuntos de datos para treinta eucariotas distintos. Para cada especie analizada, identificamos conjuntos minimales de patrones de acoplamiento entre dos sitios, que fueron capaces de reproducir, a un dado nivel de regularización, las frecuencias observadas de uno y dos sitios en secuencias donoras. Al realizar un análisis comparativo y sistemático de los sitios 5', demostramos que la información del linaje evolutivo se podía rastrear a través de las probabilidades di-nucleótido inferidas por nuestro modelo. Pudimos identificar patrones de interacción de dos sitios para plantas, animales y hongos, y proponemos que las correspondientes a los primeros dos grupos hacen eco de las diferencias previamente reportadas entre ellos en lo pertinente a la regulación del proceso de splicing.

En el capítulo 3 utilizaremos el formalismo de redes complejas para crear y explorar un grafo de conocimiento quimiogenómico de alta relevancia para tareas de priorización de fármacos. A través de la integración de bases de datos farmacológicos curados, se consolidará y evaluará la complementariedad entre distintas áreas de conocimiento, y cómo aprovechar de mejor forma la diversidad de información.

El capítulo 4 tomará provecho de los datos previamente organizados para, a través de

un enfoque de aprendizaje semi-supervisado, diseñar un algoritmo para la predicción de nuevas interacciones entre drogas y blancos terapéuticos (“targets”). El método utilizado es uno de *difusión en red*: similar a la idea del caminante sobre la red, explorado en la sección 1.2, se puede difundir un campo, equiparable a la probabilidad de que un dado caminante termine en un dado nodo. Esta idea inspira toda una familia de metodologías que explicaremos en el mismo capítulo. Estos métodos aprovechan la analogía entre la difusión de información en redes complejas y la difusión en espacios físicos, mediada por el operador “Laplaciano” en red, un operador que exploraremos en detalle.

Capítulo 2

Estadística en sitios de splicing

La mayor parte de los genes eucariotas se componen de secciones alternantes de ADN (ácido desoxiribonucleico) conocidos como exones e intrones. Los intrones son quitados del transcrito naciente (la información que estaría siendo copiada) a través del proceso llamado splicing[5, 6, 7]. Esto deja a una serie de exones, regiones codificantes de la información genética, que serán concatenadas dando lugar al transcrito maduro. El splicing resulta de una serie de reacciones bioquímicas coordinadas y secuenciales, las cuales involucran pequeñas ribonucleoproteínas nucleares (snRNPs, “small nuclear ribonucleoproteins”) que, junto con proteínas no-snRNP, forman la maquinaria molecular que se conoce como el spliceosoma[5, 8] (ver figura 2.1). Se sabe de dos tipos de spliceosoma en los eucariotas: el tipo U2 (conformado por snRNPs U1, U2, U4, U5 y U6), que procesa la mayoría de los pre-mRNAs (precursor a ARN mensajero), y un tipo pequeño U12, que realiza el splicing para una fracción más pequeña de los pre-mRNAs quienes poseen intrones de tipo U12[9].

Como ilustra la figura 2.1, la tarea del spliceosoma es identificar los sitios de splicing, quitar los intrones y empalmar los exones adyacentes. Pese a algunas variaciones específicas al linaje evolutivo de cada especie, existen por lo general cuatro indicadores en una dada secuencia para indicar la ubicación correcta donde debe ubicarse el spliceosoma sobre los transcritos inmaduros. Para la gran mayoría de los intrones U2, se reconocen los nucleótidos G y T al principio del intrón, conformando el llamado sitio de splice 5' (5'ss) o “donor”, y los nucleótidos A y G al final del intrón, llamado sitio de splice 3' (3'ss) o “aceptor”.

El sitio de splicing 5' es un elemento esencial para que suceda el splicing. Estas secuencias donoras están involucradas en un paso clave de la reacción de splicing de ARN, durante la cual se reconoce la frontera entre exón e intrón, mediante una formación de complejo que depende fuertemente del apareamiento de bases entre el 5'ss y la snRNP U1, la primera de la maquinaria spliceosómica tipo U2. Específicamente, el U1 forma un apareamiento sobre la juntura exón-intrón, potencialmente involucrando a los últimos tres sitios del exón (posiciones de secuencia -3 a -1, en nuestra nomenclatura) y

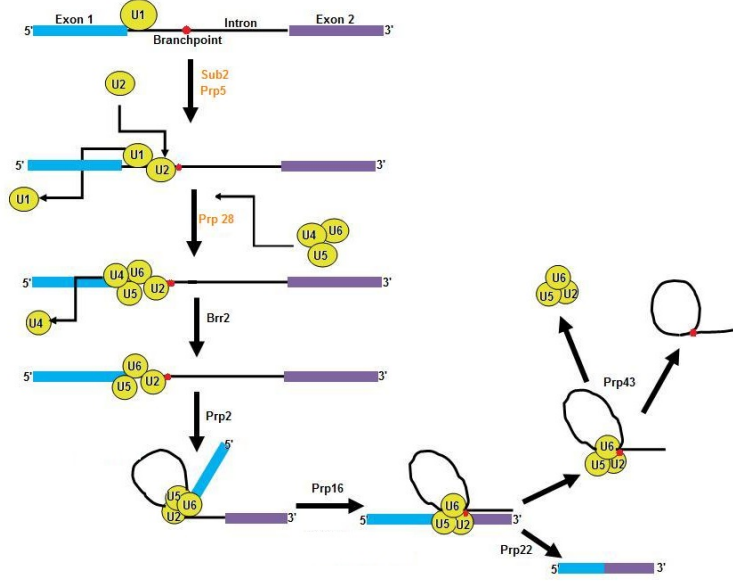


Figura 2.1: Ilustración del proceso de splicing, realizado por el spliceosoma. El mismo es formado por los complejos proteínicos U1-6, que se acoplan al transcrito inmaduro en diversos pasos hasta llevarlo a la madurez. Imagen obtenida del dominio público.

las primeras seis del intrón (posiciones de secuencia +1 a +6, donde las posiciones +1 y +2 muy frecuentemente expresan el dinucleótido GT ya mencionado como indicador importante). Este apareamiento tiene importancia también en los pasos subsecuentes del proceso, en especial con los sitios intrónicos (-3 a -1) y los sitios exónicos +5 y +6[10, 11, 12]. El funcionamiento biomolecular del mecanismo de splicing depende de forma muy fuerte, entonces, del reconocimiento de la secuencia presente en estas nueve posiciones sobre la frontera entre el exón e intrón[13, 14].

Dada la relevancia de estos sitios donores y la composición de sus secuencias, muchas investigaciones han apuntado a descubrir patrones no triviales asociados con la biología relevante. Han sido populares los enfoques de teoría de la información, en especial para inferir conocimientos respecto a la funcionalidad biológica del splicing y a su historial evolutivo. Varios estudios han utilizado la entropía y otras medidas de contenido de información para obtener intuiciones biológicas acerca del proceso de splicing. Esto incluye el uso de la entropía de Shannon en los análisis,

$$S = P(\vec{s}) \log(P(\vec{s})) \quad (2.1)$$

donde $P(\vec{s})$ es la probabilidad de observar la secuencia $\vec{s}$ (un vector de 9 dimensiones donde cada una puede tomar el valor categórico A, C, G o T) entre el sitio -3 del exón y el sitio +6 del intrón.

Uno puede estudiar la estadística de la composición de las secuencias de splicing, $\vec{s}$ considerando a cada sitio como independiente. Sin embargo este abordaje simplifica-

do no tendría en cuenta una parte no trivial de la estadística compleja embebida en estas secuencias. Varios estudios se enfocan entonces en la inferencia de patrones de correlación entre sitios, varios de los cuales encontraron asociaciones estadísticas significativas entre los pares de sitios -2 y +4, -1 y +5, y -2 y +5 [15, 16, 17]. Otros estudios buscan vincular las asociaciones encontradas a la funcionalidad bioquímica del splicing [18, 19, 20].

Sumada a la relevancia de la información y entropía en el análisis de secuencias de sitio de splicing, muchos problemas biológicos han comenzado a utilizar el principio de la máxima entropía para ajustar modelos a datos [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Esto se basa en la observación de que las correlaciones de bajo orden, incluyendo a las interacciones de pares, cumplen un rol significativo en comparación a los momentos más altos de su distribución de probabilidad correspondiente [24, 34]. En nuestro caso, utilizaremos este abordaje para encontrar un modelo de probabilidades $P(\vec{s})$ que tenga en cuenta interacciones de dos sitios. A través de este análisis se verán patrones relevantes al linaje biológico de distintos organismos, además de comportamientos generales en la estadística de las secuencias 5'ss.

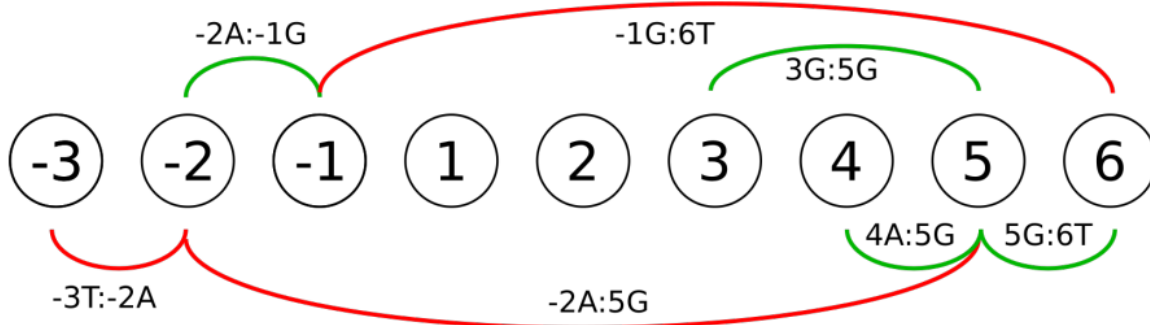


Figura 2.2: Diagrama de los nueve sitios incluidos en las secuencias analizadas (negativos exónicos, positivos intrónicos). Pueden existir interacciones negativas (rojo) o positivas (verde) entre sitios, no necesariamente adyacentes. Cada sitio puede poseer un nucleótido adenina, citosina, guanina o timina (A, C, G o T). La timina es reemplazada por el uracilo (U) luego de la transcripción a mRNA, pero por simplicidad siempre nos referiremos a él como T.

2.1. Metodología

En analogía con la mecánica estadística, se trata a nuestro sistema como una cadena de sitios que pueden mostrar distintos estados (los nucleótidos A, C, G, T), un esquema de lo cual se puede ver en la figura 2.2. En esta sección trataremos al fenómeno, y su comportamiento estadístico, en términos de ensambles y energía, y mostraremos como se avanza desde ese punto a un análisis aplicable a la biología.

Cada ensamble analizado por nuestra metodología está compuesto de N secuencias de largo 9, donde el vector $\vec{s}$ ya mencionado se puede expresar como $(s_{-3}, s_{-2}, s_{-1}, s_1, s_2, s_3, s_4, s_5, s_6)$, donde $s_i \in \{A, C, G, T\}$ (adenina, citosina, guanina, timina). Bajo la hipótesis de que la estadística subyacente se pueda describir con probabilidades marginales de un sitio $f_i(s_i)$ y de dos sitios $f_{ij}(s_i, s_j)$, planteamos una distribución de Boltzmann:

$$P(\vec{s}) = \frac{1}{Z} e^{-E_d(\vec{s})} \quad (2.2)$$

donde

$$E_d(\vec{s}) = - \sum_{i=1}^9 h_i(s_i) - \sum_{i=1, j>i}^9 J_{ij}(s_i, s_j) \quad (2.3)$$

cumple el rol del Hamiltoniano que solo depende de ocurrencias sitio-nucleótido de un sitio y co-ocurrencias de los mismos en dos sitios. La energía se ve favorecida por coeficientes de un sitio $h_i(s_i)$ grandes, mientras que un coeficiente de dos sitios $J_{ij}(s_i, s_j)$ positivo también lo favorece, y uno negativo provocará una menor probabilidad de la secuencia en cuestión al aumentar su energía. Esto implica que se tendrán 36 parámetros de un sitio (9 sitios con 4 nucleótidos posibles en cada uno) y 576 interacciones de dos sitios (16 combinaciones posibles entre dos dados aminoácidos, por 36 pares de sitios sin contar repeticiones).

El proceso para encontrar el modelo de ajuste óptimo se detalla en la sección 2.1.2. Una vez hallado un modelo óptimo para un dado conjunto de secuencias, se examina lo obtenido en base a los parámetros. La base visual de este análisis es el diagrama “circos”, utilizando el paquete **RCircos** en el lenguaje de programación **R** (detalle en la figura 2.3). Examinando estas figuras se puede inferir algunas intuiciones básicas acerca de la estadística de un y dos sitios para un dado ensamble de secuencias. Combinando este análisis, y una integración de las estadísticas características de distintas especies, podremos llegar a ciertas conclusiones significativas respecto el carácter mecánico y evolutivo del splicing.

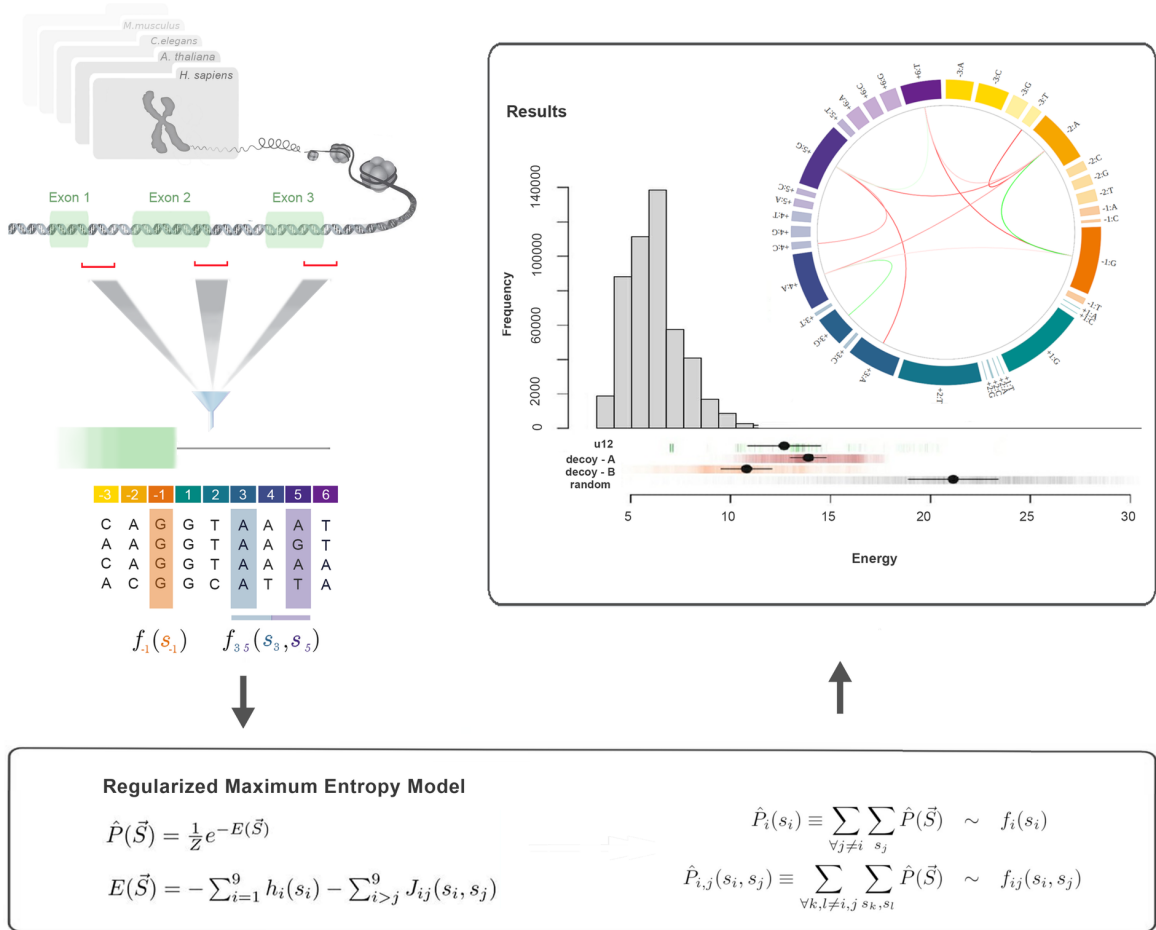


Figura 2.3: Proceso de ajuste de un dado ensamble de secuencias, extraído de datos genómicos anotados (izquierda, arriba). Se obtienen los parámetros que mejor describan las frecuencias de dos sitios (derecha, abajo) bajo la suposición de un modelo regularizado de máxima entropía (izquierda, abajo). La representación esquemática del modelo ajustado es el diagrama circos (arriba, derecha), donde cada sitio se representa por un color distinto (cálidos para el exón, fríos para el intrón), y la frecuencia de cada nucleótido en cada sitio se representa por el ancho de la entrada correspondiente en el círculo. Las interacciones no nulas obtenidas se muestran como curvas conectando dos realizaciones sitio-nucleótido distintas, donde el color rojo denota una interacción negativa y el verde una positiva. Además el histograma (arriba, derecha) muestra la distribución de energías para el ensamble analizado bajo el modelo ajustado. Son notablemente más bajas en comparación a las secuencias con intrones de tipo U12, que siguen una estadística distinta, al igual que cuando se compara con secuencias artificialmente generadas desde los parámetros obtenidos (“decoy-A” y “decoy-B”) y mucho más bajas que las secuencias completamente aleatorias (“random”).

especie	código	assembly	5'ss (únicos)	5'ss GT (únicos)
<i>Aspergillus nidulans</i>	ani	ASM1142v1	24563 (3954)	24514 (3919)
<i>Coprinopsis cinerea okayama</i>	cci	CC3	61427 (4737)	61325 (4670)
<i>Cryptococcus neoformans</i>	cne	cryp_neof_125_91_V1	36046 (2637)	35220 (2435)
<i>Magnaporthe oryzae</i>	mor	MG8	22772 (3126)	22567 (3039)
<i>Neurospora crassa</i>	ncr	NC12	16739 (2522)	16533 (2423)
<i>Arabidopsis thaliana</i>	ath	TAIR10	136036 (11286)	130069 (6803)
<i>Hordeum vulgare</i>	hvu	IBSCv2	499846 (48879)	413160 (10107)
<i>Medicago truncatula</i>	mtr	MedtrA17.4.0	156335 (11396)	148511 (6964)
<i>Oryza sativa</i>	osa	IRGSP-1.0	130702 (12688)	124646 (8418)
<i>Physcomitrium patens</i>	ppa	Phypa V3	194684 (8310)	191147 (6991)
<i>Populus trichocarpa</i>	ptri	Pop_tri_v3	179803 (10168)	176766 (9285)
<i>Solanum lycopersicum</i>	sly	SL3.0	141263 (12562)	135225 (8400)
<i>Vitis vinifera</i>	vvi	12X	108097 (9614)	105266 (7713)
<i>Anas platythynchos</i>	apl	CAU_duck1.0	152272 (9075)	146465 (5405)
<i>Bos taurus</i>	bta	ARS_UCD1.2	203086 (11919)	194432 (6624)
<i>Canis lupus familiaris</i>	clu	CanFam3.1	215912 (12437)	206340 (5521)
<i>Danio rerio</i>	dre	GRCz11	276776 (11723)	268477 (6720)
<i>Equus caballus</i>	eca	EquCab3.0	216007 (10367)	208399 (4984)
<i>Gorilla gorilla</i>	ggo	gorGor4	212738 (14620)	200262 (7697)
<i>Homo sapiens</i>	hsa	GRChv38	502197 (12129)	488939 (6206)
<i>Monodelphis domestica</i>	mdo	ASM229v1	208511 (14090)	196940 (6970)
<i>Mus musculus</i>	mmu	GRCm39	391997 (9026)	384141 (5369)
<i>Ornithorhynchus anatinus</i>	oan	mOrnAna1.p.v1	187119 (13517)	175225 (6562)
<i>Oryctolagus cuniculus</i>	ocu	OryCun2.0	147142 (12204)	138124 (5217)
<i>Sarcophilus harrisii</i>	sha	mSarHar1.11	227688 (10104)	220325 (5637)
<i>Salmo salar</i>	ssa	ICSASGv2	516921 (28445)	455502 (14209)
<i>Sus scrofa</i>	ssc	Sscrofa11.1	260642 (21094)	243064 (8300)
<i>Xenopus tropicalis</i>	xtr	Xenopus_tropicalis_v9.1	231618 (20862)	205430 (10794)
<i>Caenorhabditis elegans</i>	cel	WBcel235	127661 (7625)	124619 (5351)
<i>Drosophila melanogaster</i>	dme	BDGP6.32	63121 (4023)	62451 (3780)

Cuadro 2.1: Genomas analizados. La penúltima columna muestra la cantidad de secuencias disponibles para utilizar en el ajuste, y, entre paréntesis, la cantidad sin repetir secuencias. Lo mismo se ve en la última, excluyendo las secuencias que no hayan presentado un dinucleótido GT en los sitios 1-2 del intrón (el caso más frecuente y característico de los sitios de splicing 5'). La tercera columna muestra el nombre de referencia que tiene el ensamble de datos, en fuentes públicamente disponibles en línea. Los reinos de los hongos, de los vegetales y de los animales se separan por líneas horizontales.

2.1.1. Características de los datos

Cada ensamble de secuencias consiste en un número grande de secuencias extraídas del genoma de una dada especie, de nueve sitios cada una, donde los primeros tres (designados -3, -2 y -1) son las últimas tres posiciones del exón, y los últimos seis (de

1 a 6) corresponden a las primeras seis del intrón. Como se ve en el cuadro 2.1 los ensambles de distintas especies presentan una cantidad variable de secuencias de sitio de splicing 5', según el tamaño de su genoma.

En dicha figura vemos que contamos con datos para los genomas de cinco hongos (reino *fungi*), ocho plantas (*plantae*) y diecisiete animales (*metazoa*). A partir de esto lograremos extraer evidencia significativa respecto a las tendencias en común entre especies. Para realizar un paralelo entre estas tendencias con el desarrollo evolutivo, se obtiene a partir de una fuente independiente un *time-tree*[35]: a saber, un árbol filogenético que organiza a las distintas especies según sus puntos de bifurcación evolutivos. Esto nos muestra dicha información a lo largo de millones de años y distintas escalas temporales geológicas.

2.1.2. Características del ajuste

La tarea en base a este modelo es obtener los parámetros óptimos para describir las frecuencias de uno y dos sitios correspondientes a un dado ensamble de secuencias. Como se podría terminar con una cantidad enorme de parámetros de dos sitios J_{ij} no nulos, se implementa un esquema de regularización mediante el cual se penalizan dichos coeficientes no nulos si no ayudan a una mejor representación de la estadística subyacente.

Una iteración del proceso de ajuste genera primero, con los parámetros h_i y J_{ij} presentes, un ensamble artificial de secuencias por muestreo de rechazo según un esquema de Metropolis-Hastings, acorde a la distribución 2.2 y con la energía dada por la ecuación 2.3. Luego se calcula la probabilidad marginal de un dado nucleótido en un dado sitio (P_i) y de un dinucleótido entre dos dados sitios P_{ij} a partir del ensamble artificial. La actualización de los parámetros del modelo se realiza entonces en función del desvío registrado entre estas probabilidades y las originalmente observadas, de acuerdo a las siguientes prescripciones:

$$\begin{aligned}
 & h_i^{t+1} \leftarrow h_i^t - \epsilon_h [f_i - P_i] \\
 & \text{si } J_{ij}^t = 0 \\
 & \quad J_{ij}^{t+1} \leftarrow \begin{cases} 0 & \text{if } |\Delta| < \gamma \\ \epsilon_j [\Delta - \text{sign}(\Delta)] & \text{if } |\Delta| > \gamma \end{cases} \\
 & \text{si } J_{ij}^t \neq 0 \\
 & \quad J_{ij}^{t+1} \leftarrow \begin{cases} J_{ij}^t + \epsilon_j [\Delta - \gamma \text{sign}(J_{ij}^t)] & \text{if } \eta > 0 \\ 0 & \text{if } \eta < 0 \end{cases}
 \end{aligned}$$

donde $\Delta = f_{ij} - P_{ij}$ y $\eta = (J_{ij}^t + \epsilon_j [\Delta - \gamma \text{sign}(J_{ij}^t)]) * J_{ij}^t$. Se puede observar que γ cumple el rol de una regularización: si un parámetro J es nulo, el error Δ en sus frecuencias de dos-sitios correspondientes tendrá que superar γ para tener la posibilidad

de volverse no-nulo, mientras que si el parámetro ya es no-nulo y se acerca lo suficiente a un valor mínimo, se lo forzará a regresar a cero. De esta forma imponemos sobre el modelo un incentivo hacia la parsimonia, y podemos controlar hasta qué punto excluir una descripción completa del sistema, a favor de tener que manejar una menor cantidad de parámetros, que podrían potencialmente ser muchos.

Podemos ver el efecto de esta regularización en la figura 2.5. A medida que se relaja la constante de regularización γ , se obtiene una mayor cantidad de parámetros de interacción J no nulos, y a cambio de esa complejidad agregada en el modelo se obtiene una descripción más precisa de la información en un dado ensamble y sus señales subyacentes. Como vemos en la figura 2.6, el comportamiento se nota consistentemente a lo largo de todas las especies examinadas.

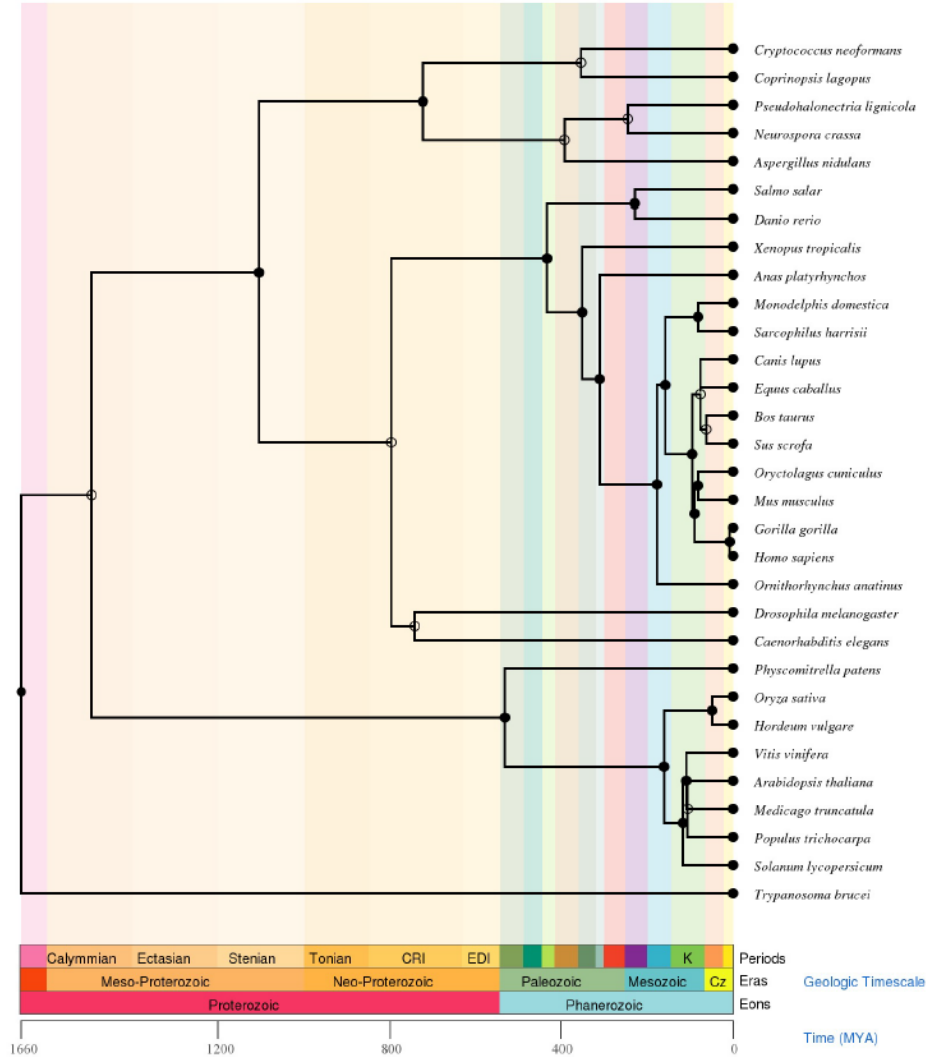


Figura 2.4: Árbol filogenético inferido a partir de las señales filogenéticas correspondientes a 30 especies analizadas. Generado utilizando timetree.org[35].

2.1.3. Convergencia y bases de atracción

Para generar las secuencias artificiales mencionadas en la sección 2.1.2 se emplea muestreo de rechazo para partir de una secuencia aleatoria y progresivamente modificarla hasta llegar a una muestra de la distribución de equilibrio de las secuencias. A la hora de plantear un equilibrio, cabe preguntar ¿existe el equilibrio termodinámico para este sistema? Es decir, ¿existe una distribución estacionaria única?

Para explorar estas preguntas en términos de recorridos estadísticos en un paisaje

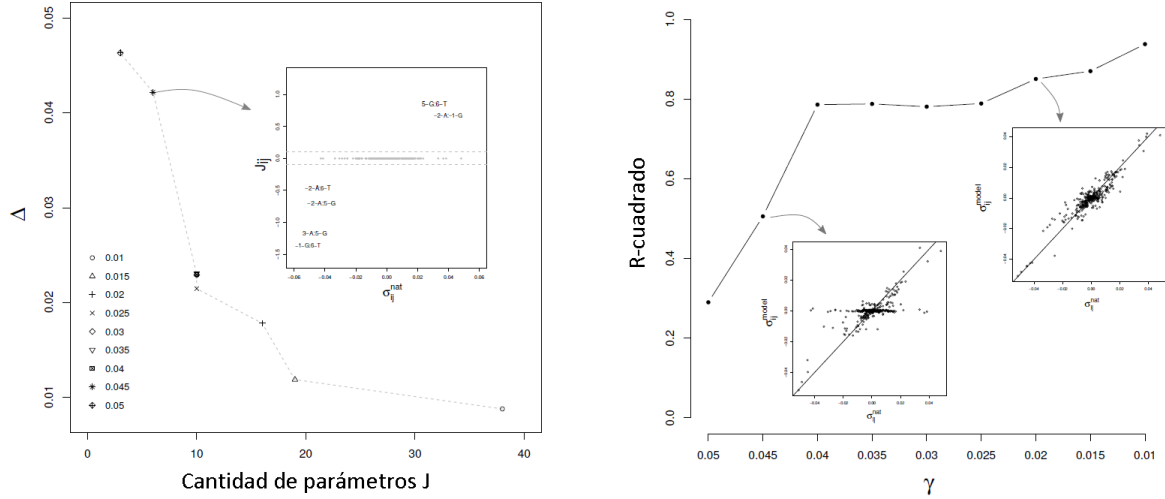


Figura 2.5: **(Izquierda)** Desvío máximo observado desde probabilidades de dos-sitios estimadas por el modelo ajustado, y sus contrapartes observadas desde los datos de los ensambles. El error disminuye a medida que la regularización permite una mayor cantidad de parámetros J de interacción, o sea, al relajar el hiperparámetro de regularización γ . **(Derecha)** Cercanía de las correlaciones estimadas por el modelo ajustado con las correlaciones observadas en los datos, a medida que se relaja la regularización. Se observa un R-cuadrado creciente en el ajuste entre ambos datos, a costa de incorporar más parámetros J en el modelo.

energético, construimos para el ensamble de *H. sapiens* una red compleja donde cada secuencia posible es representada por un nodo: es decir, al tener nueve sitios y cuatro realizaciones posibles para cada uno, se cuenta con $4^9 = 262144$ nodos. Estos se conectan entre sí solo al cumplir ambas de las siguientes condiciones: (a) varían en exactamente un sitio, exigiendo una distancia de Hamming de 1 entre dos secuencias (cada nodo tendrá, máximo, tres vecinos por sitio, o sea grado $k = 27$); (b) el nodo destino tiene una energía menor, según los parámetros convergidos, que el nodo origen.

Como consecuencia se obtiene una red como la vista en la figura 2.7. Es necesariamente dirigida acíclica, ya que si la energía disminuye en camino entre dos nodos, lo hará por todos los caminos que los pueden conectar. Esto es análogo a pensar en la energía como un campo escalar sobre el espacio discreto del grafo, y que su gradiente tendrá circulación nula (este tipo de idea exploraremos en otro contexto, en el capítulo 4).

Durante el recorrido se puede saltar de nodo a nodo en una sola dirección, sin ciclos posibles para regresar. Considerando componentes en el sentido fuerte, esto significa que el paisaje se separará en distintos componentes, cada uno correspondiente a una distinta cuenca de atracción. Podemos analizar el tamaño y las energías presentes en estos componentes, obteniendo el resultado que vemos en la figura 2.8, evaluando las energías para el subconjunto de la red cuyas secuencias aparecieron alguna vez en los

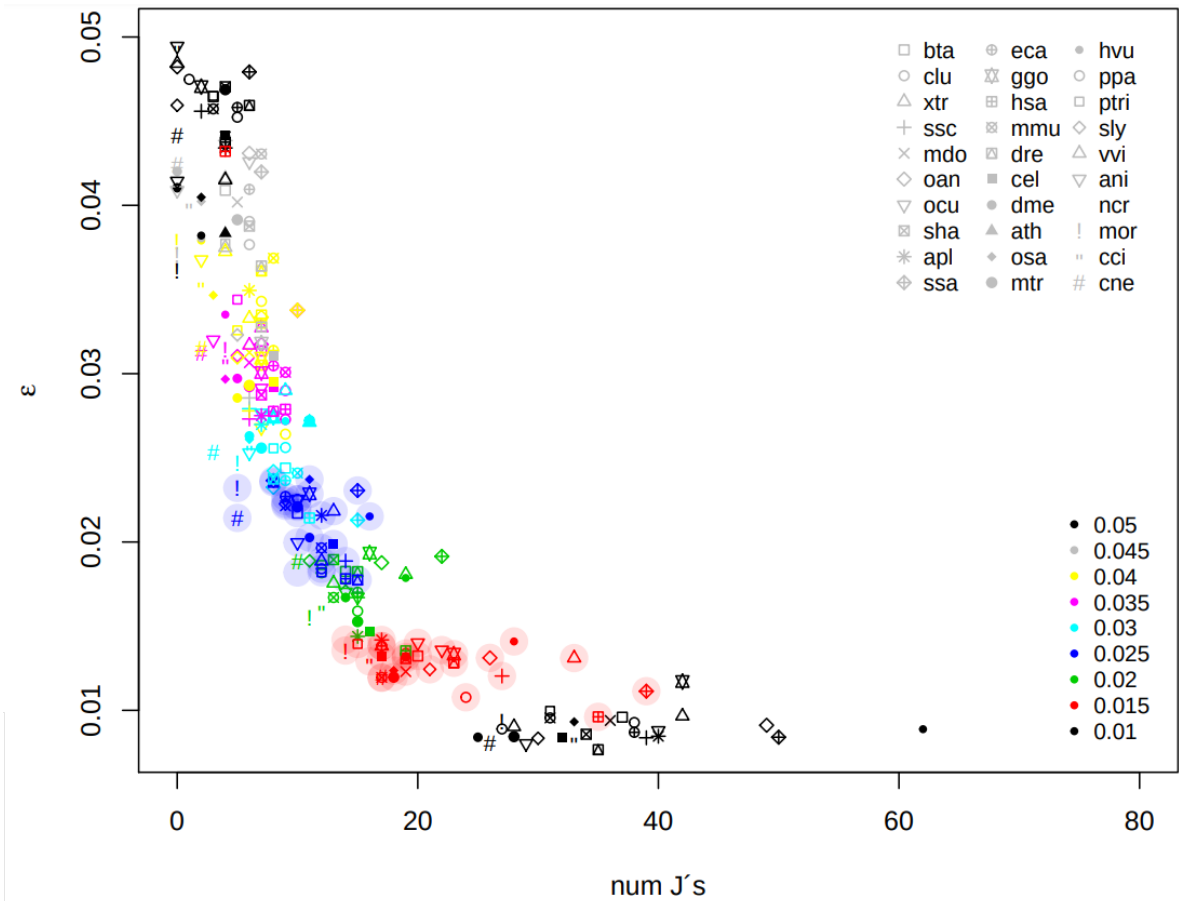


Figura 2.6: Convergencia progresiva para los modelos de todas las especies analizadas. Demuestran un comportamiento consistente en términos del coeficiente γ de regularización: cuanto más laxa la misma, mayor la cantidad de parámetros J contemplados por el modelo, y menor el error ϵ en las probabilidades estimadas de dos sitios.

ensambles. Esta figura corrobora que la gran mayoría de las secuencias existe en un componente conexo de energía drásticamente más baja que los demás, con un mínimo muy estable en comparación al resto de la red. Esto indica que, con el ruido agregado de la distribución de Boltzmann, podemos tener confianza de que el muestreo será capaz de recorrer las partes informativas de la red, y el algoritmo de ajuste nos dará parámetros que representan a la información allí contenida.

2.2. Señales conservadas

Además de obtener modelos separados para cada una de las especies mencionadas, a partir de estos modelos se pueden encontrar tendencias en común. Algunas de las

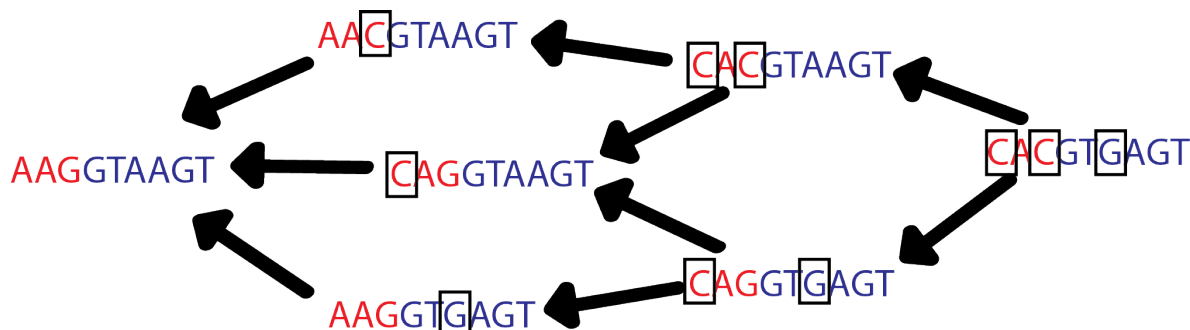


Figura 2.7: Ejemplo esquemático de una parte de una red de descenso de energía entre secuencias. Cada una de las secuencias es representada por un nodo distinto, y su descenso de energía las lleva unidireccionalmente hacia el ejemplo AAGGTAAGT. Se muestran los sitios exónicos en rojo y los sitios intrónicos en azul. Con un rectángulo se señala las letras que varían, en comparación a la secuencia de menor energía (a la izquierda).

observaciones más interesantes surgen cuando se comparan interacciones de dos sitios entre las dos regiones del sitio de splicing, el exón y el intrón.

Fuimos capaces de encontrar ciertas tendencias promedio que se mantienen consistentes entre las especies, en especial en lo que se refiere a los nucleótidos consenso de cada sitio: la realización de un sitio más frecuente, o, si se quiere, la más característica del sitio de splicing 5'. Por un lado, se observó que las interacciones entre nucleótidos consenso dentro del mismo exón son positivas; lo mismo es válido para el comportamiento interno del intrón. Esto puede no ser sorprendente: es lógico, desde la estadística de un sitio, que tanto el exón como el intrón tengan menor energía cuando se tiende a realizaciones consenso. Sin embargo la disminución en energía excede lo esperado por la estadística de un sitio, a tal punto que emergen las mencionadas interacciones de dos sitios positivos e incentivan aún más que los distintos sitios del intrón muestren consenso al mismo tiempo (y lo mismo para el exón, independientemente).

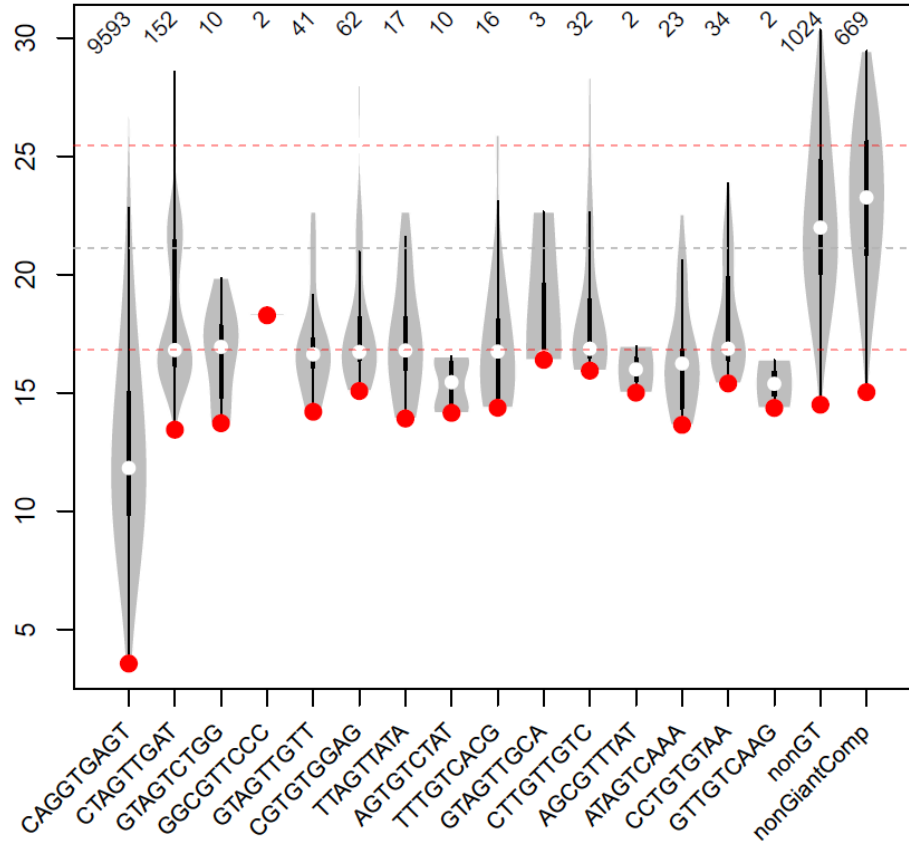


Figura 2.8: Bases de atracción sobre la red dirigida acíclica de secuencias. La base más grande es también la que tiene la distribución de menor energía, y una secuencia atractora reconocida en genética de sitios 5' como muy común, y característica de los mismos. Las demás bases de atracción muestran distintos atractores con un mucho menor número de nodos a su alcance. La penúltima distribución corresponde a las secuencias que no mostraron un dinucleótido GT al principio del intrón, mostrando la alta escala de energía que uno espera. La última distribución muestra secuencias que quedaron totalmente aisladas de todos los demás componentes en ambos sentidos de los vínculos.

Una observación fuerte y sorprendente fue lo contrario en las interacciones entre posiciones del intrón y el exón. En este caso se ven parámetros J negativos, lo que indica que, en secuencias de splicing, la presencia de sitios consenso del lado exónico suele no coincidir con la presencia de sitios consenso del lado intrónico. Estos resultados extienden observaciones previas realizadas sobre sitios de splicing de mamíferos, como humanos y ratón [17, 18, 36], y apoyan la idea de que se favorece la complementariedad del 5'ss o bien en la parte exónica o bien en la intrónica, pero no en ambas partes al mismo tiempo.

El extender este resultado a tres reinos distintos de eucariotas, de forma consistente entre las diversas especies de la tabla 2.2, puede indicar que esta característica adversarial entre el exón e intrón sea importante en general a la mecánica del splicing. Sin embargo se puede refinar progresivamente la escala en la cual buscamos patrones en común. Esto haremos en las secciones siguientes, aprovechando el poder descriptivo de los modelos ajustados, y ofreciendo perspectivas acerca del posible significado bioquímico y evolutivo de los resultados.

2.3. Divergencia de las señales de interacción según diferenciación de especies

A pesar de un grado alto de conservación en el spliceosoma, las frecuencias de un sitio han demostrado un grado no trivial de especificidad en los eucariotas [37, 38]. Nuestros modelos logran capturar un comportamiento similar en las probabilidades marginales de ocurrencias de dos sitios. Esto se puede observar mediante una comparación de los parámetros obtenidos por distintas especies, estableciendo una medida de cercanía en base a esta similaridad, y formando a partir de esto una estructura jerárquica.

Esta estructura jerárquica se puede observar en la figura 2.9, comparado a la información filogenética evolutiva que ya habíamos explicado con la ayuda de la figura 2.4. En la figura 2.9 comparamos la jerarquía de agrupación de especies via los parámetros obtenidos por sus modelos a una regularización $\gamma = 0,025$ (izquierda) con el árbol filogenético (**centro**). Podemos ver que ambas estructuras demuestran una fuerte concordancia, y por sobre todo, una separación fuerte entre plantas, animales y hongos. La similaridad entre ambas agrupaciones se corrobora mediante un análisis de Fowlkes-Mallows [39], con un p-valor $< 10^{-4}$.

El panel derecho de la figura 2.9 nos indica que ciertos parámetros de interacción tienen una presencia llamativa en ciertos reinos (plantas, animales u hongos) y una ausencia igual de llamativa en otros. Nos podemos preguntar entonces si esta diferencia es estadísticamente significativa. Para cada parámetro J_{ij} no nulo identificado en cualquier especie, en base a su presencia en los tres grupos de plantas, animales u hongos se realizó un test de Maddison-Slatkin [40] para determinar significancia estadística, con los resultados reportados en la 2.3. El test utiliza la metodología de parsimonia de Sankoff

[41] y provee los p-valores para la hipótesis nula (falta de significancia estadística).

Cuadro 2.2: Patrones conservados. Las interacciones medias entre regiones exónicas e intrónicas para los ensambles de distintas especies. EC, ENC, IC, e INC representan los sitios exón-consenso, exón-no-consenso, intrón-consenso e intrón-no-consenso, respectivamente. Entonces, por ejemplo, IC-EC resume los valores medios de interacciones entre sitios intrónicos con nucleótido consenso, y sitios exónicos con nucleótido consenso.

Species	IC-EC	IC-ENC	INC-EC	INC-ENC	IC-IC	IC-INC	INC-INC	EC-EC	EC-ENC	ENC-ENC
cne	-0.41	0	0	0	-0.14	-0.01	0	0.88	0.13	-0.13
ani	-0.5	0	0.03	0	0.21	0.02	0	0.82	0	-0.15
ncr	-0.68	0	0	0	0.26	0.03	0	0.68	0	-0.07
mor	-0.45	0	0	0	0.1	0	0	0.87	0	-0.15
cci	-0.59	0	0.04	0	0.15	0.09	0	0.7	0	-0.35
ath	-0.39	0	0.05	0	0.37	0	0	0.83	0	-0.11
hvu	0	0.09	0	-0.06	-0.13	0.92	0.09	0	-0.16	-0.29
mtr	-0.3	0	0	0	0.1	0	0	0.95	0	0
osa	-0.24	0	0.01	0	0.23	0.05	-0.04	0.94	0	0
ppa	-0.77	0	-0.01	0	-0.02	-0.01	-0.03	0.64	0	-0.08
ptri	-0.42	0	0	0	-0.02	0	0	0.91	0	0
sly	-0.36	0	-0.01	0	0.21	0	0	0.91	0	-0.07
vvi	-0.37	0	0	0	-0.02	0	0	0.93	0	-0.04
apl	-0.4	0	0	0	0.29	0.01	0	0.87	0	0
bta	-0.28	0	0	0	0.57	0	0	0.78	0	0
clu	-0.3	0	0	0	0.52	0.03	0	0.79	0	0
dre	-0.45	0	0	0	0.23	0	-0.02	0.86	0	0
eca	-0.37	0	0	0	0.48	0	0	0.8	0	0
ggo	-0.18	0	0	0	0.59	-0.01	0	0.78	-0.1	0
hsa	-0.48	0	0	0	0.14	-0.01	0	0.86	0	0
mdo	-0.22	0	0	0	0.67	0	0	0.69	-0.15	0
mmu	-0.51	0	0	0	0.08	0	0	0.86	-0.06	0
oan	-0.19	0	0	0	0.81	-0.06	0.03	0.55	0	0
ocu	-0.24	0	0	0	0.61	0	0	0.76	0	0
sha	-0.33	0	0	0	0.6	0.01	0	0.73	-0.05	0
ssa	0.14	0	-0.06	0	0.74	0.01	0.02	0.63	-0.16	0
ssc	-0.14	0	0	0	0.7	-0.04	0	0.7	-0.07	0
xtr	0.11	0	-0.04	0	0.56	-0.02	0.03	0.81	-0.13	-0.04
dme	-0.89	0	0	0	-0.23	-0.04	0	0.38	0	-0.08
cel	-0.9	0	0.05	0	-0.21	-0.01	0	-0.35	0	-0.13

Como es de esperar, desde ambas figuras podemos corroborar que esas interacciones son significativamente distintas para alguna partición de los grupos de plantas, animales y hongos. Las primeras tres entradas del cuadro 2.3 corresponden a interacciones negativas (cuando no nulas) entre consensos exónicos y consensos intrónicos, específicamente aquellos que vinculan el consenso -1G del exón con los consensos de los últimos tres sitios del intrón. La -1G:+6T es expresada en animales y hongos, la -1G:5G en plantas, y la -1G:4A en plantas y hongos. La cuarta entrada del cuadro 2.3 es una interacción estabilizadora de consenso intrón-intrón +3G:+5G presente en el 70 % de los animales y ningún hongo ni planta.

La última interacción, +3G:+4A, es negativa, está presente en todas las plantas pero pocos hongos y animales. La misma desincentiva la presencia del no-consenso +3G en plantas (que, cabe destacar, en hongos y animales tiene una frecuencia que sí es consenso) dada la presencia del consenso +4T (que, inversamente, en hongos y animales tiene una frecuencia más baja, que la vuelve no-consenso).

Estos patrones obedecen algunas de las tendencias que observamos en el cuadro 2.2: en el caso de las tres primeras, una desestabilización mutua entre consensos exónicos y consensos intrónicos, observada en la primera columna del cuadro 2.2. La cuarta interacción sigue el patrón de estabilidad interna del consenso del intrón, la tendencia más frecuente en la quinta columna de la tabla (tendencia que, cabe destacar, sobresale principalmente para los animales). La última interacción representa una desincentivación de un no-consenso en el intrón dado un consenso en el intrón: esto no se observa en el cuadro 2.2, que vuelve aún más interesante la ocurrencia de este fenómeno en las plantas. Y, por más que no sea percibido en la sexta columna de dicha tabla, sí sigue la lógica de que los consensos en el intrón busquen coexistir con otros consensos intrónicos.

Estos hallazgos indican que existe información filogenética no trivial contenida dentro de las correlaciones de dos sitios en los datos, extraída a través de nuestro algoritmo y codificada dentro de los parámetros de cada modelo ajustado. Las asociaciones filogenéticas más fuertes (los reinos de las plantas, animales y hongos) demostraron una asociación con las probabilidades estimadas de los modelos en general, y específicamente presentaron una relación significativa con las cinco interacciones vistas en el cuadro 2.3.

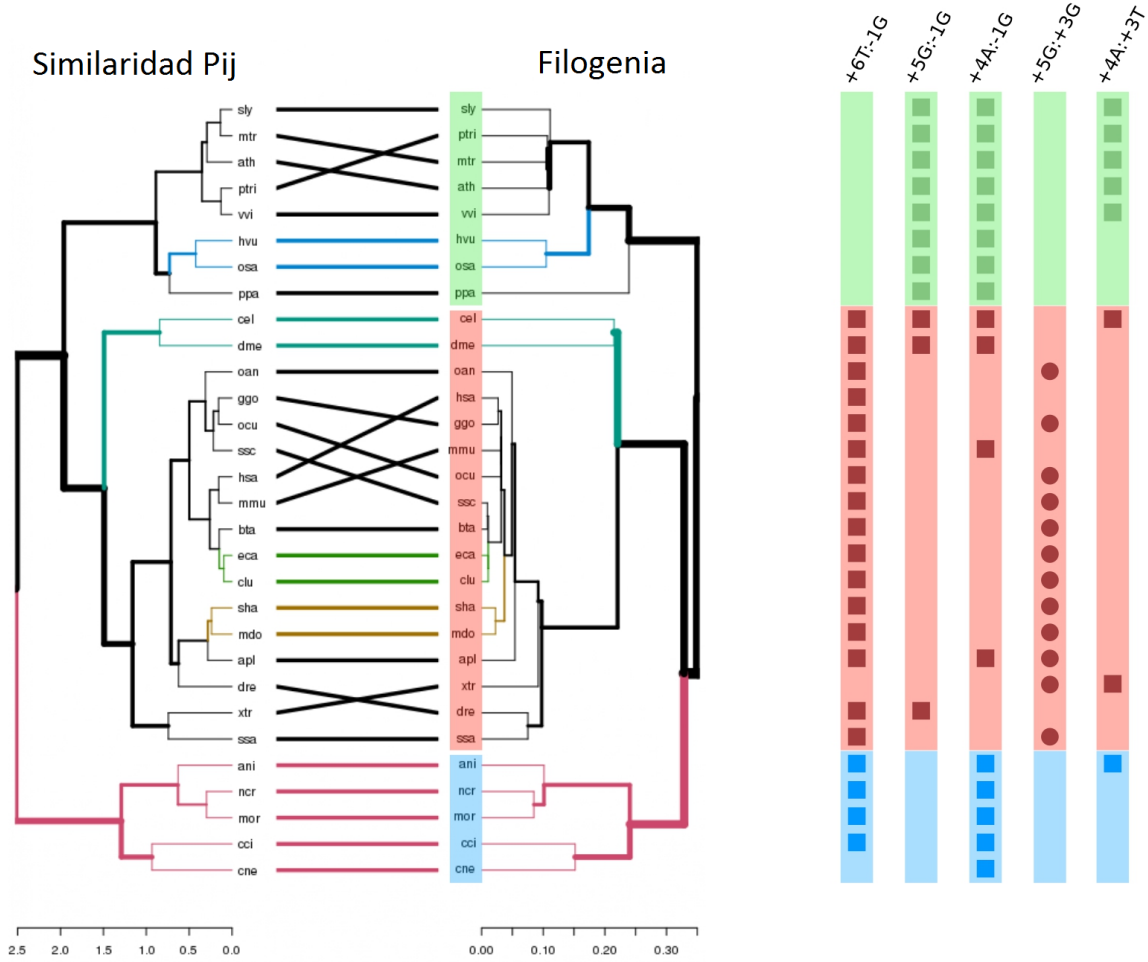


Figura 2.9: **(Izquierda)** Agrupación jerárquica de especies según la similaridad entre las probabilidades de dos-sitios que generan sus modelos ajustados, combinando las frecuencias de un sitio con las interacciones de dos sitios. **(Centro)** Árbol filogenético construido a partir de *time-tree.org*[35], comparando según la agrupación por parámetros a la izquierda. **(Derecha)** Coeficientes de interacción significativamente distintos ($p < 0,05$) entre los reinos *plantae* (verde), *metazoa* (rojo) y *fungi* (azul), según un test de Slatkin-Maddison[40].

Parámetro	<i>plantae</i>	<i>metazoa</i>	<i>fungi</i>	MS.obs	MS.null	MS.pvD
-1G:+6T	0	16	4	3	3,9,10	4.1e-3
-1G:+5G	8	3	0	3	4,9,11	0
-1G:+4A	8	4	5	3	3,10,13	4.1e-3
+3G:+5G	0	12	0	4	4,10,12	4.1e-3
+3T:+4A	5	2	1	3	3,7,7	5.0e-2

Cuadro 2.3: Parámetros de interacción estadísticamente significativos según un test de Maddison-Slatkin, como mostrados en la figura 2.9. Se muestra en cuantas plantas, animales y hongos el parámetro demostró un valor no nulo (y siempre del mismo signo). En la quinta columna se indica el número de pasos evolutivos de Sankoff inferidos como medida de parsimonia para explicar la agrupación de especies que demostraron el parámetro como no nulo [41]. En la sexta columna se muestra la cantidad mínima, mediana y máxima de pasos obtenidos al repetir este procedimiento con distinta aleatoriedad. La última columna muestra los p-valores para la hipótesis nula de que la agrupación no sea significativa, según la distribución obtenida para los pasos evolutivos mencionados (ajustados por método de Bonferroni).

2.4. Modelos consolidados

Con tal de explorar la partición plantas/animales/hongos en lo pertinente a los modelos ajustados, ideamos un nuevo análisis para corroborar si cada grupo tiene su propia firma en estadística de splicing. Para cada uno de estos grupos creamos un ensamble consolidado muestreando de manera uniforme y proporcional un número de secuencias de las distintas especies partícipes de cada grupo (800000 para animales y plantas, y 161547 para los hongos por menor cantidad de datos). Como resultado se obtuvo un ensamble híbrido para plantas, otro para animales, y otro para hongos. A cada uno se le ajustó un modelo distinto de máxima entropía para obtener las probabilidades de un sitio y las interacciones de dos sitios correspondientes, visualizadas en forma de **circos** en la figura 2.10 (modelo en $\gamma = 0,025$, regularización lo suficientemente alta para visualizar las interacciones más significativas filtrando la información de índole más ruidosa).

Observamos la re-emergencia de los patrones de interacción identificados como significativos en la sección 2.3. Además se nota que las interacciones siguen la lógica de estabilidad interna del intrón/exón e inestabilidad mutua entre los dos, que vimos en la sección 2.2.

Además se observan en sus grupos correspondientes las primeras cuatro interacciones significativas del cuadro 2.3 en sus grupos correspondientes. A saber, se repite la interacción fuerte negativa +6T:-1G en animales y hongos, reemplazada por la +5T:-1G en las plantas. Se ve también la interacción negativa +4A:-1G en plantas y hongos, y la positiva +3G:+5G solo en los hongos.

Este resultado muestra una consistencia tanto de la lógica conservada entre especies de la sección 2.2 como de las diferencias y agrupamiento entre ellas, explorada en la

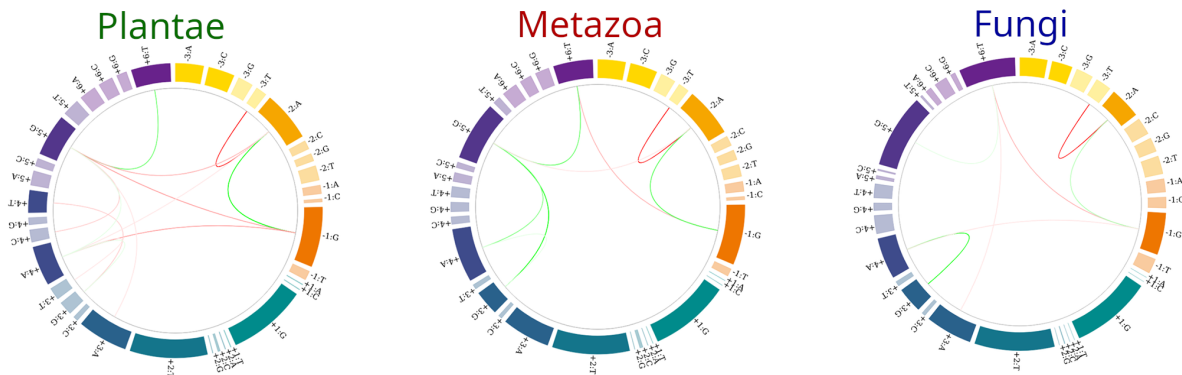


Figura 2.10: Circos representando los modelos obtenidos al ajustar los parámetros sobre un conjunto de datos ocupado por muestras aleatorias de las especies de plantas (**izquierda**), animales (**centro**) y hongos (**derecha**). Se notan diferencias importantes en las interacciones entre sitios. Entre estas, las más llamativas se repiten de la figura 2.9 y del cuadro 2.3.

sección 2.3.

2.5. Discusión

El reconocimiento molecular de los sitios de splicing 5' se enfrenta a varios desafíos. Por un lado, con la excepción de los primeros dos sitios intrónicos GT, las secuencias 5'ss son altamente degeneradas. Por el otro, el reconocimiento sucede en momentos distintos del ciclo de splicing, a través de la interacción con varios complejos que se acoplan de forma completa o parcial a la secuencia 5'. En tercer lugar, los snRNA no solo se ven involucrados en el reconocimiento, sino que muchas proteínas presentes en ellos también cobran importancia en la estabilización del acoplamiento. Finalmente, el splicing ocurre en un contexto genómico en el cual los factores de estructura genética y la presencia de las señales cis o trans puede influir significativamente sobre la regulación fina de cada sitio 5'. En esta sección apuntamos a minar señales conservadas y divergentes en secuencias 5' donoras, utilizando el criterio de que los patrones estadísticos subyacentes debían reflejar aspectos relevantes del paisaje complejo bioquímico descrito.

Nuestra estrategia de maximización de entropía nos permitió recuperar resultados previos dentro de un marco unificado y obtener nuevas perspectivas acerca de los aspectos regulares embebidos en la estadística detrás de las secuencias donoras. Por ejemplo, la escala de energías vista en la figura 2.3 (derecha, arriba), formalizada en las ecuaciones 2.2 y 2.3, sigue la misma idea fundamental que el SD-Score (definido como el logaritmo de la frecuencia de una secuencia donora) introducido por Sahashi et al con tal de predecir resultados de splicing observados en mini-genes artificiales diseñados [18]. A pesar del gran número de elementos cis y trans que contribuyen a la regulación del splicing, este hallazgo nos indica que la energía E_d podría, a cierto grado, reflejar

la fuerza de un sitio donar. En este sentido, no solo sirve para cuantificar qué tanto una secuencia se ve representada en el genoma, sino que también nos provee una escala significativa con la cual medir el grado de complementariedad con la maquinaria del spliceosoma. La mayor parte de las secuencias observadas presentaron valores intermedios de E_d , indicando una aversión estadística a la concordancia perfecta con los nueve sitios. Esto es consistente con ciertas observaciones previas que afirman que un mínimo de 5-6 pares de Watson-Crick son necesarios para el reconocimiento de un sitio de splicing, pero que un apareamiento excesivo (> 7 bases) es contraproducente [17]. Este acoplamiento suelto podría favorecer la procesividad de la reacción de splicing, y favorecer escenarios donde el binding efectivo es regulado por terceros.

Con la ayuda de nuestro modelo también identificamos interacciones generales de dos sitios, que indican que las energías también dependen de la distribución espacial de pares a lo largo de la secuencia del sitio de splicing. A pesar de tener cierto grado de especificidad por especie, los acoplamientos entre sitios de consenso en el exón estuvieron sesgados hacia los valores positivos, así como los del intrón entre sí (columnas 5 y 8 del cuadro 2.2). En contraste, los sitios consenso del exón tuvieron interacciones predominantemente negativas con los sitios consenso del intrón (columna 1 del mismo cuadro). Estos resultados presentan un acuerdo con ciertas observaciones previas observadas a partir de datos de secuencias humanas y de ratón [17, 18, 19] y apoyan la idea de que un grado alto de complementariedad se favorece alternativamente en la sección intrónica o exónica del sitio de splicing.

La identificación consistente de los acoplamientos de dos sitios indica que las ocurrencias de nucleótidos en distintas posiciones del 5'ss no son independientes. Esta observación ha sido reportada previamente, y fuimos capaces de mostrar que las probabilidades conjuntas entre pares de nucleótidos contienen información biológicamente significativa. Esto se ve a través del dendrograma construido a partir de las distribuciones de probabilidad de los modelos ajustados, que siguen de forma cercana la relación filogenética entre especies analizadas (figura 2.9). Las secuencias donoras constituyen sitios de reconocimiento no solo para el snRNA U1 en las primeras etapas del complejo spliceosómico, sino también para U6 y U5 en el paso pre-catalítico de la reacción. Se pueden esperar mecanismos de compensación útiles para asegurar la fidelidad del splicing incluso frente a la presencia de una mayor variabilidad en los sitios. En este contexto, nuestros hallazgos indican que las regularidades obtenidas en el proceso de splicing hacen eco de los procesos evolutivos divergentes vinculados a las especificidades estructurales utilizadas por la maquinaria del splicing para definir el borde exón-intrón. Sin embargo, puede resultar una tarea difícil el establecer una relación directa entre las interacciones obtenidas por el modelo y los detalles del reconocimiento en splicing. Otros estudios como el presentado por Schwartz et al [38] no han sido capaces de mostrar de forma definitiva una correlación entre variaciones de secuencia en sitios 5' y la secuencia complementaria en el snRNA U1. Esto es, en parte, porque el reconocimiento de sitios de splicing resulta de una cantidad enorme de factores. Por ejemplo un estudio

reciente [42] determinó que la fidelidad del spliceosoma está influenciada por un gran número de factores accesorios que no son indispensables para que el proceso de splicing tome lugar.

Nuestros resultados también indican que las señales filogenéticas pueden ser capturadas por los parámetros de interacción de nuestro modelo. Estas asociaciones poseen una robustez que las volvió detectables tanto en modelos de regularización fuerte como débil, cuando se analizaron especie por especie, y cuando dichas especies fueron consolidadas en grupos separados para las plantas, los animales y los hongos.

Varias de las interacciones detectadas ya han sido reportadas en la literatura, en el contexto de ciertos estudios más específicos al ser humano, al ratón o a otros mamíferos [15, 16, 17, 18, 19]. Por ejemplo, nuestros modelos capturaron fácilmente las interacciones intrónicas positivas +4:+5 y +5:+6, además de una interacción negativa -2:+5 de exón a intrón. También hallamos una interacción -1:+5 negativa en plantas; sin embargo, a diferencia de los resultados obtenidos por otros estudios sobre humanos y otros mamíferos, no obtuvimos dicha interacción para los animales en general. Ambos sitios vinculados en esta interacción son altamente informativos; Yeo y Burge obtuvieron la misma relación negativa entre ellos, analizando 12700 intrones en 1821 transcritos no-redundantes [20], y Carmel et al infirieron la relación compensadora -1G:5G desde un análisis comparativo de 8869 exones homólogos entre humano y ratón. Wong et al encontraron recientemente un resultado distinto, al cuantificar la actividad de 32768 5'ss artificiales en tres contextos genómicos, y vieron una interacción epistática positiva entre esos sitios [14]. Si bien nuestro análisis comparativo logró detectar esta interacción en algunos animales, *H. sapiens* no fue uno de ellos; nuestro modelo tampoco pudo recuperar dicha interacción cuando fue entrenado con el ensamble animal consolidado. Esto tiene su apoyo en la baja correlación observada entre las ocurrencias -1G y +5G en los datos para dicho ensamble, indicando que esta discrepancia viene de los datos, y no del entrenamiento del modelo. Este resultado fue robusto, sin importar si se utilizaba el conjunto completo de 5'ss anotados, o restringido a las secuencias presentando el dinucleótido canónico GT al principio de la sección intrónica.

De acuerdo a nuestro modelo, el comportamiento compensatorio entre los últimos sitios exónicos e intrónicos se puede entender según un grupo de interacciones estabilizadoras (+5G:+6T, -1G:-2A) e inestabilizadoras (-2A:+5G, -1G:+6T). Este patrón complejo es consistente con las observaciones recientes de Artemyeva y Porter, quienes reportaron que la composición de pares en las posiciones -1 y -2 se veían significativamente alteradas según la ocurrencia (o falta de) de un nucleótido +5G [13]. Además, la relevancia de la posición +6 ya había sido notada por Carmel et al, conectada a ciertas aberraciones de splicing que llevan a la disautonomía familiar [17]. Este estudio da evidencia experimental de que un apareamiento de bases en la posición -1 impide el splicing aberrante del 20avo intrón del gen IBKAP causado por un apareamiento fallido de la posición +6 con el snRNA U1. Nosotros observamos un mecanismo compensatorio similar entre las posiciones intrónicas (+5 y +6) y exónicas (-2 y -1) en plantas y ani-

males. En el caso de las plantas, sin embargo, detectamos un acomplamiento negativo -1G:5G que reemplazó el -1G:6T observado en animales. Estas interacciones afectan posiciones intrónicas de relativamente bajo contenido de información en plantas, indicando que las interacciones podrían cumplir un rol importante en como estas especies tratan con la variabilidad en las secuencias.

Se han resaltado muchas diferencias entre las plantas y los animales en términos de sus mecanismos de splicing. Quizá una de las más directas es la diferencia en largos típicos de longitud de intrón y la prevalencia de distintos tipos de eventos de splicing: salteo de exones en animales y retención de intrones en plantas [9, 43]. Muchos estudios recientes también han señalado no solo la arquitectura de genes, sino también las diferencias funcionales entre el splicing alternativo en plantas y animales [44, 45, 46, 47]. Por ejemplo, aunque el spliceosoma en las plantas aún no ha sido aislado, el número de factores de splicing identificados en *A. thaliana* es casi el doble del observado en seres humanos [48]. Otro punto de interés es que la prevalencia de eventos de retención de intrones junto con la degradación del transcripto NMD y la secuestación nuclear indica que, en contraste a los organismos animales, el splicing en las plantas podría jugar un rol funcional regulatorio, relacionado a la respuesta al estrés, además de expandir la diversidad proteómica [9, 43, 49, 50, 51]. En este contexto, nuestros resultados resaltan diferencias significativas en interacciones de dos sitios que involucran occurrencias posición-nucleótido relevantes a ciertas consideraciones evolutivas y funcionales. Basado en nuestros hallazgos, creemos que es una vía fértil de investigación la conexión entre estas diferencias y los rasgos mecánicos del proceso de splicing en las plantas y los animales.

2.6. Conclusión

Este capítulo empleó un enfoque regularizado de máxima entropía para obtener una serie de modelos generativos probabilistas para las secuencias de sitio de splicing donador 5' en 30 especies eucariotas distintas. Este modelo incorporó una escala de energía basada en los datos, que capturaba la abundancia de una dada secuencia dentro de un genoma. Éste estadístico de energía sirvió como una medida práctica para caracterizar secuencias, representando el continuo entre el orden completo y el estado desordenado de las secuencias.

También mostramos que las probabilidades di-nucleótido en secuencias donores contienen información relevante al linaje evolutivo. Con la ayuda de nuestros modelos ajustados, fuimos capaces de identificar conjuntos minimales de patrones de interacción que permitían recuperar, a un dado nivel de regularización, las frecuencias de dos sitios observadas en el conjunto anotado de secuencias. El análisis de estas interacciones a lo largo de las especies nos permitió identificar patrones que diferencian a las plantas, los animales y los hongos. Estas firmas en composición de secuencias estaba em-

bebida en las interacciones de dos sitios, en especial aquellas que involucraban a los últimos sitios del intrón, indicando que podrían estar relacionadas con ciertos rasgos taxón-específicos de las primeras etapas del spliceosoma y de sus pasos catalíticos.

Con este capítulo vimos un ejemplo práctico de cómo las interacciones entre pares pueden revelar un orden paralelo a los mecanismos bioquímicos subyacentes en el proceso de splicing. A continuación se utilizará esta racional para un problema distinto: la caracterización de moléculas asociadas a drogas de uso biomédico y sus interacciones con distintas proteínas. Se ampliará el problema de las relaciones entre pares a un análisis más complejo, por lo que emplearemos conceptos más avanzados de redes complejas para observar los patrones contenidos en los datos.

Capítulo 3

Consolidación de datos bioquímicos heterogéneos

El proceso de splicing, examinado en el capítulo 2, conforma una parte importante de la maquinaria celular para la manufactura de proteínas. Como explicamos en la sección 1.1, las proteínas son el bloque fundamental mediante el cual se construyen las estructuras de la célula y de sus organelas, y sus interacciones dan el soporte que sostiene la amplia gama de procesos moleculares que da lugar a la vida.

Una proteína interactúa no solo con otras proteínas, sino también con moléculas pequeñas, sea como parte de una vía metabólica o por una función de señalización. Esto nos lleva a dos de las preguntas centrales de este capítulo: las interacciones que una proteína tiene con un dado conjunto de moléculas, ¿qué dice de ella? Si dos proteínas interactúan en gran parte con las mismas moléculas, ¿qué implica esto acerca de la similaridad estructural o funcional de la proteína?

Se pueden postular preguntas similares por el lado de las moléculas. Cuando un dos ligandos tienen una alta similaridad química entre ellos, y uno posee interacciones conocidas con una proteína, ¿se puede inferir algo acerca del otro ligando?

Esto marca una posible vía de investigación para la búsqueda racional de fármacos, estableciendo ciertas proteínas como blancos terapéuticos e identificando moléculas con las cuales drogar esas dianas. En el capítulo 4 abordaremos ese problema, pero antes, en este capítulo adquiriremos una familiaridad íntima con los datos que fundamentan estos análisis. Revisaremos qué paralelos se obtienen al analizar de distintas formas los vínculos entre drogas y targets. Y, con una vista integral de esta panorámica, realizaremos las observaciones que consideramos más pertinentes para consolidar esta gran diversidad de datos.

3.1. Metodología

Nuestro análisis se sostiene sobre una base de conocimiento heterogéneo. Conviven en paralelo nociones de similaridad estructural química, información de interacciones conocidas droga-target, y agrupamientos de proteínas según criterios estructurales o funcionales. En esta sección daremos la base de ellas, como se obtuvieron los datos, y qué significan para la investigación que sigue.

3.1.1. Medidas de similaridad estructural química

Las interacciones entre una droga y una proteína se rigen por un panorama fisicoquímico complejo entre ellas, incluyendo tanto la estructura intrincada de la proteína como las propiedades químicas de la droga. Las últimas vienen en una variedad enorme, entonces, ¿cómo se establece a dos de ellas como “parecidas”?

El primer criterio contemplado es la similaridad de Tanimoto, considerada una de las medidas más apropiadas para la comparación estructural entre moléculas [52]. El concepto detrás de la misma se puede visualizar en figura 3.1. Allí vemos dos moléculas con una serie de grupos en común, y una diferencia notable: el ejemplo superior posee un grupo ciclohexano que en el inferior permanece incompleto.

El criterio de Tanimoto establece la similaridad en base a un conjunto de subestructuras de la molécula. Más específicamente, examina una serie de “fingerprints” presentes en las moléculas de interés[53], comparándolas en un enfoque “todas contra todas” empleado también en áreas como taxonomía de plantas [54]. Estas huellas moleculares incluyen información de estructura y características fisicoquímicas como los tipos de uniones interatómicas presentes. En nuestro caso se utiliza una base construida como parte de un trabajo previo con el programa **checkmol**[55][56].

En el ejemplo de la figura 3.1, podemos simplificar y considerar como “huellas” una cadena hexano, un grupo amina, y un grupo metilo en las mismas posiciones. Se considera el conjunto g_A de grupos moleculares presentes en la droga A, y g_B para la droga B, y calcula un índice de Jaccard entre ambos:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3.1)$$

o sea, intersección sobre unión. Se entiende a partir de esto que si las dos moléculas no comparten ningún grupo, su similaridad de Tanimoto será nula (ya que el denominador seguirá siendo finito, al ser A y B de tamaño no nulo). Por el lado contrario, si comparten todos sus grupos, entonces $A = B$, entonces $J(A, B) = 1$. El umbral mínimo de uso para una similaridad de Tanimoto significativa es 0,8, establecido como consenso en el campo de la quimioinformática [55].

Cabe preguntar, entonces, ¿qué nivel de detalle ofrecen las estructuras contempladas en la medida de Tanimoto? Por un lado llegan a la escala de los grupos identificados

en la figura 3.1. Por el otro, no distinguirán entre isótopos de un mismo elemento. Y, además, no distinguirán estereoisomería (enantiómeros, diaestereoisómeros, etc.). Esto implica un detalle importante para el propósito de las drogas: la similaridad de Tanimoto ignora la quiralidad. En muchos casos conocidos, incluyendo al metabolismo de la sacarosa, un estereoisómero es metabolizado de forma distinta a otro. Para la interacción de un ligando con una proteína, no cuesta ver por qué podría suceder esto: aquellas moléculas que encajan en un dado sitio fisicoquímico de una proteína cuando tienen una configuración, posiblemente no lo hagan en otra.

De esta manera y otras, el criterio de Tanimoto identificará como “idénticas” (similaridad 1) a pares de moléculas que no lo son. Esta limitación se acepta como parte del criterio, reconociendo que luego de cualquier análisis que realicemos, tendremos que dar cuenta de las diferencias que puedan tener, desapercibidas por Tanimoto. Al conjunto de drogas consideradas idénticas según el criterio de similaridad se las agrupa, y a ese grupo se lo denomina un *cluster identitario* (CID) de Tanimoto. Cada miembro de un cluster identitario tendrá los mismos grupos identificados como subestructuras de la molécula, y compartirán las mismas similaridades con las demás moléculas, no idénticas, del conjunto completo de datos. Se pueden representar entonces a todas esas drogas en un mismo cluster identitario, y así también asignar las similaridades de sus miembros individuales al cluster entero, de forma colectiva.

Cuando se consideran relaciones asociadas a valores de Tanimoto menores a la unidad, se comienza a abarcar un panorama más extenso. Entre los miembros de dos clusters identitarios distintos (y, por ende, entre los dos clusters mismos) puede existir una similaridad en el intervalo $[0, 1)$. Desde estas relaciones simples de pares comienza a surgir una estructura más grande: una red compleja de vínculos pesados no dirigidos (lo último se ve por la simetría en los operadores de intersección y unión).

Un ejemplo de la estructura resultante se puede visualizar en la figura 3.2. Allí además ilustramos en color distinto las diferentes comunidades detectadas por un algoritmo de detección de comunidades de Louvain[58]. Notamos entonces una estructura modular en esta red, donde las moléculas se asocian primero con un subconjunto de las más similares, y luego con otras comunidades que se agruparon también de la misma manera. De esto resulta una pregunta que responderemos en las próximas secciones: ¿qué características moleculares en común dan lugar a las similaridades preferenciales entre las moléculas miembro de distintas comunidades?

Una segunda medida de similaridad explorada es la asociación por subestructura. Este criterio separa a una molécula en sus estructuras características (“fingerprints”) y las plantea como nodos de un grafo. Repite este mismo proceso para otra molécula, y compara los dos grafos resultantes. Si el primer grafo es subgrafo del segundo, se considera que la primera es subestructura de la segunda [56]. Esta comparación heurística evita que el problema computacional se vuelva impráctico para las cantidades de información en la base de datos.

La ilustración en la figura 3.3 ayuda a explicar esta idea. Una conclusión directa

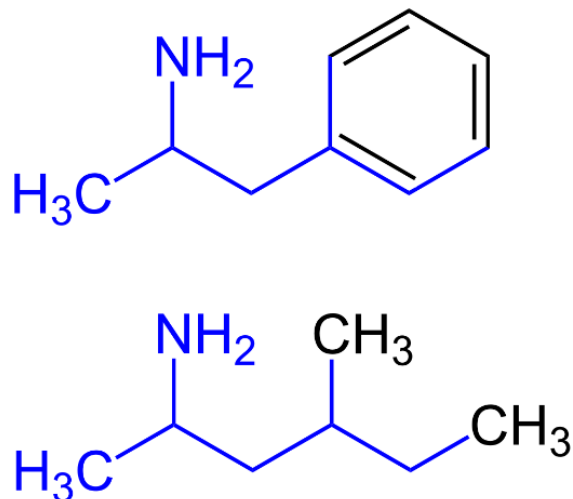


Figura 3.1: Dos moléculas con similaridad de Tanimoto significativa. En azul se notan las componentes en común que poseen. Imagen utilizada bajo licencia CC BY-SA 3.0, agradeciendo la contribución original[57].

de este tipo de vínculo es que la red completa debe ser acíclica: si la molécula B es subestructura de la molécula A, y la molécula C es subestructura de la molécula B, entonces por transitividad C debe ser subestructura de A, no viceversa. Esto se cumple, pero por simplicidad en nuestra red obviamos los saltos a más de un paso (por ejemplo, de A a C) cuando se puede explicar en cambio por una serie de pasos intermedios (de A a B, y de B a C). En el caso de una red por subestructura, se considerarán vínculos pesados según la diferencia en sus pesos moleculares ($A_{ij} = w_i/w_j$). Cuanto más similares los pesos, más cercana se considera la molécula a su subestructura.

En tercer lugar tenemos la similaridad por esqueleto molecular en común, o *scaffold* molecular. Los scaffolds se definen por el criterio de Bemis-Murcko[60][61], y calculados de forma computacional por la implementación pre-existente **scaffoldgraph**[62]. Se puede ver en la figura 3.4 un ejemplo de scaffold molecular que comparten dos drogas distintas, con variaciones en una estructura periférica.

El esqueleto sobre el cual se construye un fármaco tiene una gran importancia en su diseño, afectando de forma crucial la habilidad que una molécula tiene para interactuar con un dado sitio de una proteína. Sin embargo las estructuras periféricas también tienen una gran relevancia, por lo que el panorama de diseño de drogas es más complejo que cualquiera de estas partes. En este contexto, las tres medidas de similaridad planteadas (Tanimoto, subestructura y scaffold) se complementan entre sí. Cada una ofrece una visión parcial del mismo fenómeno: las características químicas de un dado fármaco, y cómo a través de ellas dos drogas se pueden parecer. Cada una también permite considerar una red compleja de características diferentes: la red por similaridades de

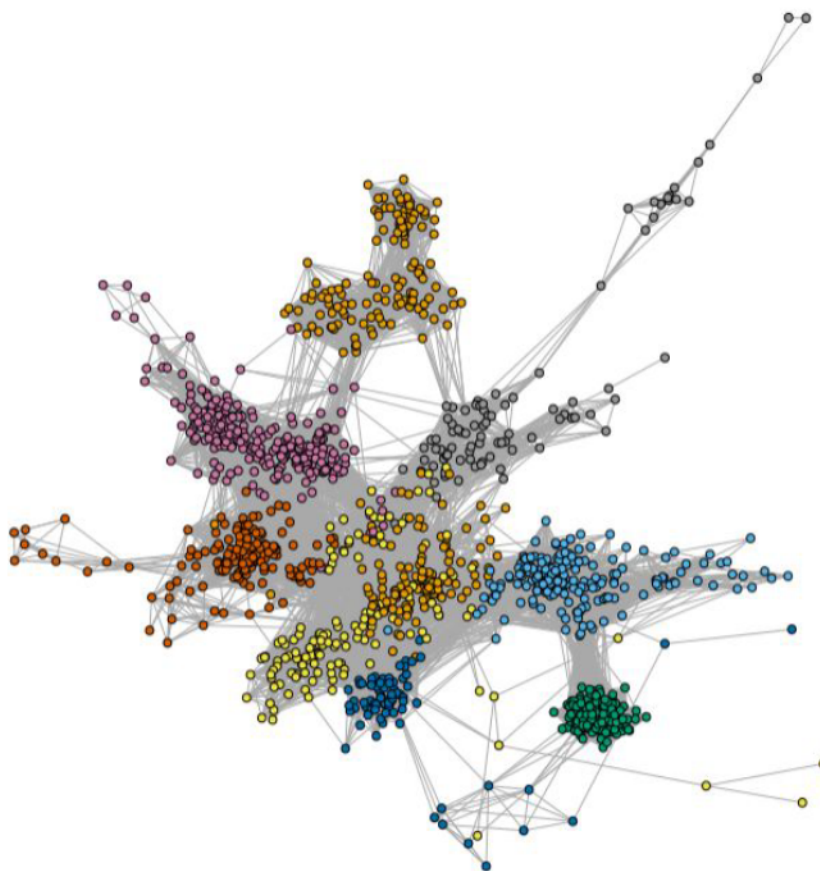


Figura 3.2: Una sección de la red compleja formada por las similitudes de Tanimoto entre moléculas, considerando similitudes mayores a 0,8. Cada color distinto representa una comunidad distinta de la red, detectada por algoritmo de Louvain. Desde esta visualización resulta evidente la estructura modular de la red por asociación de similitud de Tanimoto.

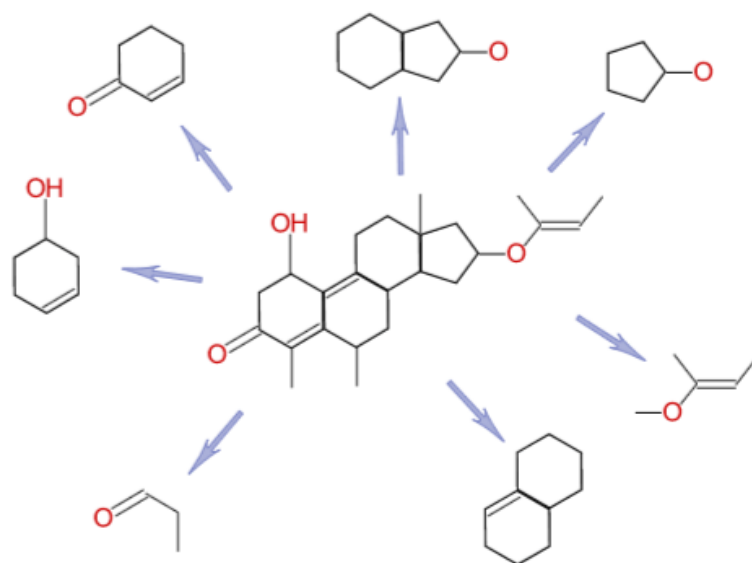


Figura 3.3: Descomposición de una dada molécula en sus varias subestructuras. Cada una de ellas puede también dividirse aún más, formando una red dirigida con estructura de árbol (acíclica salvo por casos de identidad). Imagen cortesía de la documentación de OpenBabel [59].

Tanimoto tiene una estructura altamente modular, mientras que por subestructura se obtiene una red dirigida acíclica y por scaffolds en común se obtendrá una red de co-pertenencia, con la alta densidad de cliques correspondiente.

Estas tres medidas de similaridad toman en cuenta conceptos muy diferentes para caracterizar la cercanía entre drogas, y sin embargo es de esperar que tengan un grado de complementariedad. Analizaremos esta idea en la próxima sección, comparando los clusters identitarios de las redes de Tanimoto y subestructura, evaluando la relación de similaridad entre distintos clusters para ambas medidas, y observando un paralelo entre la jerarquía de scaffolds moleculares y la estructura modular de la red de Tanimoto.

3.1.2. Consistencia entre medidas de similaridad química

La mayor parte de los clusters identitarios contienen una sola molécula: esto no es sorprendente, ya que la identidad, uno espera, debería reflejar una relación con sí misma y nadie más. En la práctica, sucede que existen clusters identitarios de tamaño mayor a uno, fenómeno que debemos analizar para los criterios de similaridad basados en Tanimoto y subestructura por separado.

En la figura 3.5 mostramos dos distribuciones importantes para la caracterización de los clusters identitarios de Tanimoto y subestructura. El panel izquierdo muestra la

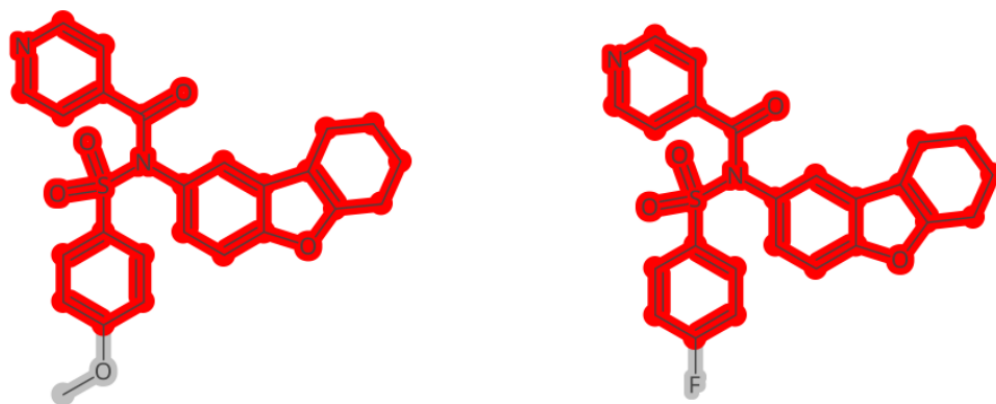


Figura 3.4: Dos moléculas con un scaffold grande en común (rojo). A diferencia de la medida de Tanimoto, el esqueleto de las moléculas se construye alrededor de las estructuras de anillo, por criterio de Bemis-Murcko. Cambiar pequeñas estructuras periféricas como en este ejemplo puede tener grandes efectos a la hora de interactuar con proteínas.

distribucion de sus tamaños. Como podemos ver allí, la gran mayoría (por al menos un orden de magnitud) de los clusters, sean Tanimoto o subestructura, contienen una sola droga. Existen clusters de mayor tamaño, cuyo número decae fuertemente con la cantidad de elementos que incluyen. Notamos también que la distribución de los clusters de subestructura se sesga más hacia el tamaño 1 y se ve muy disminuida en el resto de los tamaños. Esto también tiene sentido: si bien la similaridad de Tanimoto toma en cuenta una variedad de estructuras presentes en las moléculas comparadas y calcula un coeficiente de solapamiento entre los conjuntos, la medida de subestructura es mucho más estricta, exigiendo que una de las moléculas sea directamente parte de la otra. Esto le da a la medida de subestructura muchas menos maneras de ver equiparadas a dos moléculas distintas.

Por otro lado, el panel izquierdo de la figura 3.5 nos muestra una medida de persistencia de clusters identitarios. La persistencia, en este caso de Tanimoto, se calcula como la proporción de clusters Tanimoto que ven a su conjunto de drogas totalmente contenido dentro de un cluster de subestructura. Lo inverso aplica a la persistencia de subestructura. En el gráfico ilustramos la persistencia cuando se segmentan los cluster identitarios según tamaño. Podemos ver que todos los cluster de tamaño 1 se corresponden completamente con algún cluster de la red opuesta, dando la máxima persistencia de 1. Pero, en el caso de Tanimoto, apenas se llega a los clusters de tamaño > 1 se ve que solo aproximadamente un cuarto de los clusters se ve contenido por uno de subestructura. Esta tendencia es consistente a lo largo de los distintos tamaños, sugiriendo que no depende de esta variable, sino que por lo general esa será la proporción de clusters identitarios de Tanimoto que se ven enteramente incluidos por uno de subestructura. En

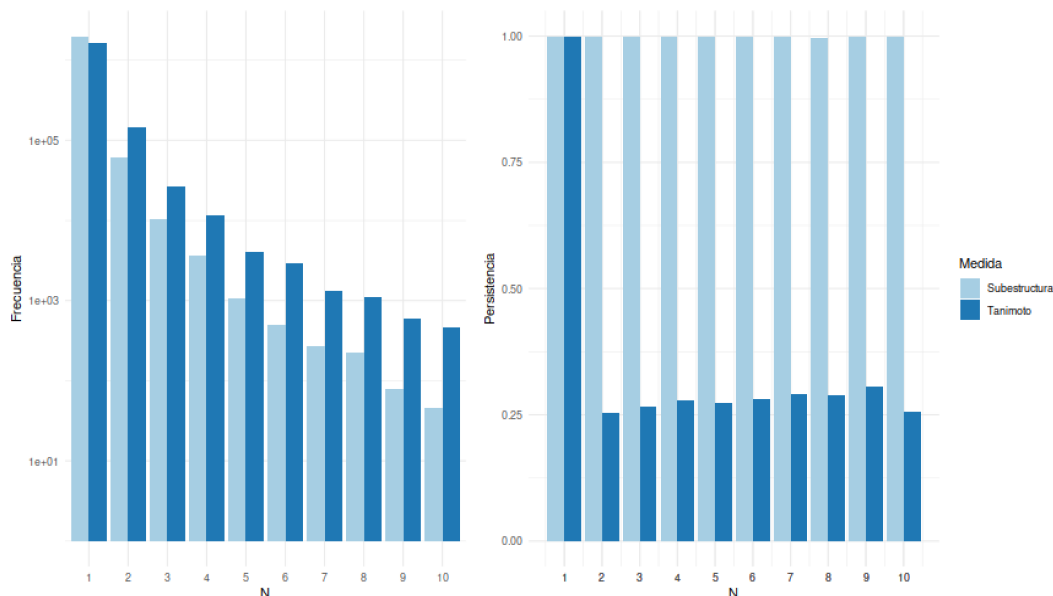


Figura 3.5: **(Izquierda)** Distribución de tamaños de clusters identitarios bajo la medida de subestructura y Tanimoto (eje y logarítmico). **(Derecha)** Cambio en la proporción de los clusters identitarios que demuestran persistencia, en base a su tamaño.

el caso de subestructura, por otro lado, la persistencia a distintos tamaños se mantiene muy cercana a 1. Esto se debe a que el criterio de identidad por subestructura es más exigente: aquello que está junto según subestructura, tiene una chance mayor de estar junto según Tanimoto, mientras que la implicancia recíproca es mucho más débil.

Cuando se extiende el análisis más allá de los cluster identitarios, el próximo paso es considerar qué tan similares son las drogas contenidas en diferentes clusters identitarios. Estas interacciones de a pares acaban por formar estructuras complejas de red, como la observada en la figura 3.2. La red formada por estas interacciones es enorme, con más de siete millones de nodos. Sin embargo en esta figura nos restringimos a una comunidad de esta red, identificada según el algoritmo Louvain de detección de comunidades. Ella es separada nuevamente en sub-comunidades según el mismo algoritmo, revelando una estructura jerárquica y modular en la red de Tanimoto.

Además de las relaciones entre Tanimoto y subestructura, resta abarcar la complementariedad de los scaffolds moleculares con estas medidas de similitud química. Ya que la red de similitud entre drogas por scaffolds en común conforma una red de co-pertenencia, se puede ver al análisis de scaffolds como una clasificación de nodos. Una pregunta pertinente es si los nodos de la red de Tanimoto tienden a agruparse de una forma que sugiera la misma clasificación que por scaffolds. Con el propósito de visualizar y responder esta pregunta, se tomó otra comunidad de la red, subdividida nuevamente en comunidades como en la figura 3.2, y se compararon en base al scaffold más grande compartido por todas sus drogas contenidas. El resultado es la figura 3.6,

donde mostramos esta sección de la red, dividida por color según comunidad, con los scaffolds característicos (o sea, presentes en todas las moléculas del grupo) de cada una a su lado. Una observación inmediata es que las drogas de una comunidad, por más numerosa, pueden llegar a compartir scaffolds de gran tamaño. Más aún, si las comunidades son cercanas en términos topológicos en red, entonces en muchos casos los scaffold característicos que les corresponden también compartiran sub-scaffolds entre ellos, es decir, subestructuras características en común. De esta forma revelamos una relación coherente entre la similaridad por scaffolds y por medida de Tanimoto: la red de Tanimoto se organiza de manera jerárquica, por comunidades y subcomunidades; de manera inversa, a medida que se quitan estructuras periféricas en las moléculas (el proceso de cálculo de scaffolds Bemis-Murcko), se encuentran scaffolds que caracterizan comunidades en la red de Louvain. Cuanto más veces se itera la separación en scaffolds, más pequeñas son las estructuras resultantes, pero más comunes a lo largo de comunidades más grandes. Vemos entonces que la jerarquía en especificidad de scaffolds hace eco de la jerarquía comunitaria de la red de Tanimoto.

Este análisis de comunidad no tiene su paralelo con la red de subestructura, ya que la misma es acíclica. Según los conceptos explicados en la sección 1.2, no posee triángulos, y sin ellos, la modularidad lograda por un análisis en comunidades es baja. Como consecuencia, las mismas no reflejan una estructura significativa.

Vemos en resumen una coherencia entre las distintas medidas de similaridad quimiinformáticas entre drogas, y una riqueza particular de información en la medida de Tanimoto, que emplearemos en más detalle. Esta es solo una parte del rompecabezas: por otro lado, debemos explorar las relaciones entre drogas y proteínas, el foco y centro tanto de este capítulo como del próximo.

3.1.3. Interacciones droga-target

Como introdujimos en la sección 1.1.1, existen configuraciones en las cuales una molécula puede interactuar con una proteína, ocupando una posición adyacente a su estructura. Estas interacciones son de enorme interés bioquímico y farmacológico, y se pueden explorar y corroborar de diversas maneras. En todas, el objetivo es determinar si la droga se ha ligado al target de forma que su actividad normal se vea afectada. Cuando se busca una interacción de este tipo, se llama “ligando” a la molécula pequeña interactuante, y “target”, o “blanco” a la proteína de interés.

Un ejemplo es el ensayo enzimático. En este procedimiento se prepara una solución con una dada concentración de la droga. Luego se agrega progresivamente la proteína de interés, en este caso una enzima. Después de un dado tiempo, se mide nuevamente la concentración de la droga en la solución. De verse reducida en una cantidad significativa, se considera que la molécula se liga a la proteína.

Otro tipo de ensayo es la cristalografía por rayos X. Como resulta obvio desde el nombre del método, para poder realizar este procedimiento es necesario primero lo-

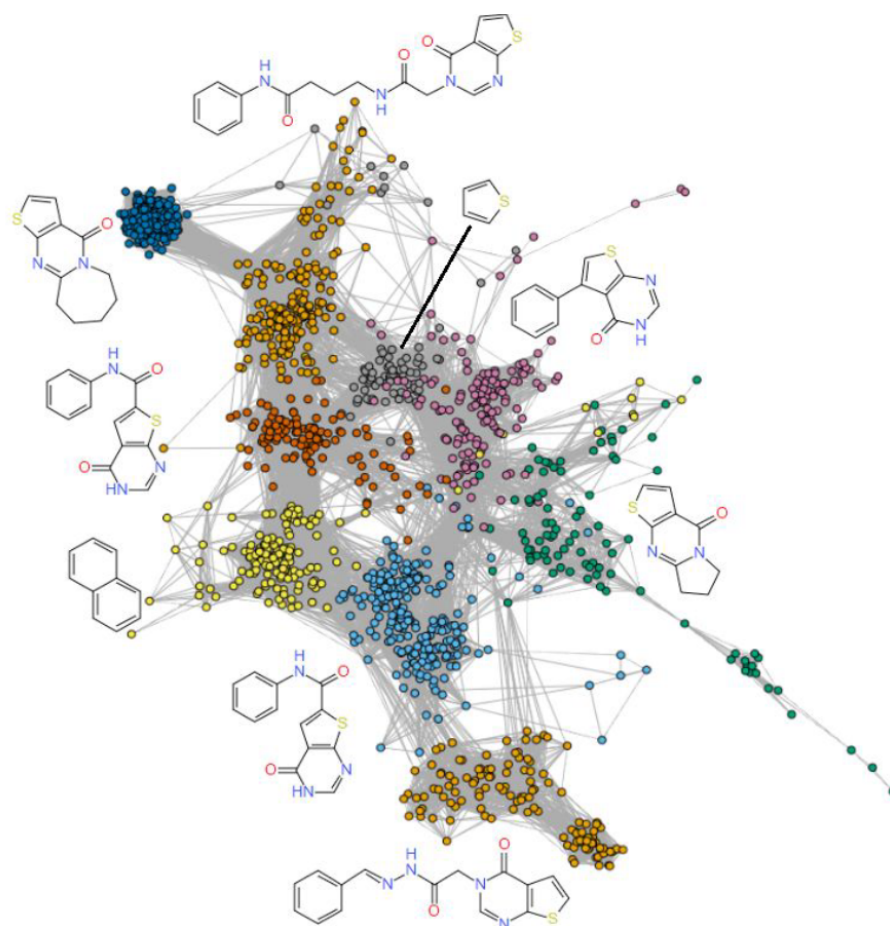


Figura 3.6: Una comunidad de la red por similitud de Tanimoto, separado en comunidades detectadas al correr el algoritmo de Louvain nuevamente dentro de sí misma. Se señalan los scaffolds más grandes detectados para todas las moléculas contenidas en una comunidad. La mayoría de los scaffolds característicos de las comunidades comparten un sub-scaffold entre ellos; sin embargo existen excepciones, como podemos ver en las comunidades amarilla y gris.

grar cristalizar a la proteína de interés en complejo con la droga estudiada y cualquier otro ligando presente en la solución. Esto es algo extremadamente complejo de lograr, dependiente de la temperatura, las concentraciones, y diversos factores que deben ser controlados con toda la precisión posible. Si se logra establecer el complejo y cristalizarlo, luego se realiza la técnica cristalográfica de rayos X como se conoce en la física. Al final del proceso se cuenta con lo que es considerado la evidencia más robusta y confiable de la interacción entre una proteína y un ligando: la estructura cristalina en tres dimensiones de la macromolécula de la proteína junto a la molécula pequeña de la droga, y posiblemente otros ligandos.

Así como estos dos métodos, existen diversos otros, todos con mayor o menor confiabilidad, pero con una limitación en común: el que una proteína y una droga interactúen en condiciones *in vitro* no garantiza un efecto terapéutico cuando se administra a un ser vivo. Este otro paso depende de la disponibilidad del fármaco en el individuo, del metabolismo, y de un panorama mucho más complejo al que se explorará en este capítulo. Nuestros análisis se limitarán a la observación de las interacciones entre drogas y targets en condiciones de laboratorio, reconociendo que para llevarlo al reino práctico de las aplicaciones terapéuticas se depende de un recorrido mucho más largo.

La base de datos empleada en el marco de esta tesis proviene de **TDR Targets**[56]. Para recopilar esta base, Uran et al incorporaron bioactividades (información de interacción droga-target) de otras bases de datos como **ChEMBL**, conservando los datos de cada ensayo considerado, que incluye la inhibición (u otro efecto) sobre la proteína y la concentración con la cual se observó el efecto. Para una dada interacción entre una droga y una proteína target, pueden existir varios ensayos contribuyendo evidencia diversa. Esta evidencia fue curada como parte del trabajo de **TDR Targets**, llevando las interacciones a una clasificación ternaria de “activo”, “inactivo”, e “incierto”. Este procedimiento llevó a un número de seis millones de actividades droga-target curadas, de confianza.

Aún así, esto es una parte crucial en el flujo de trabajo farmacológico: sin predecir nuevas interacciones droga-target, sería muy difícil abarcar la recomendación de fármacos existentes para nuevos propósitos. Más aún, es necesario un enfoque de sistemas al realizar este análisis: se deben considerar las diversas medidas de similaridad química mencionada, se debe contar con un cuerpo de interacciones droga-target corroboradas de buena confianza, y además se deben incorporar criterios funcionales y estructurales acerca de las mismas proteínas. Esto presentaremos en la próxima sección, planteando medidas de similaridad entre proteínas que ayudarán al flujo de información en nuestra metodología.

3.1.4. Similaridad entre proteínas

Como vimos, las proteínas son macromoléculas complejas, como la que ejemplificamos en la figura 1.3. Es posible reconocer entre ellas relaciones basadas en similaridad e

secuencia y/o estructura 3D que hace posible definir familias de proteínas que muchas veces se puede suponer como producto de un linaje en común.

La base de datos PFAM (Protein FAMily) de InterPro propone la identificación de dominios protéicos a partir de un análisis de familias de proteínas[63][64]. A través de alineaciones múltiples de secuencias aminoacídicas de proteínas, y utilizando modelos de Markov de variable oculta (hidden Markov models, o HMM), se reconoce la presencia de regiones específicas compartidas denominadas dominios PFAM. La identificación de dichos dominios en numerosas proteínas habla del valor evolutivo de dichas estructuras, muchas veces relacionado con un rol asociado a una función biológica en particular. En nuestro análisis, utilizaremos las anotaciones PFAM para vincular proteínas a partir de la inclusión común de estos elementos estructurales.

Otro criterio según el cual vincular proteínas viene dado por las anotaciones de ortología. La información para sintetizar una proteína se encuentra codificada en la secuencia de un dado gen. Entre dos proteínas presentes en especies distintas, una relación de ortología significa que cada uno de sus genes correspondientes evolucionó a partir de un gen proveniente evolutivamente de un ancestro en común. Una ilustración del concepto se puede ver en la figura 3.7. Este tipo de relación puede ser útil para transferir información de interacciones droga-target de una especie a otra. Para la misma contamos con información de la base de datos OrthoMCL[65], que calcula relaciones de ortología entre genes a partir de alineaciones de secuencia y clustering Markoviano en red [66].

Ambos tipos de anotación se basan en similitudes de secuencia, con distintos algoritmos empleados para extraer la información allí contenida. Compararemos a los dos como parte de nuestro análisis, pero en última instancia nos apoyaremos en la agrupación PFAM por sus asociaciones con regiones específicas en las secuencias de aminoácidos de las proteínas. Utilizando su capacidad de agrupar las mismas, le podremos dar un significado más puntual a las interacciones de una dada droga (o grupo de drogas) con los targets que comparten por una dada anotación. Contamos también con otro conjunto de datos respecto a la especificidad con la cual un dado fármaco interactúa con una proteína, en base a las regiones significativas (dominios) con los que interactúa: esto discutiremos a continuación.

3.1.5. Interacción específica ligando-proteína

La actividad terapéutica de una droga puede ser caracterizada no solo en términos del conjunto de proteínas con las cuales interactúa, sino los sitios o estructuras significativas dentro de las mismas con los cuales presenta dicha interacción. En el contexto en el cual una molécula se liga a una proteína, se le refiere como “ligando”, y la información de región donde interactuó, sumado con las anotaciones previamente mencionadas PFAM (que toman en cuenta características estructurales relevantes) pueden resultar valiosas a la hora de encarar la tarea de la priorización.

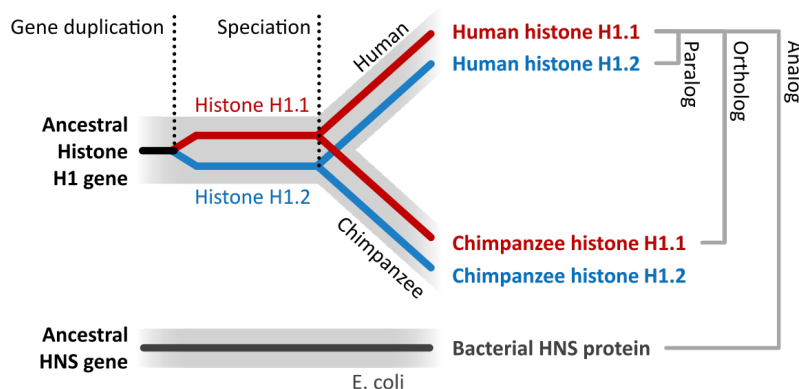


Figura 3.7: Distintas relaciones de similaridad posibles entre genes. Un gen es análogo a otro si cumple un rol similar en distintos organismos, sin partir de un origen en común. En contraste, un ortólogo (rojo con rojo u azul con azul) indica un rol paralelo en distintos organismos, con una relación confiable a un gen antepasado en un ancestro evolutivo en común. Cuando un gen se duplica en un mismo organismo, se lo denomina un parólogo. Imagen utilizada bajo licencia CC BY 4.0, agradeciendo la contribución original [67].

Por desgracia conocer los sitios específicos donde interactúa un ligando es difícil, accesible en pocos casos (por ejemplo, cuando el complejo ligando-proteína exhibe una estructura cristalina y se presta a la cristalografía de rayos X). Estos datos sin embargo existen, y por su valor enorme como un conjunto de validación se los elige incluir dentro de un análisis para relacionar drogas con grupos PFAM.

Se utiliza a la base de datos BioLiP (Ligand-Protein)[68], que proporciona para una cantidad de interacciones molécula-proteína los sitios en la cadena de aminoácidos donde la interacción tomó lugar. Combinando estos índices con los índices característicos de un dominio PFAM según InterPro, fuimos capaces de asociar específicamente drogas a anotaciones PFAM a través de mediciones directas en una base de datos de confianza. Esto será de gran utilidad en la sección 3.3, cuando examinemos las asociaciones posibles entre grupos PFAM y las drogas que típicamente son capaces de afectar las proteínas allí agrupadas.

Esta es una de las varias fuentes de datos que consideramos en este trabajo. El desafío propuesto es hallar patrones en común entre ellas y extraer observaciones significativas, no solo para el propósito de la priorización en red, sino también como pantallazo del mundo del diseño de fármacos y sus interacciones con entes biológicos. A continuación comenzaremos desde esta faceta en particular: compararemos la información contenida en las interacciones droga-target con aquello que puede ser inferido desde las distintas medidas de similaridad química.

Las fuentes de información detalladas en esta sección forman una base de datos heterogénea, cuya estructura se puede describir en términos de diversas entidades y relaciones. Resumimos las características de esas redes en la tabla 3.1, y con esta base

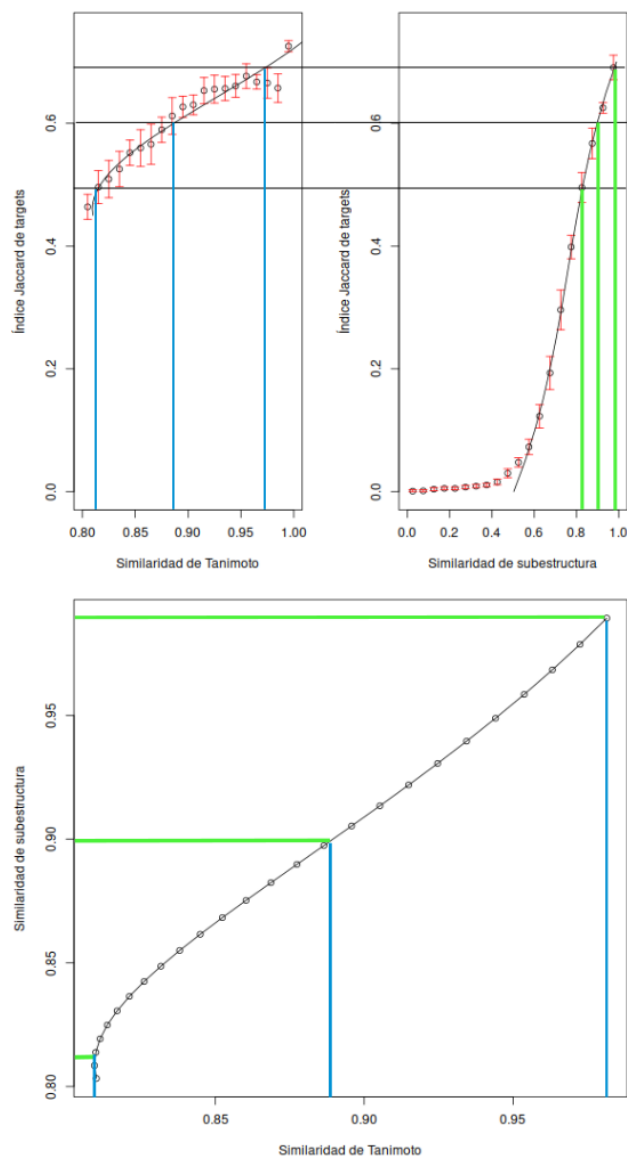


Figura 3.8: **(Izquierda)** Comparación entre un índice de Jaccard cuantificando los targets que dos drogas tienen en común (eje x) y la similaridad de Tanimoto entre ellas (eje y), restringida por encima de 0.8 como es común en este tipo de análisis. **(Centro)** Mismo concepto comparando esta vez con la similaridad por subestructura entre dos drogas (la diferencia en pesos moleculares, sumado al hecho de ser efectivamente una subestructura). En ambos casos se nota que a mayor similaridad química, mayor cantidad de targets en común. **(Derecha)** Relación entre la similaridad por subestructura entre drogas y la similaridad de Tanimoto. Esto indica no solo consistencia entre ambas medidas, sino también que las drogas más similares por Tanimoto tienden a tener pesos moleculares parecidos.

Red	Entidades contenidas	Proveniencia	Nodos	Vínculos	Bipartita
Clusters Tanimoto	Droga/CID	TDR Targets	2178504/1855022	2178504	Si
Clusters Subestructura	Droga/CID	TDR Targets	2178505/2073829	2178505	Si
Tanimoto	CID/CID	TDR Targets	1733336/1733336	24541032	No
Subestructura	CID/CID	TDR Targets	2008740/2008740	193419431	No
Scaffold	Droga/scaffold	scaffoldgraph	3346423/1485495	5351661	No
Interacción	Droga/target	TDR Targets	952699/5767	2638773	Si
PFAM	Target/ anotación	InterPro	332492/10611	449340	Si
Ortología	Target/ anotación	OrthoDB	483502/78184	483502	Si
Target-Especie	Target/especie	TDR Targets	564534/549	564534	Si
Interacción espacial	Droga/target (sitios)	BioLiP	36584/36963	864060	Si

Cuadro 3.1: Resumen de características de redes utilizadas en el análisis integral de datos bioquímicos. Las redes con un solo tipo de nodo muestran un solo número en la columna “nodos”, mientras que en el caso de dos distintos tipos, se da el número del primer y segundo tipo respectivamente. Una red con dos tipos de nodos puede no ser bipartita: la red droga/scaffold, por ejemplo, contiene varias drogas que son asimismo scaffolds de otras drogas. Esta distinción se hace en la última columna.

sentada podemos avanzar hacia nuestro análisis integrativo de los datos.

3.2. Consistencia entre interacciones bioquímicas y similitudes químicas

En el capítulo 4, el objetivo integral será tomar provecho del criterio de “culpa por asociación” en red para, desde interacciones conocidas droga-target, predecir nuevas interacciones posibles. Este criterio, sin embargo, depende de un supuesto no trivial: que los nodos cercanos en una dada red efectivamente manifiesten de forma similar la característica buscada (en este caso, interacción con targets en común). Por esta razón, y como parte de nuestra caracterización de los paralelos entre las redes quimiinformáticas y bioinformáticas, nos es de interés responder la siguiente pregunta: las drogas similares, ¿poseen una cantidad significativa de targets en común con los que interactúan?

Para cuantificar esto en primera instancia se toma el índice de Jaccard de la ecuación 3.1, comparando para dos drogas el conjunto de targets con los cuales cada una interactúa. Obteniendo la intersección y unión de los mismos se calcula este coeficiente, y para cada par de drogas se compara dicho coeficiente con los pesos que obtienen las similitudes de Tanimoto y subestructura a través del vínculo que conecta a ambas drogas como nodos en la red.

Podemos ver el resultado en la figura 3.8, donde se calculó la similitud Tanimoto promedio entre pares de drogas que comparten targets con un dado índice de Jaccard (izquierda), y se repite el cálculo para la similitud por subestructura (centro). Finalmente se grafican las similitudes promedio según ambos criterios, dado el mismo puntaje de Jaccard para el conjunto de pares de drogas en cuestión (derecha). El gráfico correspondiente a la similitud de Tanimoto comienza desde un valor de similitud

0,8, como ya mencionamos que haríamos. En el caso de subestructura, un mayor peso de similaridad indica que los pesos moleculares de ambas drogas son más parecidos. Lo que vemos en ambos casos es una relación creciente en la similaridad química a medida que aumentan los targets en común correspondientes al par de drogas. Esta relación es monótona, dando peso a la idea de que a mayor similaridad química se comparten más targets.

El panel derecho además muestra una correlación importante entre ambas medidas de similaridad, que se puede tratar como una “calibración” a través de los bins de los índices de Jaccard de targets compartidos entre cada dado par de drogas.

Los primeros dos paneles no solo le dan una justificación concreta al criterio de “culpa por asociación” que aplicaremos en el capítulo 4, sino que también nos habla de la información complementaria entre nuestros datos. En el caso de la relación de Tanimoto, lo que este resultado nos dice es que, si dos drogas comparten una mayor cantidad de grupos funcionales característicos entre sí, entonces se vuelve más frecuente que también interactúen con más de las mismas proteínas. Por el lado de la relación de subestructura, en cambio, este resultado nos informa que si una subestructura es de peso molecular más parecido a su superestructura (o sea, se “restó” o excluyó una porción más chica de su estructura para obtenerla), entonces ambas tenderán a interactuar con más proteínas en común. Esto a su vez reafirma la elección que hicimos sobre la ponderación de enlaces de subestructura a partir del cociente de pesos moleculares en juego.

Como hablamos en la sección anterior, distintas proteínas pueden verse agrupadas por características en común, sea una similaridad entre sus secuencias y la presencia de estructuras significativas en común (PFAM), o por estar vinculados por un linaje evolutivo (ortología). Dado que las drogas más parecidas tienden a interactuar con más proteínas en común, lo próximo a explorar es si una dada anotación, conteniendo un dado conjunto de proteínas, puede asociarse con drogas químicamente similares. En las secciones siguientes examinaremos estas ideas, partiendo de una noción de “drogabilidad” de las proteínas correspondientes a una anotación.

3.3. Drogabilidad de grupos PFAM

Como ya mencionamos, las anotaciones PFAM agrupan proteínas según similitudes en sus secuencias y dominios significativos en común. Esto en muchos casos se traduce a una similitud en las funciones que desempeñan en sus respectivos organismos. Desde un punto de vista de redes, un grupo PFAM se vincula con sus proteínas como una red bipartita. En la figura 3.9 se muestra una visualización de las varias capas que vinculan anotaciones PFAM a proteínas, proteínas a drogas a través de sus datos curados de interacción, y drogas a cluster identitarios.

La próxima capa de esta red es la de interacciones droga-target, visible en la misma

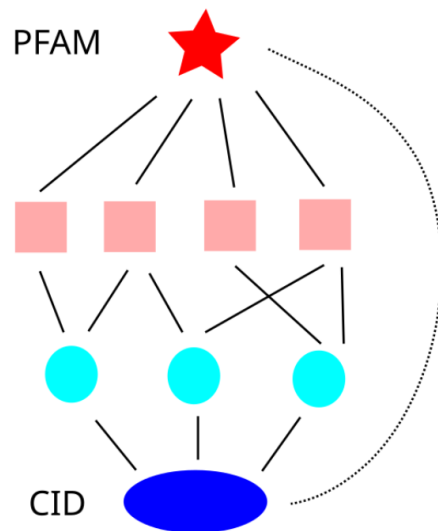


Figura 3.9: Ilustración simplificada del procedimiento para establecer relaciones significativas entre CID (cluster identitario de drogas) y grupos PFAM (denotado por una estrella), a través de la capa química de drogas representadas por un CID (círculos) y las proteínas que interactúan con ellas (cuadrados), contenidas dentro del grupo PFAM de interés. Se consideran los caminos que se puede tomar a través de la red para llegar del CID al PFAM. Cada uno representa una droga característica del CID que interactuó con una proteína contenida dentro del conjunto PFAM. La cantidad de estas interacciones se compara contra una hipótesis nula hipergeométrica para obtener un p-valor de Fisher (luego ajustado por procedimiento de Holm).

figura. Para cada grupo PFAM se contó el número de sus proteínas que tuviese alguna interacción conocida con alguna droga: a esta proteína se le dice entonces “drogable”, y la cantidad de ellas dentro de un grupo PFAM es característica de la “drogabilidad” de dicho grupo.

No es suficiente tener una gran cantidad de proteínas drogables para considerar drogable al grupo PFAM. Por mayor que sea la cantidad, podrían ser una proporción pequeña de todos los targets asociados a la anotación en cuestión. Se usa entonces un test de Fisher para determinar si la drogabilidad de un dado grupo PFAM es significativa. Esto se hace comparando respecto a una hipótesis nula. La ocurrencia de un grupo PFAM no-drogable que tenga asociado por azar un número mayor o igual a $\tilde{k}$ targets que presenten bioactividades sigue una probabilidad hipergeométrica

$$P(k \geq \tilde{k}) = \sum_{k \geq \tilde{k}} \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad (3.2)$$

donde N es el número de proteínas en la base de datos, K es el número de proteínas activas dentro de ellas, n es el número de proteínas contenidas dentro del grupo PFAM, y k es el número de ellas que efectivamente interactúa con alguna droga, y es, por lo

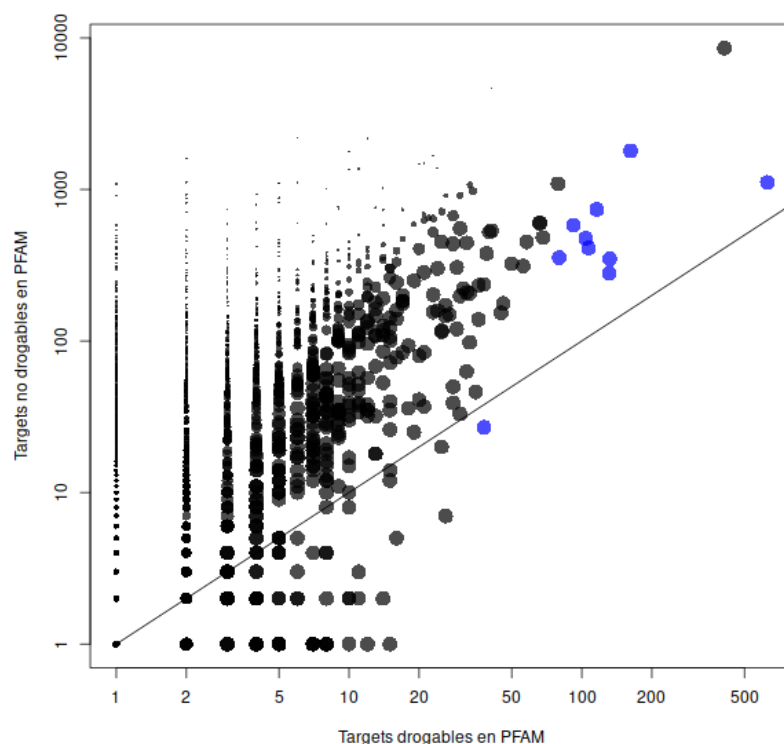


Figura 3.10: Comparación entre la cantidad de targets drogables por al menos una droga (eje x) contenidos en un dado grupo PFAM (círculos) comparado con los targets no drogables allí contenidos (eje y). La línea identidad nos revela que la tendencia es tener más targets no drogables que drogables. Esto ilustra la conveniencia de utilizar la información de un target para otro target dentro del mismo grupo PFAM.

PFAM	Nombre	T. drogables	No drogables	p-valor
PF00001	7 transmembrane receptor (rhodopsin family)	627	1115	0
PF02932	Neurotransmitter-gated ion-channel transmembrane region	131	280	1.36e-107
PF02931	Neurotransmitter-gated ion-channel ligand binding domain	132	348	4.88e-99
PF00104	Ligand-binding domain of nuclear hormone receptor	107	410	8.80e-67
PF00105	Zinc finger, C4 type (two domains)	104	475	1.28e-58
PF00520	Ion transport protein	116	737	4.99e-52
PF01094	Receptor family ligand binding region	80	355	2.84e-46
PF13354	Beta-lactamase enzyme family	38	27	5.88e-45
PF07714	Protein tyrosine and serine/threonine kinase	162	1797	1.07e-42
PF00089	Trypsin	92	582	9.67e-42

Cuadro 3.2: Diez dominios PFAM más drogables, según el análisis estadístico de Fisher. Se incluyen los nombres, relacionados con la función que desempeñan en general las proteínas que los contienen. Una mayor drogabilidad observada no solo refleja sobre qué tan accesible es el grupo de proteínas desde una perspectiva fisicoquímica, sino también acerca de cuánta evidencia se tiene al respecto, según las prioridades de la industria farmacológica.

tanto, drogable. Una probabilidad de obtener una dada drogabilidad (o más) por azar se denomina en este contexto un p-valor. Debajo de cierto umbral, un p-valor indica un criterio suficiente para rechazar la hipótesis nula y considerar significativa, en este caso, la drogabilidad de un grupo PFAM. El umbral más común utilizado es de 0,05.

En la figura 3.10 podemos observar los resultados de este análisis, visualizando a cada grupo PFAM como un círculo de tamaño proporcional al p-valor de drogabilidad, codificando en color azul un resultado significativo. En el eje horizontal vemos el número de proteínas drogables (que están asociadas a al menos una droga) dentro del grupo PFAM, mientras que el eje vertical muestra el número de proteínas no drogables. Lo primero a notar es que cuantas más proteínas drogables tiene un grupo PFAM, más proteínas no drogables también tendrá: esto no es cierto viceversa. Esto nos indica una de dos conclusiones: o bien (a) la mayoría de los targets pertenecientes a grupos grandes nunca se evalúan en ensayos de interacción con fármacos (entonces la no drogabilidad es en realidad incerteza acerca de si es drogable o no), o (b) la mayoría de las proteínas en un grupo PFAM no resultarán drogables. Esto habla de una baja densidad de información de interacciones, una temática que notaremos recurrente en materia de priorización de fármacos.

Se puede ir más allá del p-valor de Fisher, planteando un puntaje no binario con tal de comparar en una escala continua la drogabilidad de distintos grupos PFAM, y pesarlos dentro de contextos no solo descriptivos, sino también de aprendizaje automatizado. Este es el llamado puntaje de relevancia, y resulta sumamente útil en tareas de priorización, donde ya se ha utilizado[55], incluyendo nuestro análisis en la sección 4.3. Se organiza a los grupos en orden ascendente de p-valor, y se le asigna el puntaje máximo de relevancia al 20 % más bajo. Adicionalmente, se considera también un puntaje no nulo para los grupos aunque existan por encima de este umbral, según el criterio:

$$R_j = \begin{cases} 1 & \text{si } p_j \in C_{20\%} \\ \left[\frac{-\log_{10}(p_j)}{\max_j(-\log_{10}(p_j))} \right]^\alpha & \text{si } p_j \notin C_{20\%} \end{cases} \quad (3.3)$$

donde p_j representa al p-valor obtenido para el grupo PFAM j , α arbitrario se fija en un valor de 1, $C_{20\%}$ representa al conjunto de los grupos PFAM con el 20 % más bajo de p-valores, y el segundo caso de la función nos muestra una penalización logarítmica sobre un p-valor más alto (un p-valor unitario, el menos significativo posible, da un puntaje de relevancia nulo), normalizados a lo largo de los distintos grupos de proteínas y elevado a una potencia α arbitraria.

Basándonos en este análisis podemos extender nuestras observaciones a la red entera de la figura 3.9, llegando hasta los cluster identitarios de la capa química de Tanimoto. Seguimos el esquema de la figura de una punta a la otra para relacionar un dado grupo PFAM con las drogas con las cuales interactúan sus proteínas miembro, y luego con el cluster identitario al cual pertenecen (esto último ya que presentan una similitud perfecta de acuerdo a la medida quimiinformática utilizada, y se puede consolidar en

un cluster la información de varias interacciones separadas). Surge entonces la pregunta clave: ¿cuánta evidencia se tiene de que un dado grupo PFAM pueda asociarse significativamente con las drogas de un cluster identitario? Esto es relevante como indicio de que las drogas en dicho cluster podrían estar actuando específicamente en ese dominio de las proteínas sobre las cuales tienen efecto. En otras palabras, podría contener información acerca de dónde en la proteína interactúa un dado ligando.

Se aplica nuevamente el test de Fisher utilizado como parte del cálculo del puntaje de relevancia, pero esta vez se compara el conjunto de proteínas con los cuales una dada droga interactúa, y el conjunto de proteínas contenidas en un grupo PFAM, para así obtener una respuesta acerca de si esa droga interactúa de forma significativamente superior al azar con los miembros de dicho grupo. Nuevamente se utiliza la ecuación 3.2 para calcular la probabilidad de que una distribución hipergeométrica presente el mismo número o mayor de coincidencias $\tilde{k}$. Pero esta vez, N será el número total de proteínas con las cuales el cluster identitario podría interactuar, K es el número que coincidiría con una proteína miembro del grupo PFAM considerado, n es el número de proteínas con las cuales el cluster identitario efectivamente interactúa, y k es el número de ellas que efectivamente forma parte del grupo PFAM en cuestión.

Esto se realizó para todas las combinaciones posibles de grupo PFAM con clusters identitarios, teniendo la precaución de solo evaluar los casos donde existe alguna interacción. Los p-valores resultantes fueron ajustados según el método de Holm, ya que controla por tasa de errores de tipo I por comparación múltiple (*family-wise error rate*, o FWER), y tiene una potencia uniformemente mayor al método de Bonferroni.

Podemos ver un ejemplo de conjunto de drogas características en la figura 3.11. Tres de ellas (hileras superior) comparten un mismo scaffold. Del resto (inferior) las primeras dos se basan en esqueletos similares, pero no idénticos. La tercera es totalmente distinta, e incluso representa a su propio scaffold, que tiene ciertas similitudes con el esqueleto de las primeras tres moléculas mencionadas (superior). En general cabe destacar las similitudes en estructura química subyacente entre algunas de ellas. Puede resultar significativo a un PFAM un conjunto de drogas con scaffolds variados, como vemos en el ejemplo. Sin embargo es común que varias de ellas compartan un esqueleto molecular, o bien que los esqueletos mismos tengan ciertas similitudes entre ellos. Este ejemplo ilustra la idea detrás de nuestros próximos análisis: evaluar si la similitud química entre clusters identitarios tiene una correspondencia con haber sido asociados de forma significativa a un grupo PFAM en común.

Como el primero de estos análisis, tomamos para un dado grupo PFAM las drogas características y evaluamos las similitudes químicas de Tanimoto que demuestran entre ellas. Repitiendo esto para todos los grupos se obtiene una distribución de probabilidad de similitudes intra-PFAM, como se ve en la figura 3.12. Esta distribución se compara con las similitudes cuando los cluster identitarios no son característicos de un mismo grupo PFAM, tomada como distribución de probabilidad nula. En la figura queda cualitativamente en evidencia la diferencia entre ambas distribuciones, y

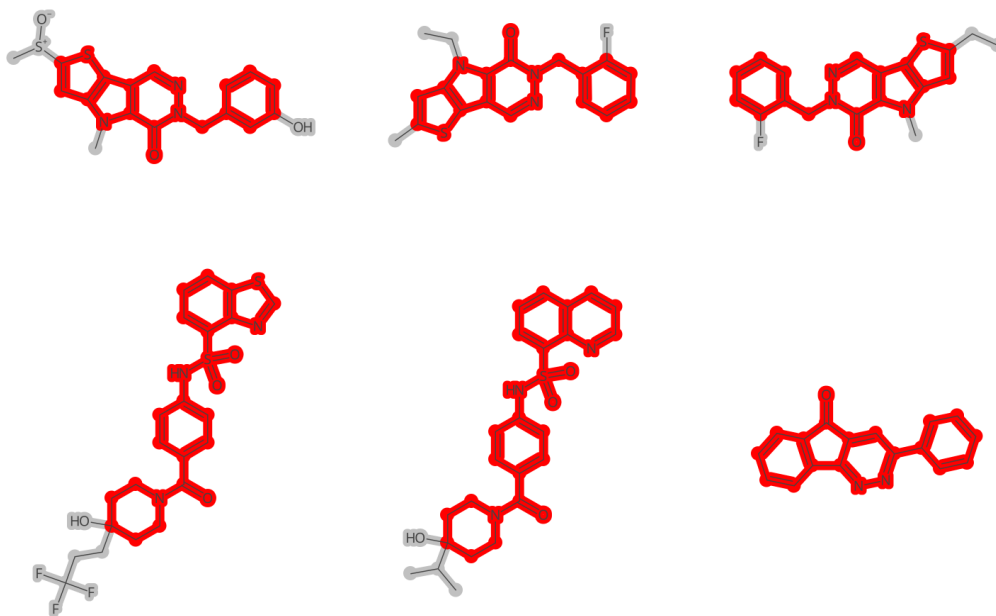


Figura 3.11: Ejemplo de drogas significativas halladas para el grupo PF00224 (“Pyruvate kinase, barrel domain”), con scaffold molecular marcado en rojo.

al proceder con un test no paramétrico de Kolmogorov-Smirnov, se obtiene un p-valor muy por debajo de 0,05, indicando que la diferencia entre ambas distribuciones no sucedió por azar, y se rechaza la hipótesis nula de que la distribución de similaridades de Tanimoto entre drogas características de un mismo grupo PFAM siga la misma estadística que las que no se ven relacionadas por un mismo grupo. Esto nos indica que la similaridad química de Tanimoto no solo privilegia clusters identitarios cuyas drogas tienden a interactuar con targets similares, como vimos en la sección 2.2, sino que esta relación se mantiene cuando se agrupa proteínas por dominios en común según la base de datos PFAM, a tal punto de que una prueba estadística corrobora la significancia de la asociación. Con cierta confianza entonces podemos proponer que la presencia de una anotación PFAM en común entre proteínas puede utilizarse como indicio que sufre que los mismos podrían ser blanco de drogas similares.

Sigue la incógnita, sin embargo, de la relación inversa: si dos drogas son químicamente similares según Tanimoto, ¿qué dice eso de que interactúen con proteínas agrupadas según grupos PFAM en común? Para hacer esto tomamos la red como vista en la figura 3.9 y planteamos para cada capa una matriz de incidencia, ya que, entre una y la otra, la red construida es bipartita. El producto matricial de las matrices de incidencia, desde la capa de clusters identitarios hasta la capa de grupos PFAM, nos da una matriz de incidencia efectiva. Multiplicando la misma por su transpuesta nos da una matriz de

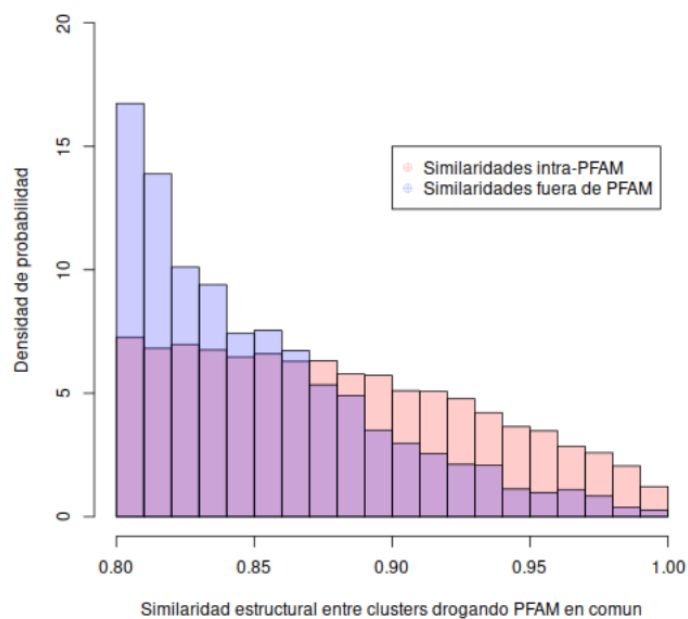


Figura 3.12: Histograma de la probabilidad de obtener un par de drogas con una dada similaridad de Tanimoto entre ellas, cuando ambas tienen una relación significativa con un dado PFAM (rojo) y cuando no tienen una relación significativa con un PFAM en común (azul). Un test de Kolmogorov-Smirnov para comparar ambas distribuciones entrega un p-valor debajo de 0,05, indicando que las drogas con PFAM significativos en común tienden a mostrar una mayor similaridad química de Tanimoto entre ellas.

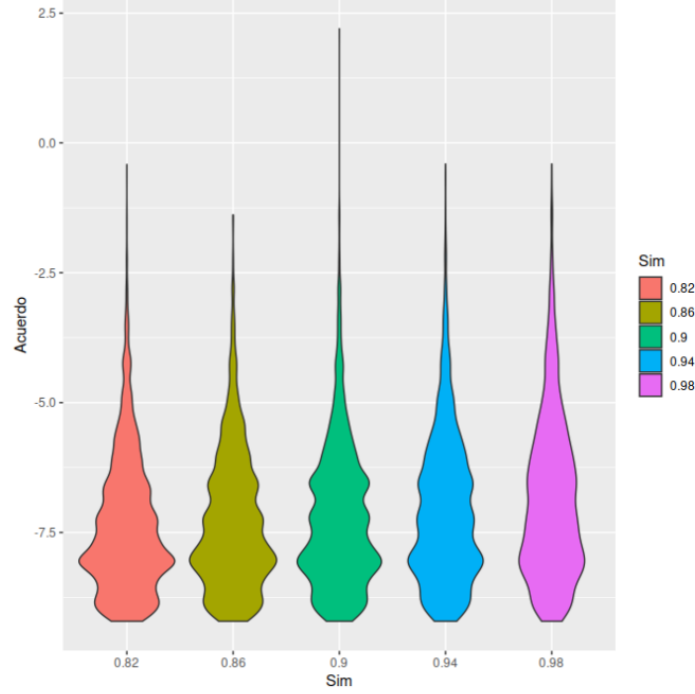


Figura 3.13: Distribución del acuerdo PFAM definido, en escala logarítmica, para distintos bins de similaridad de Tanimoto entre drogas. La tendencia es ligera, pero apunta a una mayor varianza y mayor media en el acuerdo PFAM a medida que se toman pares de drogas de mayor similaridad estructural.

adyacencia entre clusters identitarios, a través de sus posibles relaciones con grupos PFAM en común:

$$\overline{\overline{A}}_{CID}^{(PFAM)} = \overline{\overline{B}}_{CID \rightarrow D} \overline{\overline{B}}_{D \rightarrow P} \overline{\overline{B}}_{P \rightarrow PFAM} \overline{\overline{B}}_{P \rightarrow PFAM}^T \overline{\overline{B}}_{D \rightarrow P}^T \overline{\overline{B}}_{CID \rightarrow D}^T$$

donde $\overline{\overline{B}}_{CID \rightarrow D}$ vincula un cluster identitario a las drogas que lo componen, $\overline{\overline{B}}_{D \rightarrow P}$ expresa las interacciones entre una droga D y una proteína target P , y $\overline{\overline{B}}_{P \rightarrow PFAM}$ indica la pertenencia de una proteína P a un dado grupo PFAM.

La matriz de adyacencia resultante cuantifica por cuántos caminos (cuántas drogas, a través de cuántas interacciones con targets, con un mismo grupo PFAM) se ve vinculada una droga a la otra a través de grupos PFAM en común. Al elemento (i, j) de la matriz $\overline{\overline{A}}_{CID}^{(PFAM)}$ lo llamamos el “acuerdo PFAM” entre el cluster identitario i y el j . Cuando comparamos este acuerdo con la similaridad química de Tanimoto entre dos drogas, obtenemos el resultado de la figura 3.13. En todos los niveles de similaridad quimioinformática se ve una distribución concentrada en los niveles bajos de acuerdo PFAM, con un aumento muy gradual. Sin embargo, cuando se muestrea una porción de la población con similitudes más altas entre sí, se tiende a ver que la distribución de

sus acuerdos tiende a estirarse hacia los valores más altos, cobrando más presencia. Esto es una evidencia menos contundente que la anterior, pero acompaña a la relación cercana entre la similaridad química de Tanimoto entre clusters identitarios y las bioactividades con targets en común.

Además de observar la concordancia química entre las drogas, nos podemos preguntar por la concordancia física de las relaciones significativas droga-PFAM, comparada con los sitios concretos donde una molécula se liga a una dada proteína: dato que obtuvimos con la base BioLiP, mencionada en la sección 3.1.5. Esta base se construyó a partir de interacciones vistas con cristalografía de rayos X, considerada de la evidencia más fiable y específica en materia de asociaciones droga-target.

Para validar los resultados obtenidos de drogabilidad específica, consideramos evaluar la significancia de asociaciones cluster identitario a PFAM respecto a la capacidad de otorgar un puntaje más alto a aquellas asociaciones que se encuentren consignadas en la base de datos BioLiP, referida a interacciones verdaderamente observadas entre fármacos y sitios aminoacídicos dentro de un dado dominio PFAM. Para este análisis será preferible un gradiente continuo de confiabilidad, entonces volvemos a emplear el puntaje de relevancia de la ecuación 3.3, que organizará a las interacciones específicas PFAM-droga (cluster identitario de Tanimoto) desde una máxima relevancia de 1 hasta una mínima de 0, comenzando la etapa continua de la función luego del primer percentil 20 de la distribución de p-valores obtenidos por el test de Fisher. Luego podemos evaluar la capacidad que tiene este ranking de puntajes para predecir interacciones verdaderamente observadas entre fármacos y los sitios correspondientes a dominios PFAM. Hacemos uso de la base calculada en la sección 3.1.5 como conjunto de validación de máxima confianza (referido a menudo como “ground truth”). Luego en la figura 3.14 podemos estudiar el rendimiento de nuestro ranking comparado con esta validación.

La curva ROC (“Receiving Operator Characteristic”) compara la ganancia en verdaderos positivos (“TPR, True Positive Rate”) contra la ganancia en falsos positivos (FPR, “False Positive Rate”) a medida que se desciende en el ranking de confiabilidad. Una recta de pendiente 1, con área bajo la curva de 0,5, representaría un filtro tan bueno como el azar, ya que por cada verdadero positivo se obtendrían aproximadamente tantos falsos positivos como estén presentes en la muestra completa. Todo por debajo de ella es peor que el azar. Se calcula el área debajo de la curva (AUC, Area Under Curve) como indicador del rendimiento total del filtro en términos de TPR y FPR: un AUC de 1 indica un filtro perfecto, mientras que 0,5 delata a un “no-filtro”, o sea, azar. En todos estos aspectos, la curva ROC observada en 3.14 indica un buen clasificador.

La curva PR, por otro lado, indica el balance entre la precisión (equivalente a la tasa de verdaderos positivos, TPR) comparado al “recall”, la recuperación, medida como la proporción de los verdaderos positivos que son efectivamente recuperados. Se espera entonces que comience en lo alto, comenzando de una alta especificidad sobre las predicciones correctas, pero una baja proporción de ellas. Luego a medida que desciende por el ranking, sacrifica precisión hasta llegar a un mínimo, donde recupera el

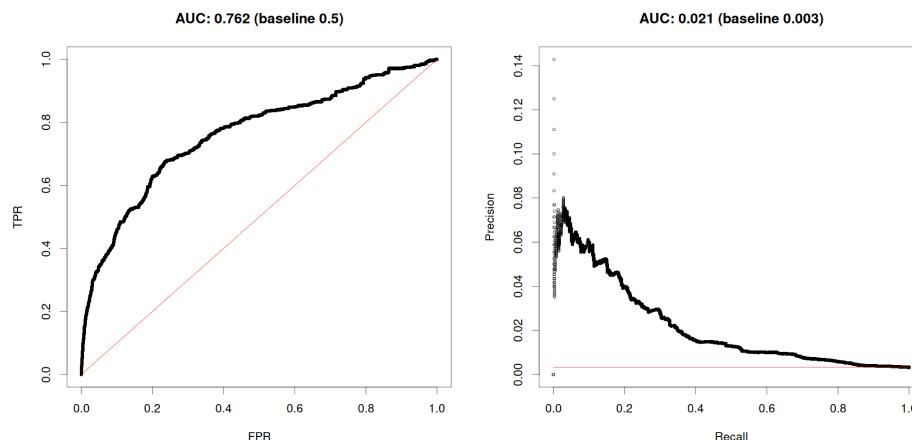


Figura 3.14: **(Izquierda)** Curva ROC y **(derecha)** curva Precision-Recall (precisión-recuperación) para un ranking de interacciones predichas con un test de Fisher para evaluar significancia de la relación PFAM-droga (cluster identitario de Tanimoto). Las interacciones ya no se incluyen de forma binaria si el $p\text{-valor} \leq 0,05$, sino que se implementa el criterio del puntaje de relevancia de la ecuación 3.3 para obtener un puntaje continuo y clasificable en orden descendiente. Se compara contra el conjunto de validación construido desde las interacciones BioLiP de la sección 3.1.5, como “ground truth” del fenómeno físico que se busca observar/predecir.

conjunto completo de datos. Este mínimo es la precisión de una muestra azarosa, o sea, la proporción de muestras positivas en el conjunto de validación. Esta cifra representa la línea de un clasificador azaroso, y también es el AUC debajo de esta curva base a superar. Podemos ver en la figura 3.14 que las predicciones estadísticas de interacción droga-PFAM superan este filtro azaroso de forma consistente, por lo que también en materia de precisión-recuperación nuestro filtro tiene buen rendimiento.

En este caso, como en los que veremos en el capítulo 4, es importante saber en qué circunstancias cada curva aporta la mayor información. La ROC es más útil cuando el conjunto de validación presenta un balance entre positivos y negativos, mientras que la precision-recall es más confiable en el caso inverso. Como podemos ver en la figura 3.14, la precisión de una curva base sería 0,003, equivalente a la proporción de positivos: por ende estamos en el segundo caso, un desbalance, y fue necesaria la curva PR para afirmar que nuestro filtro tuviese buen rendimiento. La conclusión es que el puntaje de relevancia de interacciones PFAM-droga nos otorga una medida de asociación consistente con las validaciones construidas a partir de BioLiP, o sea, reflejan las posiciones significativas que se sabe, con la mayor confiabilidad posible, que ocupan las moléculas al ligarse con las proteínas.

En esta sección exploramos relaciones entre las proteínas asociadas a una anotación PFAM y las drogas con las que interactúan, resumidas en un cluster identitario de Tanimoto. De esta información obtuvimos conclusiones significativas: por un lado, las asociaciones significativas con un dominio PFAM tienden a asociarse a una mayor

similitud química entre los clusters identitarios en cuestión. Por el otro, estas asociaciones significativas, obtenidas puramente desde consideraciones estadísticas, coinciden con lo observado con cristalografía de alta precisión, respecto a qué secciones de una proteína son ocupadas por un ligando en interacción. Esto será especialmente útil en el capítulo 4, y en general lleva a preguntarnos cómo las interacciones de una dada proteína pueden informar nuevas asociaciones con otras proteínas en su mismo grupo PFAM. Esto cobra mayor importancia aún cuando consideramos que muchas especies poseen poca información de interacciones en sus targets, y se verían beneficiadas de poder comunicar hacia ellas la información de otras especies. Evaluaremos esta idea en la sección siguiente, con una cuantificación de la diversidad en especies que contiene un dado grupo PFAM.

3.4. Entropía inter-especie y flujo de información

Como vimos en la sección anterior, las anotaciones de proteínas sirven para agruparlas no solo según los criterios de la anotación en particular (similitud en secuencias/dominios, linaje evolutivo) sino que también poseen drogas características, con las cuales las proteínas agrupadas interactúan a un nivel estadísticamente significativo. Esto coloca un peso importante en el rol de las anotaciones para agrupar información de interacciones droga-target, y realza la importancia de los grupos PFAM y las relaciones de ortología como posibles puentes mediante los cuales propagar información entre especies.

Uno puede encontrarse con un problema de tratamiento farmacológico donde el organismo de interés ha sido muy poco estudiado y por ende no dispone de muchos datos experimentales. Se podría contar con otra especie, un organismo modelo (por ejemplo *M. musculus*) para el cual muchas proteínas cuentan con un amplio repertorio de drogabilidad. A través de una asociación por anotaciones, la integración de datos planteada en esta tesis en forma de red multicapa hace posible considerar las interacciones presentes en el segundo organismo y utilizarlas sobre el primero, justificado por las similitudes que propone el criterio de anotación. No solo eso, sino que también puede aplicarse con confianza el criterio de culpa por asociación para priorizar en red a través de la capa química, ya que las observaciones de la sección 3.3 establecieron una relación entre la similitud química y la co-pertenencia a un grupo PFAM como drogas significativas.

Si bien un mismo dominio PFAM puede aparecer asociado a proteínas de distintas especies, no todos los grupos PFAM involucran a la misma diversidad de especies. Para cuantificar esta propiedad utilizamos una entropía de Shannon, que cuantifica la heterogeneidad media esperada en un observable:

$$S_A = \sum_{i \in \text{species}} f_i^{(A)} \log(f_i^{(A)})$$

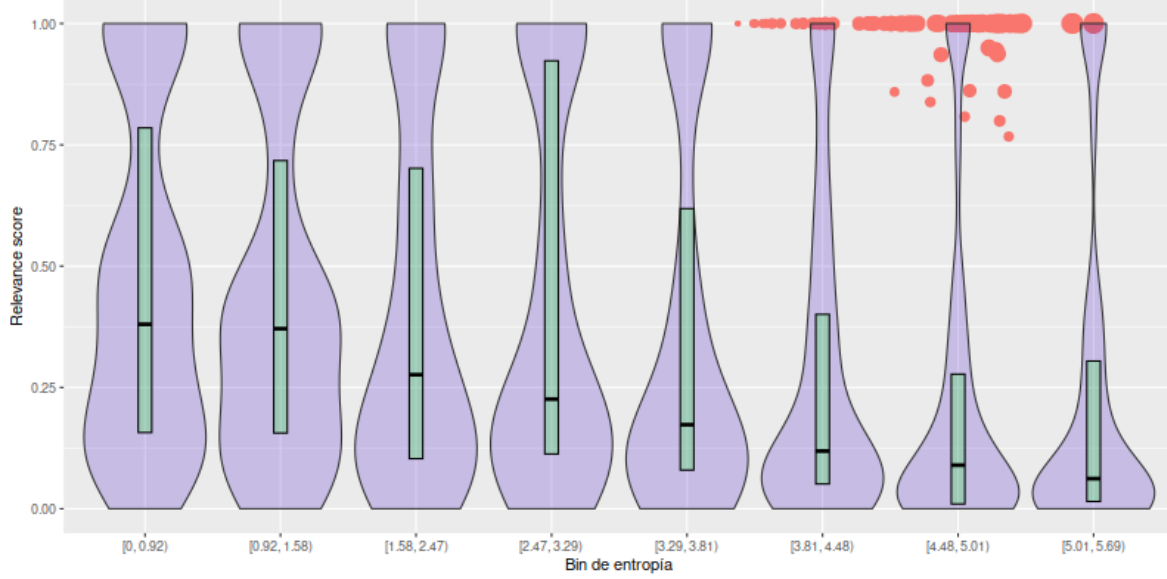


Figura 3.15: Diagrama violin (con cuartiles boxplot) para comparar entre bins de entropía inter-especie de un dado grupo PFAM (eje x) y el puntaje de relevancia calculado a partir de la drogabilidad del grupo PFAM (eje y). A medida que aumenta la entropía inter-especie (indicando un mayor beneficio para comunicar la información de los targets de una especie hacia otra) decae la drogabilidad del grupo PFAM (indicando una menor cantidad de información para compartir). Los grupos PFAM excepcionalmente relevantes y diversos en especies se encuentran en el extremo de arriba a la derecha.

donde $f_i^{(A)}$ es la frecuencia relativa de proteínas presentes de la especie i dentro del grupo A evaluado. Se puede entonces calcular la *entropía inter-especie* para cada grupo PFAM. En conjunto al puntaje de relevancia discutido en la sección 3.3, ambos números nos entregan un panorama que permite evaluar al mismo tiempo cuánta información relevante se ve contenida por la anotación y la capacidad para transmitir dicha información entre especies.

Podemos visualizar este panorama en la figura 3.15. Organizadas por columnas según la entropía inter-especie que demostraron distintos grupos PFAM, se muestra para cada una su distribución de puntaje de relevancia, en formato barplot. Observamos que de manera casi monótona (salvo por el primer intervalo) la mediana de las distribuciones de puntaje de relevancia disminuyen a medida que aumenta la entropía inter-especie. Esto se ve acompañado de cierta dispersión, pero la misma disminuye a medida que se avanza en el gráfico. Lo que se puede observar a partir de esta figura es que existe por lo general un balance entre el puntaje de relevancia y la entropía inter-especie, de forma que a medida que un grupo PFAM obtiene más presencia entre especies diversas, al mismo tiempo cae la drogabilidad de las proteínas contenidas en dicho grupo. Existen pocas excepciones que tienen ambos puntajes altos, donde más se puede aprovechar la transferencia de información. En el cuadro 3.3 resumimos algunas de estas

PFAM	Nombre	Entropía/Máxima	RScore	Targets totales
PF00186	Dihydrofolate reductase	5.69/5.98	1	63
PF00753	Metallo-beta-lactamase superfamily	5.58/7.99	1	259
PF00204	DNA gyrase B	5.57/6.64	1	100
PF00432	Prenyltransferase and squalene oxidase repeat	5.32/7.10	1	137
PF01048	Phosphorylase superfamily	5.31/7.28	1	157
PF00326	Prolyl oligopeptidase family	5.28/7.64	1	201
PF00348	Polyprenyl synthetase	5.25/7.32	1	160
PF01233	Myristoyl-CoA:protein N-myristoyltransferase, [...]	5.21/5.58	1	48
PF02852	Pyridine nucleotide-disulphide oxidoreductase, [...]	5.21/5.21	1	172
PF00850	Histone deacetylase domain	5.19/5.19	1	228

Cuadro 3.3: Diez dominios PFAM con el mayor producto del RScore con la entropía inter-especie, dando la mejor posibilidad de transferir información de targets drogables entre especies distintas. Se abrevia el nombre de algunos dominios. Se comparan las entropías con las máximas posibles, si los targets estuviesen repartidos de forma equitativa por las especies presentes en el PFAM en cuestión.

excepciones, los 10 PFAM que más logran no solo una alta drogabilidad, sino también verse representados en una gran diversidad de especies.

Vista tanto la factibilidad como las dificultades en la transferencia de información de interacciones entre especies, podemos concluir, junto a los demás capítulos, que existe una base concreta de datos desde la cual encarar el problema de la priorización que veremos en el capítulo 4. En el proceso de caracterizar las bases de datos pertinentes hemos descubierto características llamativas, las cuales discutiremos a continuación.

3.5. Discusión

En este capítulo no solo sembramos las bases para los análisis del capítulo 4, sino que también pudimos extraer conocimientos interesantes acerca de la interacción compleja entre la información de similaridad química entre drogas, las clasificaciones de proteínas según criterios bioquímicos, y las interacciones conocidas entre ellas y las moléculas de droga.

Establecimos las similitudes y diferencias entre la similaridad por criterio de Tanimoto, por proporción de pesos moleculares en subestructura, y por compartir esqueletos moleculares (scaffolds) en común. Notamos una compatibilidad entre las tres, y en especial observamos una estructura jerárquica en la topología de red compleja que se construye a partir de las interacciones a pares de Tanimoto. Esta estructura jerárquica es evidenciada con un análisis de comunidades, donde notamos que cada comunidad tiende a asociarse con un scaffold en particular, y que las comunidades cercanas tienden a identificarse con scaffolds similares.

Estas observaciones resultarán útiles durante el proceso de priorización y recomendación de nuevos fármacos para interactuar con proteínas específicas. Tener en cuenta a los esqueletos moleculares es una clave del diseño racional de fármacos, tanto en en-

foques de variación de scaffold como el “scaffold hopping” [69][70][71] como cuando se buscan compuestos de esqueleto similar, variando estructuras periféricas o combinando scaffolds [72]. Una gran parte de estos enfoques utiliza aprendizaje automatizado sobre estos conjuntos de datos para hacer el máximo provecho posible de la información de esqueletos moleculares. Habiendo establecido su consistencia con otras medidas de similaridad química, esto nos prepara para, en el capítulo 4, combinar distintas fuentes heterogéneas de información para el propósito de la priorización y recomendación.

De vital importancia en nuestros análisis, y los próximos, fue la concordancia no solo entre las similaridades químicas, sino también entre ellas y las interacciones conocidas droga-target. La “culpa por asociación” establecida entre drogas por similaridad química es una base fundamental para la priorización, sea directamente relacionando fármacos con enfermedades [73] o con targets[74]. Estas no son las únicas formas de relacionar drogas: por ejemplo, Yang et al usaron perfiles de efectos secundarios para establecer una similaridad entre ellas, con el propósito de examinar indicaciones de fármacos para enfermedades [75]. Fuera de las medidas químicas, veremos en el capítulo 4 que podemos asociarlas según interacciones con targets similares, aplicando el índice de Jaccard a las interacciones en común como visto en la sección 3.2.

Muchos autores no se limitan a las asociaciones directas entre drogas y targets, sino que propagan a distintas capas de información. Rutherford et al, por ejemplo, utilizan las interacciones droga-target conocidas en conjunto con asociaciones gen-enfermedad para proponer dianas terapéuticas específicas a enfermedades [76]. En una veta similar, nosotros elegimos proponer asociaciones entre drogas y grupos PFAM, que asocian proteínas por similaridades en secuencia y en dominio, regiones significativas dentro de su estructura. Optamos por un procedimiento estadístico utilizando un test de Fisher para evaluar la cantidad de proteínas que forman un “puente” entre las capas de red de drogas y de anotaciones PFAM, ajustando por comparaciones múltiples según metodo de Holm.

Este análisis nos reveló para muchos grupos PFAM un conjunto de drogas que se pueden considerar significativamente asociadas con ellos. Esta observación no es meramente estadística: al comparar dichas moléculas entre ellas se nota que a menudo comparten scaffolds en común, y las similaridades químicas dentro de uno de estos conjuntos es significativamente más alta que en el caso de asociar drogas que no comparten un mismo grupo. Notamos una relación inversa coherente, aunque más débil, cuando separamos a las asociaciones droga-droga según sus pesos de similaridad Tanimoto, y evaluamos cuánto comparten grupos PFAM a las que están significativamente asociadas (3.13). Esta serie de resultados nos indica que con un procedimiento estadístico logramos concentrar información de interacciones droga-target por grupos PFAM, de una manera que hace eco también de las asociaciones químicas de similaridad, las cuales no tuvieron ingerencia al calcular dicha significancia estadística.

Además comparamos la información de asociación PFAM-droga con un conjunto de validación “ground truth” obtenido en base a BioLiP, una base de datos que coloca

las interacciones de drogas con proteínas en los sitios específicos que ocupaban sobre la secuencia, de acuerdo a la estructura cristalina asociada. En combinación con la información de qué sitios engloban un dado grupo PFAM, pudimos establecer una base de datos de gran confianza acerca de la interacción física entre un fármaco y las regiones asociadas a un grupo PFAM. Al comparar, vimos que nuestro análisis estadístico, combinado con el puntaje de relevancia expresado en la ecuación 3.3, forma un filtro de buen poder predictivo sobre las interacciones físicas de máxima confiabilidad. Llevaremos esto un paso más adelante en la sección 4.3, aprovechando la difusión en red que veremos en el capítulo 4 para mejorar el poder predictivo logrado. Esto servirá como un ejemplo claro de la mejora que se puede lograr si se cuenta con una base de datos confiable y una aplicación meditada de la difusión en red.

Por último, este capítulo estableció la facilidad y las dificultades en propagar información desde organismos modelo hacia especies con menos datos de interacciones droga-target. Para distintos grupos PFAM se calculó una medida de entropía para cuantificar la diversidad de especies representadas dentro del mismo. Combinado con el puntaje de relevancia de drogabilidad de la sección 3.3 observamos un intercambio compensatorio entre ambos: los grupos PFAM con mayor drogabilidad, es decir, mayor cantidad de información de interacciones, también fueron los que menor diversidad presentaban en sus especies. El inverso era cierto, donde los mayores puentes entre especies a menudo tenían poca información para compartir. Esto no impide la migración de los datos: existen grupos intermedios en ambos campos que pueden ser aprovechados, y algunos excepcionales que rinden bien bajo ambos criterios.

Más que revelar una barrera a la priorización, esto nos informa que tendremos que ponderar bien cómo tomar el máximo provecho de la información disponible y de los criterios para difundirla. Lo mismo se puede decir de todos los resultados de este capítulo: nos han revelado las ventajas y la naturaleza del juego complejo entre las distintas fuentes de datos con las cuales contamos.

3.6. Conclusión

Este capítulo ha sido en parte una preparación para el próximo, dejando una base sólida de datos y una comprensión global del juego complejo entre sus distintas fuentes. Los criterios que esto ha aportado son extremadamente valiosos: el aprendizaje automatizado depende no solo de los algoritmos a utilizar, sino de la calidad de los datos y de su incorporación correcta.

Sin embargo, esto no fue solo un fundamento para lo que sigue. Las exploraciones de este capítulo han revelado características interesantes y relevantes de los datos de interacción droga-target, y han dejado en evidencia que la detección de una interacción no es un acto de azar, sino que se ve íntimamente relacionada con las similitudes químicas entre las drogas involucradas y las anotaciones pertinentes a las proteínas que

se considera. Estos son conocimientos no triviales en el campo del diseño farmacológico, y tienen provecho mucho más allá de lo que veremos en el capítulo 4.

Capítulo 4

Priorización de fármacos

Lejos, en el dominio celestial [...] existe una red maravillosa que ha sido colgada por algún artesano hábil, de tal manera que se extiende infinitamente en toda dirección. [...] el artesano ha colgado una joya brillante de cada nudo en la red [...]. Si elegimos arbitrariamente una de estas joyas y la observamos con cuidado, descubriremos en su superficie pulida el reflejo de todas las otras joyas en la red, de número infinito. No solo eso, sino que cada joya reflejada en ella también refleja todas las demás, por lo que sucede un proceso infinito de reflexión.

—F. H. Cook, *“El budismo Hua-Yen: La red enjoyada de Indra”*

La difusión en red, también llamada “label propagation”, es un método frecuente para la recomendación personalizada en redes, y un caso particular, muy explorado, del aprendizaje semi-supervisado. Se justifica mediante la premisa de la “culpabilidad por asociación”: la idea de que la similaridad entre entidades (sean usuarios de una plataforma, entes biológicos complejos o ubicaciones geográficas) en un espacio de conocimiento puede extenderse a otro dominio. Puntualmente, en este capítulo se trabajará bajo la suposición (corroborada en el capítulo 3) de que dos drogas con estructura química similar (entre otras similitudes) interactuarán con mayor probabilidad con el mismo conjunto de proteínas.

Desde el enfoque de redes, cuando buscamos proponer un nuevo fármaco en relación a una dada proteína, nos referimos a poder establecer esa nueva relación a partir de una transferencia de información sobre la actividad ya reportada de drogas conocidas. Dada, por ejemplo, una proteína de interés, esto implica inferir un campo de valores numéricos para cada nodo droga, que refleje el grado de asociación inferido a partir de la información ya conocida. En la sección 4.1 trabajaremos con las características que debe satisfacer este puntaje y su distribución, tratándolo como un campo escalar en el espacio discreto de la red compleja. Utilizando la culpabilidad por asociación como

base explicaremos la idea detrás del aprendizaje semi-supervisado en red, y desde las nociones análogas en la mecánica del continuo, formalizaremos la idea de la difusión en red para garantizar que se cumplan ciertas intuiciones físicas.

4.1. Metodología

En una red, los enlaces son la evidencia directa de una relación de similitud o asociación conocida. Estas nociones se pueden extender a pares de nodos no relacionados como primeros vecinos, basándose en la transitividad que ofrecen las relaciones a pares concatenadas. La difusión en red toma esta idea para plantear un flujo de información que puede comenzar en un nodo y llegar a cualquier parte de la red. Este procedimiento se puede obtener desde una analogía directa con la física de la difusión continua, como veremos en subsecciones adicionales. Sin embargo, al ser las redes espacios discretos no-Euclidianos, esto nos motiva a generalizar el operador Laplaciano a un espacio discreto e irregular.

Los puntajes propuestos también se suelen considerar tales que se respete la llamada hipótesis de suavidad. Esto implica que el resultado de la difusión sea consistente, respetando al mismo tiempo la información contenida en los datos preexistentes. Un vector $\hat{y}$ (donde cada dimensión representa el campo escalar localizado en un dado nodo) cuantifica esta información previa: por ejemplo, se puede exigir que lleven un valor de 1 los nodos correspondientes a drogas que interactúan con una proteína de interés, y los demás 0, denotando una falta de conocimiento. Este vector influirá sobre la parte *supervisada* del aprendizaje, mientras que la topología de la red, y la suavidad del campo en ella, forma una componente *insupervisada*. El balance entre ambas contribuciones le da a esta modalidad de aprendizaje el nombre *semi-supervisado*[77]. Si el vector $\vec{y} \in \mathbb{R}^n$ denota la información aprendida a través del proceso, buscamos optimizar un criterio de costo cuadrático

$$E[\vec{y}] = \alpha |\vec{y} - \hat{\vec{y}}|^2 + (1 - \alpha) \vec{y}^T \bar{\bar{L}} \vec{y} \quad (4.1)$$

donde el primer término le da prioridad a la cercanía de lo aprendido a lo ya conocido (componente supervisada) y el segundo término a la suavidad según la topología de la red (componente insupervisada), que se cuantifica a través del operador Laplaciano en red $\bar{\bar{L}} : \mathbb{R}^n \times \mathbb{R}^n$, también llamado matriz Laplaciana, semidefinida positiva, que veremos más en detalle en la sección 4.1.4. El balance entre supervisión e insupervisión se maneja a través de la constante escalar α , que se puede tratar como un hiperparámetro del algoritmo cuando se realiza una validación cruzada como mostraremos en el capítulo 4.

Cuando se optimiza la ecuación 4.1 respecto al campo escalar $\vec{y}$, vemos que

$$0 = \alpha(\vec{y}^* - \hat{\vec{y}}) + (1 - \alpha) \frac{\bar{\bar{L}} + \bar{\bar{L}}^T}{2} \vec{y}^*$$

donde $\vec{y}^*$ representa el estado estacionario de un sistema dinámico

$$\frac{d\vec{y}}{dt} = \alpha(\hat{\vec{y}} - \vec{y}) - (1 - \alpha) \frac{\overline{\overline{L}} + \overline{\overline{L}}^T}{2} \vec{y} \quad (4.2)$$

Para asegurar un aprendizaje correcto es crucial elegir de forma correcta la matriz Laplaciana. En el capítulo 3 nos dedicamos a la curación e interpretación de los datos que conforman nuestras redes de interés, asegurando su fiabilidad y relevancia al problema de priorización de fármacos. Esto nos permite una base firme para el trabajo de priorización de este capítulo, donde más pesarán nuestras decisiones a la hora de construir la matriz Laplaciana. En esta sección mostraremos un enfoque motivado por la física para realizar esta tarea. Comenzaremos por la conocida condición de continuidad, y seguiremos para construir la familia completa de Laplacianos que son capaces de respetar los conceptos presentes en la física continua, trasladada a un espacio discreto no-Euclideo como las redes complejas.

4.1.1. Continuidad y circulación en un espacio continuo

En un espacio continuo n -dimensional real, la relación de continuidad para una dada cantidad escalar de densidad $\rho(\vec{x}, t) : (\mathbb{R}^n, \mathbb{R}) \rightarrow \mathbb{R}$ se puede expresar como

$$\int_V \left(\frac{\partial \rho(\vec{x}, t)}{\partial t} - \sigma(\vec{x}, t) \right) dV - \oint_S \vec{j}(\vec{x}, t) \cdot d\vec{S} = 0$$

donde $\vec{j} \in \mathbb{R}^n$ es una cantidad vectorial representando el flujo de ρ en el espacio, y σ representa una distribución de fuentes/sumideros en el espacio (donde un caso particular serían las fuentes/sumideros puntuales). Esta relación expresa que una cantidad solo puede aumentar/disminuir dentro de un dado volumen V si dicha cantidad fluye hacia adentro/afuera por su superficie S . El teorema de la divergencia nos dice que

$$\oint_S \vec{j}(\vec{x}, t) \cdot d\vec{S} = \int_V \nabla \cdot \vec{j}(\vec{x}, t) dV \quad (4.3)$$

entonces podemos reemplazar y obtener

$$\int_V \left[\left(\frac{\partial \rho(\vec{x}, t)}{\partial t} - \sigma(\vec{x}, t) \right) - \nabla \cdot \vec{j}(\vec{x}, t) \right] dV = 0 \quad (4.4)$$

Al mismo tiempo, se espera que $\vec{j}$ sea una función de ρ dentro del sistema en un dado instante t , entonces solo depende del tiempo a través de ρ . Por esta razón es un campo gradiente, y satisface

$$\oint_C \nabla \rho = 0 \quad (4.5)$$

que significa que la circulación neta del campo escalar en un circuito cerrado debería ser nula. Tomamos que será proporcional a ρ (en dirección opuesta para promover la distribución suave del campo), y notando que el integrando de la ecuación 4.4 debe ser nulo combinamos con la ecuación 4.5 para obtener

$$\frac{\partial \rho(\vec{x})}{\partial t} + D \nabla \cdot \nabla \rho(\vec{x}, t) = \sigma(\vec{x}, t) \quad (4.6)$$

que es la forma general de una ecuación de calor. Se puede definir entonces un operador Laplaciano como

$$\nabla^2 f(\vec{x}, t) = -\nabla \cdot \nabla f(\vec{x}, t) \quad (4.7)$$

y obtenemos

$$\frac{\partial \rho(\vec{x})}{\partial t} = -D \nabla^2 \rho(\vec{x}, t) + \sigma(\vec{x}, t) \quad (4.8)$$

El operador Laplaciano depende entonces de dos cosas: el teorema de la divergencia, y el teorema del gradiente. En la sección 4.1.2, obtendremos una analogía discreta para el operador gradiente que satisface esta condición, y en la sección 4.1.3 hacemos lo mismo con la divergencia. A través de este análisis podremos construir un operador Laplaciano discreto, aplicable a grafos.

4.1.2. Operador gradiente en grafos

Sea una red compleja con N vértices (nodos) v_i con $i \in [1, N]$. Expresaremos un campo escalar $f_i(t)$ correspondiente a cada nodo v_i . Como vimos en la sección 4.1.1, lo suponemos un campo gradiente. Una forma general para gradiente del campo f entre los vértices v_i y v_j se puede plantear como

$$\text{grad}[f](i, j) = \frac{1}{d(v_i, v_j)} g(f_i, f_j; v_i, v_j) \quad (4.9)$$

donde $d(v_i, v_j) : (\cdot, \cdot) \rightarrow \mathbb{R}^+$ representa una distancia entre los nodos v_i y v_j , tomando en cuenta la topología de la red y devolviendo un número real no negativo, y $g(f_i, f_j; v_i, v_j) : (\mathbb{R}, \mathbb{R}, \cdot, \cdot) \rightarrow \mathbb{R}$ es cualquier operación que compare f_i y f_j de acuerdo a la topología demostrada por v_i y v_j .

Esta distancia en principio no debe satisfacer todos los requerimientos de una métrica. Sin embargo el gradiente sí debe cumplir la circulación nula expresada en la ecuación 4.5. Convirtiendo la integral al discreto, tratando a los nodos como una ubicación espacial, obtenemos el teorema de circulación para cualquier ciclo cerrado de nodos:

$$\sum_{i,j}^{\text{ciclo}} \text{grad}[f](i, j) \cdot d(v_i, v_j) = 0 \quad (4.10)$$

donde la distancia d_{ij} representa la magnitud del intervalo de integración entre los dos nodos, un espacio donde el campo no puede acumularse sino solo fluir. Si sustituimos allí dentro la ecuación 4.9 vemos que

$$\sum_{i,j}^{ciclo} g(f_i, f_j; v_i, v_j) \frac{1}{d(v_i, v_j)} d(v_i, v_j) = \sum_{i,j}^{ciclo} g(f_i, k_i, f_j, k_j) = 0 \quad (4.11)$$

Podemos notar que si el gradiente y la integración utilizaran distintas nociones de distancia, entonces sería imposible satisfacer esta condición en general. En otras palabras, la ecuación

$$\sum_{i,j}^{ciclo} g(f_i, f_j; v_i, v_j) d^{-1}(v_i, v_j) d'(v_i, v_j) = 0$$

solo tiene una solución para cualquier topología de red $g(f_i, f_j; v_i, v_j)$ si se cumple que

$$d(v_i, v_j) = \alpha d'(v_i, v_j)$$

para alguna constante α . Nos queda entonces la libertad de construir la función g tal que se cumpla la ecuación 4.11. Para un ciclo de tamaño 1, el principio nos dice que un nodo no tiene gradiente respecto a sí mismo, que tiene sentido. Para un ciclo de tamaño 2, obtenemos que

$$g(f_i, f_j; v_i, v_j) = -g(f_j, f_i; v_j, v_i) \quad (4.12)$$

que nos dice que g debe ser antisimetrica frente a un cambio en dirección. Para un ciclo de tamaño 3,

$$g(f_i, f_j; v_i, v_j) + g(f_j, f_l; v_j, v_l) = g(f_i, f_l; v_i, v_l) \quad (4.13)$$

que nos da una condición de transitividad para el operador. De cumplirse, uno puede usar la condición 4.13 para reducir cualquier ciclo de tamaño $N \geq 3$ a tamaño $N - 1$. Por lo tanto, es suficiente satisfacerla junto a la condición 4.12 para garantizar que se cumpla para cualquier ciclo.

Esto solo se cumple si g toma la forma

$$g(f_i, f_j; v_i, v_j) = \alpha (r(f_j; v_j) - r(f_i; v_i)) \quad (4.14)$$

donde cualquier α se absorbería luego en la distancia $d(v_i, v_j)$. En otras palabras, la dependencia sobre distintos nodos debería ser separable.

Además suponemos linealidad sobre el campo para el operador diferencial. Expresando g como transformación lineal del campo f en distintos nodos obtenemos

$$g(f_i, f_j; v_i, v_j) = (h(v_j)f_j - h(v_i)f_i) \quad (4.15)$$

donde $h(v) : \mathfrak{R} \rightarrow \mathfrak{R}$ ocupa el rol de un peso asignado a cada v rtice acorde a sus propiedades individuales. Sustituyendo en la ecuaci n 4.9 obtenemos

$$\text{grad}[f](i, j) = \frac{h(v_j)f_j - h(v_i)f_i}{d(v_i, v_j)} \quad (4.16)$$

para cualquier funcion cont nua real $h(\vec{v})$ y cont nua real no-negativa $d(\vec{v}, \vec{v})$. Cada par de nodos distinto le dar  un valor distinto a este operador. Es decir, cada valor posible corresponde a un arista posible distinta en el grafo. La ecuaci n 4.16 muestra la forma matem tica m s general que debe cumplir un operador gradiente sobre el espacio de una red.

Esto da a lugar al concepto de densidad de flujo, donde dicho flujo se hace m s f cil cuanto m s peque a la distancia $d(v_i, v_j)$. Invirtiendo la direcci n (debe ir en contra del gradiente) y agregando una constante libre de proporcionalidad, obtenemos el flujo $G[f](i, j)$ desde el nodo i hacia el nodo j como

$$G[f](i, j) = D \frac{h(v_i)f_i - h(v_j)f_j}{d(v_i, v_j)} \quad (4.17)$$

donde $D \in \mathfrak{R}$. En la secci n 4.1.3, realizaremos un an lisis paralelo a esta secci n, respecto al concepto de divergencia.

4.1.3. Operador divergencia en grafos

Para transformar a la divergencia continua en una aplicable a grafos, adaptamos el teorema de la divergencia (ecuaci n 4.3) a un espacio discreto:

$$\sum_{E(i,j)}^{superficie} F_{ij} dS_{ij} = \sum_{v_i}^{volumen} \text{div}[\overline{\overline{F}}]_i dV_i \quad (4.18)$$

donde F_{ij} es un campo vectorial sobre el grafo, con direcci n y orientaci n de un nodo al otro, $E(i, j)$ es la arista que conecta a los nodos v_i y v_j , dS_{ij} es un peso del integrando superficie an logo al ‘ rea’ que puede existir entre dos nodos, y dV_i cumple una analog a similar con el ‘v lumen’ de un nodo.

Es suficiente reducir esto al caso de un nodo para obtener una expresi n para la divergencia:

$$\text{div}[\overline{\overline{F}}]_i = \sum_{j \neq i} F_{ij} \frac{dS_{ij}}{dV_i} \quad (4.19)$$

y sustituyendo en la ecuaci n 4.18 obtenemos

$$\sum_{E(i,j)}^{superficie} F_{ij} dS_{ij} = \sum_{v_i}^{volumen} \sum_{j \neq i} F_{ij} \frac{dS_{ij}}{dV_i} dV_i = \sum_{v_i}^{volumen} \sum_{j \neq i} F_{ij} dS_{ij}$$

La sumatoria interna conecta el nodo i con todos los demás nodos del grafo, salvo por sí mismo. Sin embargo ayuda al análisis separar a esta sumatoria entre los j contenidos dentro del mismo volumen que los v_i , y los exteriores al mismo:

$$\sum_i^{vol} \sum_{j \neq i}^{vol} F_{ij} dS_{ij} + \sum_j^{ext} F_{ij} dS_{ij} = \sum_{i,j \neq i}^{vol} F_{ij} dS_{ij} + \sum_i^{vol} \sum_j^{ext} F_{ij} dS_{ij} = \sum_{i,j > i}^{vol} (F_{ij} dS_{ij} + F_{ji} dS_{ji}) + \sum_{E(i,j)}^{sup} F_{ij} dS_{ij}$$

donde la primera sumatoria contiene los flujos internos al volumen considerado, y la segunda se limita al flujo a través de la superficie, de adentro hacia afuera. Si la superficie entre dos nodos es simétrica frente a un cambio de sentido, entonces es suficiente pedir que el flujo sea antisimétrico para anular la primera sumatoria, y se cumple entonces la equivalencia con el lado izquierdo de la ecuación 4.18. Esto deja libre la función de superficie entre nodos $dS(v_i, v_j)$ siempre y cuando cumpla una simetría invirtiendo i y j .

4.1.4. Matriz Laplaciana en grafo

Como el gradiente establecido en la ecuación 4.17 es antisimétrico (por la condición 4.12, cortesía de partir de un campo gradiente), tenemos las condiciones necesarias para calcular la divergencia del gradiente. Por analogía con la ecuación del calor (ecuación 4.8) sabemos que el flujo del campo escalar $\bar{f}$ debe fluir opuesto a la dirección del gradiente $grad[f](j, i)$ (dirigido *desde i hacia j*) que evalúa cuánto más alto es el campo escalar en el nodo destino i que en el nodo origen j . Procedemos a calcular la divergencia de este gradiente. Considerando F_{ij} de la ecuación 4.19 y sustituyendo la ecuación 4.9 evaluada desde j hacia i , obtenemos

$$\begin{aligned} div[grad[f](j, i)]_i &= \sum_{j \neq i} (grad[f])_{ji} \frac{dS_{ij}}{dV_i} = \sum_{j \neq i} \frac{h(v_i)f_i - h(v_j)f_j}{d(v_j, v_i)} \frac{dS_{ij}}{dV_i} \\ &= - \sum_{j \neq i} \frac{h(v_j)}{d(v_i, v_j)} \frac{dS_{ij}}{dV_i} f_j + \sum_{j \neq i} \frac{h(v_i)}{d(v_i, v_j)} \frac{dS_{ij}}{dV_i} f_i \\ &= \left(\sum_{j \neq i} \frac{\eta(v_i)}{\delta(v_i, v_j)} \right) f_i - \sum_{j \neq i} \left(\frac{\eta(v_j)}{\delta(v_i, v_j)} \frac{dV_j}{dV_i} \right) f_j \end{aligned}$$

donde primero se substituyó desde las ecuaciones mencionadas y se expandió el resultado. Lo próximo fue notar que la función expresada es lineal sobre las componentes

del vector $\bar{f}$. Por ende, pudimos separar las contribuciones diagonales que multiplican a la componente f_i , de las contribuciones no diagonales que multiplican a las componentes f_j , para tratar a estos coeficientes como una matriz. Con esa idea se reorganizó la sumatoria como se ve en la segunda línea. Luego realizamos la siguiente transformación que resultará conveniente: $\eta(v_i) = \frac{h(v_i)}{dV_i}$ es una función arbitraria (que incluye una normalización análoga al concepto espacial de “volumen” asociado a un nodo) y $\delta(v_i, v_j) = \frac{d(v_i, v_j)}{dS_{ij}}$ es una función que refleja la distancia entre pares de nodos (que incluye una normalización por “elemento de superficie” entre nodos). Aplicando esto obtuvimos la última línea de la ecuación, que separa de forma nítida los coeficientes que multiplican cada componente del vector $\bar{f}$. Separamos las contribuciones de las diagonales (primer término, proporcional a f_i) y las contribuciones del resto de los elementos de la matriz (segunda sumatoria, proporcional a distintos f_j). A partir de esto resulta más directo obtener los elementos de la matriz Laplaciana:

$$\mathcal{L}_{ij} = \begin{cases} \eta(v_i) \sum_{l \neq i} \delta^{-1}(v_i, v_l) & i = j \\ -\eta(v_j) \delta^{-1}(v_i, v_j) \frac{dV(v_j)}{dV(v_i)} & i \neq j \end{cases}$$

Para simplificar este análisis, podemos notar que la matriz de adyacencia, y su contraparte pesada, representan un concepto inverso a la distancia: una adyacencia nula implica distancia infinita, y un mayor peso sobre un vínculo implica una mayor cercanía entre nodos. Se puede llamar W_{ij} a un peso generalizado inverso a $\delta(v_i, v_j)$, y como consecuencia se expresa la matriz Laplaciana como

$$\mathcal{L}_{ij} = \begin{cases} \eta_i \sum_{l \neq i} W_{il} & i = j \\ -\eta_j W_{ij} \frac{V_j}{V_i} & i \neq j \end{cases} \quad (4.20)$$

A partir de esta matriz se puede plantear una ecuación análoga a la 4.8. Recordando que el campo escalar aumenta o disminuye en dirección opuesta al Laplaciano, resulta la ecuación análoga en redes:

$$\frac{d\bar{f}}{dt}(t) = -D\bar{\mathcal{L}}\bar{f}(t) + \bar{\sigma}(t) \quad (4.21)$$

donde con la elección correcta de fuente/sumidero $\bar{\sigma}(t) = \alpha(f(t) - \hat{f}(t))$ y $D = (1 - \alpha)$ se recupera la ecuación 4.2. Queda entonces planteada la forma general de matrices Laplacianas compatibles con la intuición física y el criterio de costo cuadrático presentado al principio de esta sección.

Una elección posible para el peso generalizado es $\tilde{W}_{ij} = A_{ij}$, la matriz de adyacencia, mientras que si elegimos que $V_i = 1$ para todo i , obtenemos

$$\mathcal{L}_{ij}^{un} = \begin{cases} \sum_{l \neq i} A_{il} = k_i & i = j \\ -A_{ij} & i \neq j \end{cases} \quad (4.22)$$

donde k_i es el verdadero grado del nodo i . En otras palabras, este es el Laplaciano en grafo más frecuente, el no normalizado $L = D - A$, con matriz diagonal D de grados. Otra elección posible es $W_{ij} = A_{ij}k_i^{-1}$ y $V_i = 1$, y obtenemos

$$\mathcal{L}_{ij}^{probs} = \begin{cases} \sum_{l \neq i} \frac{A_{il}}{k_i} = 1 & i = j \\ -A_{ij}/k_i & i \neq j \end{cases} \quad (4.23)$$

una elección también frecuente que penaliza a la difusión según la promiscuidad de la fuente del flujo (el grado del nodo i). Cambiando ligeramente a $W_{ij} = A_{ij}k_i^{-1}$ y $V_i = k_i^{-\lambda}$ se obtiene

$$\mathcal{L}_{ij}^{(\lambda)} = \begin{cases} 1 & i = j \\ -A_{ij}k_j^{-\lambda}k_i^{-(1-\lambda)} & i \neq j \end{cases} \quad (4.24)$$

que constituye una version adecuadamente normalizada del “Laplaciano híbrido” utilizado por Zhou et al[78]. Es fácil ver que en el caso de $\lambda = 0$ se recupera el Laplaciano anterior. En el caso de $\lambda = 0,5$ penaliza por igual al nodo origen y nodo destino, por lo que tanto ese caso como el de la ecuación 4.22 plantean Laplacianos simétricos que no se verían alterados en el paso que requiere el aprendizaje semi-supervisado desde su costo 4.1 a la ecuación de evolución 4.2. No así el Laplaciano de la ecuación 4.23, que se vería promediado con su transpuesto para obtener

$$\mathcal{L}_{ij}^{probsim} = \begin{cases} 1 & i = j \\ -A_{ij} \frac{k_j + k_i}{2k_i k_j} & i \neq j \end{cases} \quad (4.25)$$

Cabe preguntarse entonces por las características que cumple un Laplaciano de esta categoría. ¿Debe ser simétrico, o sirve aunque no lo sea? ¿Es condición suficiente el haber seguido nuestra construcción para garantizar las cualidades necesarias para la difusión en red? Esto examinaremos en la sección siguiente, limitándonos a lo más indispensable para el análisis futuro.

4.1.5. Propiedades del Laplaciano en red

En base a la ecuación 4.20 podemos plantear el Laplaciano en general como una multiplicación de matrices

$$\overline{\mathcal{L}} = \overline{V}^{-1} (\overline{D} - \overline{W}) \overline{\eta} \overline{V} \quad (4.26)$$

donde $\overline{V}$, $\overline{\eta}$ y $\overline{D}$ son matrices diagonales con entradas v_i , η_i y $\sum_{l \neq i} W_{il}$ respectivamente. Sabemos que $\overline{D} - \overline{W}$ es real y simétrica, y por lo tanto diagonalizable. Además se puede expresar como forma cuadrática:

$$\bar{x}^T (\bar{D} - \bar{W}) \bar{x} = \sum_i \sum_{l \neq i} W_{il} x_i^2 - \sum_{i \neq j} W_{ij} x_i x_j = \sum_{i > j} (W_{ij} x_i^2 + W_{ji} x_j^2 - W_{ij} x_i x_j - W_{ji} x_j x_i)$$

y como consecuencia de su simetría

$$\bar{x}^T (\bar{D} - \bar{W}) \bar{x} = \sum_{i > j} W_{ij} (x_i^2 - x_j x_i + x_j^2) = \sum_{i > j} W_{ij} (x_i - x_j)^2 \geq 0$$

por lo que, al ser $W_{ij} \geq 0$, la matriz $\bar{D} - \bar{W}$ es semidefinida positiva. Esto nos dice que todos sus autovalores son positivos o cero, que significa que en la ecuación 4.21, el operador de flujo tendrá autovalores negativos o nulos, que muestra una tendencia hacia un estado estacionario. En el caso homogéneo de dicha ecuación, el estacionario será el autovector correspondiente al autovalor nulo del operador Laplaciano (o uno de ellos, si su multiplicidad es mayor).

Como $\bar{D} - \bar{W}$ es semidefinida positiva, y $\bar{\eta}$ también lo es (por virtud de ser diagonal de entradas positivas), entonces su producto también es diagonalizable con autovalores positivos o nulos (sin garantía de que la base de autovectores sea ortogonal). Si diagonalizamos el producto de las dos, lo podemos expresar como $\bar{C}^{-1} \bar{K} \bar{C}$, donde $\bar{C}$ es el cambio de base y $\bar{K}$ es diagonal con entradas positivas o nulas. Agrupando $\bar{C}' = \bar{C} \bar{V}$ podemos decir que

$$\bar{V}^{-1} (\bar{D} - \bar{W}) \bar{\eta} \bar{V} = \bar{V}^{-1} \bar{C}^{-1} \bar{K} \bar{C} \bar{V} = \bar{C}'^{-1} \bar{K} \bar{C}'$$

tiene los mismos autovalores, con una base distinta (completa) de autovectores dada por $\bar{C}'$. Como $\bar{V}$ es diagonal, su efecto sobre $\bar{C}$ es multiplicar cada columna por un escalar distinto. Comprobamos entonces la propiedad más importante del Laplaciano, necesaria y suficiente para que la ecuación 4.21 converja siempre: tiene autovalores positivos o nulos.

Además se puede afirmar que, por cada componente desconexo de la red, se tendrá N autovalores donde N es el número de nodos allí contenidos; esto se observa porque cada componente separado tiene su propio bloque cuadrado de la matriz de adyacencia o conectividad, y que al multiplicar por las demás matrices de la descomposición, conservan su estructura como bloques separados.

Cada uno de estos bloques también tendrá un autovector correspondiente a un autovalor nulo. Esto se puede ver ya que $(\bar{D} - \bar{W})$ siempre tiene como autovector el vector constante dentro del componente en cuestión, y nulo por fuera. Restringiendo el cálculo a los nodos dentro del componente:

$$\left[(\bar{D} - \bar{W}) \bar{1} \right]_i = \sum_{j \in \text{comp}} \left(\sum_{l \neq i} W_{il} \delta_{lj} - W_{ij} \right) = \sum_{l \neq i} W_{il} - \sum_{j \neq i} W_{ij} = 0$$

este resultado se extiende fácilmente al dividir por las diagonales correspondientes a las matrices $\bar{\bar{V}}$ y $\bar{\bar{\eta}}$, por lo que en el estado estacionario, cada nodo tendrá un valor de campo de $(\eta_i V_i)^{-1}$. De esta forma se garantiza la convergencia de la evolución mediante el Laplaciano, que en el caso homogéneo de la ecuación 4.21 tenderá al estado que acabamos de describir, y en el caso forzado semi-supervisado 4.2 se apartará del mismo, hacia la información previa supervisada, según la intensidad del hiperparámetro α .

4.1.6. Difusión en red

A partir de la base que hemos explicado del operador Laplaciano y el aprendizaje supervisado, será breve explicar la idea de la difusión en red. Tomando el ejemplo del problema de encontrar nuevas drogas que interactúen con una dada proteína, basándonos en datos del estilo de los vistos en el capítulo 3, podemos partir encontrando en nuestra red a utilizar el nodo que corresponda a la proteína de interés. En términos de aprendizaje supervisado, el conocimiento que deseamos imponer sobre el sistema es que busquemos respuestas asociadas a dicha proteína: por lo tanto planteamos que el vector de supervisión $\vec{y}$ es nulo para todo índice salvo el que le corresponde al nodo de nuestra proteína, donde colocaremos una entrada 1. A este nodo se le llama “semilla”, al ser el origen de la difusión, y difundirá desde una capa de red de proteínas a una capa de red de drogas. No es necesario que sea único: en la mayoría de los casos que mostraremos, las semillas son un conjunto que se expresa de forma similar, en el vector de supervisión, con una normalización adecuada.

Luego de determinar dicho vector regresamos a la ecuación 4.2, obviando la simetrización del Laplaciano:

$$\frac{d\vec{y}}{dt} = \alpha(\hat{\vec{y}} - \vec{y}) + (1 - \alpha)\bar{\bar{L}}\vec{y} \quad (4.27)$$

y permitimos que el vector $\vec{y}$ evolucione desde un valor inicial $\vec{y}$ según dicha dinámica. Elegimos de forma apropiada la constante de supervisión-insupervisión α , que regula la rigidez con la cual se respeta la condición inicial en comparación a la topología de red.

Luego de suficientes iteraciones (sea por un criterio de convergencia o por un número fijo), se obtiene un vector final $\vec{y}^*$, con cada índice asociado a un nodo distinto, una droga posible para recomendar a la proteína de interés. Las mismas se colocan en orden descendiente de puntaje, y se quitan las interacciones conocidas en el conjunto de datos de entrenamiento (las presentes en la matriz de adyacencia/incidencia desde la cual se construyó el operador Laplaciano). Este ranking de puntaje representa a la recomendación que se podría proponer a un laboratorio experimental, o reportar en una base de datos, como resultado.

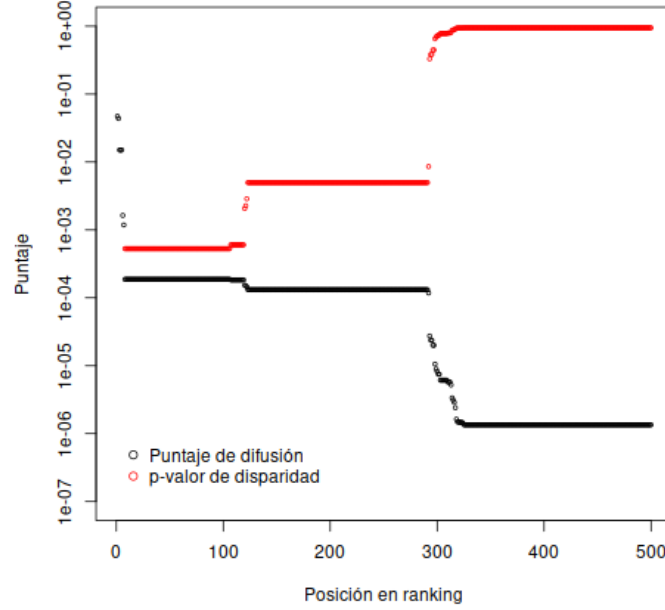


Figura 4.1: Ejemplo de ranking resultante de una difusión en red. El puntaje (negro) es el valor final del vector $\vec{y}$ luego de concluida la dinámica de la ecuación 4.27, y ordenada en orden descendiente, donde un puntaje más alto representa a una recomendación más deseable. El p-valor del filtro de disparidad (rojo) se organiza en sentido opuesto, donde uno más bajo indica una mayor confiabilidad.

Con este ranking uno puede proponer los primeros N elementos, o publicar el orden entero, sin mayor manipulación de los números: la decisión de qué recomendaciones dar dependería solamente de un umbral.

Un método útil en estos casos es el llamado filtro de disparidad (“disparity filter”)[79]. En nuestro contexto utilizamos este criterio para definir un umbral asociado a una situación de control (hipótesis nula) donde se reparte un recurso extensivo entre N elementos. La probabilidad de que un dado nodo reciba un puntaje x al azar partiendo del nodo j , la densidad probabilidad de recibirlo de parte del nodo i es

$$\rho(x_i)dx_i = (k-1)(1-x_i/s_j)^{k-2} \quad (4.28)$$

donde el “strength” s_j del nodo j es una extensión del grado al continuo, y representa la suma de los pesos que emergen de dicho nodo. La ecuación 4.28 entonces nos muestra la distribución de probabilidad que tiene el nodo i de recibir una proporción x_i/s_i del peso total que el nodo j tiene para distribuir entre k nodos.

Esta distribución se puede obtener pensando en una partición azarosa de un intervalo $[0, 1]$ con $k-1$ puntos de separadores, colocados según una distribución uniforme. Marcaccioli et al también la han conceptualizado como una aproximación a un modelo

de Pólya [80]. Integrando esta expresión podemos obtener la probabilidad de que el nodo i obtenga del j , por azar, un peso normalizado mayor o igual al recibido en la práctica $\tilde{x}_i^{(j)}$:

$$p_i^{(j)} = (1 - \tilde{x}_i^{(j)})^{k_j - 2} \quad (4.29)$$

que constituye el llamado “filtro de disparidad” (disparity filter), al asignarle un menor p-valor a un peso muy desparejo con el resto, como es fácil ver desde el modelo nulo de la ecuación 4.28.

4.2. Difusión sobre redes sintéticas

Para obtener intuiciones acerca del funcionamiento del operador Laplaciano, y para formar una idea acerca de qué podemos esperar en general, comenzamos por estudiar redes sintéticas de modularidad controlada. Estas redes se construyen a partir de dos pautas. Por un lado, se asignan clases excluyentes a los distintos nodos (por ejemplo, rojo/verde/azul). Se plantea una dada probabilidad para que los nodos de una misma clase se conecten entre sí, y otra para que se conecten con las demás, con la primera mayor a la segunda. Esto lleva a la modularidad deseada. Por otro lado, se exige que siga una distribución de grado libre de escala. Un ejemplo de estas redes se observa en la figura 4.2); para más detalle en el proceso de simulación de redes con estructura modular, ver [81].

Como vimos en la sección 4.1, el operador Laplaciano se puede definir a partir de un peso genérico W_{ij} , recíproco a cualquier distancia que se quiera utilizar. Por un lado consideraremos el caso más básico, limitado a la adyacencia, donde $W_{ij} = A_{ij}$. Analizamos además el impacto al utilizar distintas medidas de mayor complejidad, todas basadas en comparar las conexiones entre dos nodos con el número de caminos a dos pasos que los conectan, $\sum_l A_{il}A_{jl}$. La primera es la medida de similaridad por coseno:

$$W_{ij}^{cos} = \frac{\sum_l A_{il}A_{jl}}{\left(\sum_l A_{il}^2\right)^{\frac{1}{2}}\left(\sum_l A_{jl}^2\right)^{\frac{1}{2}}} = \cos(\theta_{ij})$$

donde el cálculo equivale a un coseno por analogía con un producto interno. También se considera la correlación de Pearson entre filas de la matriz de adyacencia:

$$W_{ij}^{Pea} = \frac{\sum_l A_{il}A_{jl} - \frac{k_i k_j}{n}}{n}$$

y, por último, una medida de *topological overlap* (solapamiento topológico):

$$W_{ij}^{TO} = \frac{\sum_l A_{il} A_{jl} + \Theta(A_{ij})}{\min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

donde $\Theta(x)$ es la función de Heaviside.

Como mencionamos, las tres medidas se basan en analizar los caminos que conectan dos nodos a través de un primer vecino. La similaridad por coseno los compara al número de caminos que conectan al nodo i con si mismo ($\sum_l A_{il}^2$), así como el mismo número para el nodo j . La similaridad por correlación de Pearson lo compara con un modelo nulo donde la probabilidad de conectar al nodo i y j es proporcional al grado de ambos ($\frac{k_i k_j}{n}$). Y, en el caso del topological overlap, la función de Heaviside tomará su valor más alto cuando una conexión A_{ij} coincida con un número de caminos a dos saltos $\sum_l A_{il} A_{jl}$ entre ambos nodos, proporcional a cuantas conexiones existen que podrían haberse solapado de esa manera, $\min(k_i, k_j)$.

A partir de la matriz de distancia, se construye el Laplaciano simétrico basándonos en la ecuación 4.24, extendiendo al grado k como la suma de intensidades de peso W_{ij} en un dado nodo i ($k_i = \sum_j W_{ij}$).

El enfoque que consideramos luego de esto fue un análisis espectral. El autoestado nulo del Laplaciano es una solución trivial de la difusión en red, por lo que en la figura 4.2 comenzamos a partir del primer armónico para cada uno de los distintos operadores Laplacianos considerados. La primera observación que salta a la vista es que cada autoestado responde de forma distinta a la estructura modular de la red, ya sea resaltando uno de los módulos o excluyendo a uno en particular. Lo segundo a notar es que el primer armónico señala un fenómeno parecido entre los distintos Laplacianos (el Laplaciano por adyacencia simple resalta la comunidad verde mientras que los demás la excluyen). En el segundo armónico, sin embargo, los resultados ya varían: el Laplaciano por adyacencia resalta la comunidad azul, mientras que los demás resaltan la verde, con el de Pearson llegando a difundir ligeramente hacia algunos de los nodos de las otras dos comunidades.

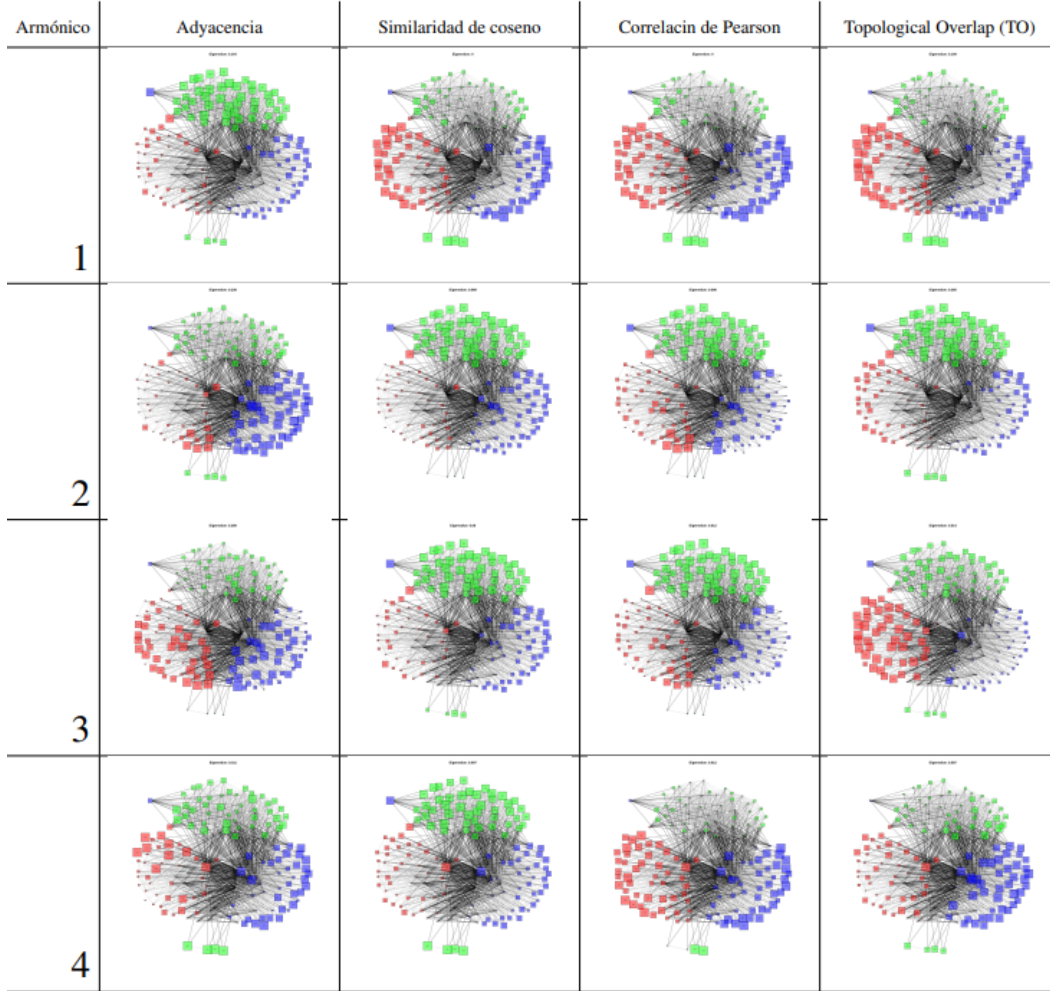


Figura 4.2: Primer a cuarto armónico (excluyendo el fundamental trivial) del operador Laplaciano construido a partir de cuatro medidas distintas de distancia (columnas). El tamaño de un nodo muestra la intensidad que lleva el campo escalar en ese sitio, para ese armónico. En una red modular, estos autoestados tienen cierto paralelo con los módulos distintos.

Los próximos dos armónicos varían aún más los resultados. Tenemos entonces que los autoestados de un operador Laplaciano resaltan información relevante a la topología de la red, en especial la estructura de comunidades. Esto coincide para distintos Laplacianos creados desde distintas medidas de similitud (pero la misma forma simétrica respecto al grado) cuando se empieza por los autovalores más bajos, pero diverge a medida que se va a los más altos. Esto indica una naturaleza consistente en su función, y complementaria en la información que se puede obtener de ellos.

Por sobre todo es importante habernos asegurado de que, al variar la medida de similitud, los operadores Laplacianos siguen cumpliendo las mismas características básicas: sus autoestados de menor autovalor son más homogéneos, y a medida que esto

aumenta se observa una tasa de cambios cada vez más abruptos en las vecindades de los nodos. Esto es fundamental para garantizar el comportamiento convergente a la hora de combinar distintas matrices de adyacencia, como haremos en la sección 4.4. Antes analizaremos un caso más simple, y en cambio exploraremos otras facetas al variar el problema de la priorización.

4.3. Priorización de asociaciones droga-PFAM

En la sección 3.3 del capítulo 3 vimos que es posible asociar con un criterio estadístico distintas drogas significativas a un grupo PFAM, según las proteínas allí contenidas que interactúan con una dada droga. No solo eso, sino que con las observaciones de la figura 3.14 determinamos que dichas asociaciones reflejan una realidad física, acerca de con qué sitios de una proteína una droga elige para interactuar.

Viendo la complementariedad entre ambos conjuntos de datos, nos proponemos utilizar a la información inferida por el método estadístico para enriquecer los datos físicos “ground truth” obtenidos con ayuda de la base de datos BioLiP[68].

La idea en general del procedimiento se puede visualizar en la figura 4.3. Se parte de una red de interacciones estadísticamente inferidas entre dos tipos de nodos: drogas (ligandos) y dominios significativos de proteínas (PFAM). Estas asociaciones estadísticas se obtienen según el procedimiento con test de Fisher y ajuste de p-valores detallado en la sección 3.3, sin considerar interacciones con evidencia física provenientes de la base de datos BioLiP. Esta segunda fuente se separa en conjuntos de difusión (semillas), de entrenamiento de hiperparámetros, y de evaluación final. Luego, para cada grupo PFAM se difunde independientemente, utilizando como semillas a los ligandos contenidos en la parte pertinente del conjunto de difusión. El puntaje obtenido luego de difundir se procesa a través de un filtro de disparidad (ver sección 4.1.6), que cumple el rol de un ajuste, sin tener que recurrir a un método de ajuste de p-valores como FDR, Holm o Bonferroni.

Esta difusión se ve afectada por una serie de hiperparámetros. El primero de los mismos deriva de la construcción del operador Laplaciano a utilizar al difundir. Con las pautas de la sección 4.1.4, se elige considerar resultados para el caso general de los Laplacianos dados por la ecuación 4.24, con valores de $\lambda \in \{0, 0,25, 0,5, 0,75, 1\}$. Esto conforma el primer hiperparámetro.

Como se mencionó red sobre la cual difundiremos será la obtenida desde nuestro método estadístico. La naturaleza de las conexiones tiene su fuerza codificada en el puntaje de relevancia. De la misma manera que la figura 3.14 se construyó en orden decreciente de puntaje de relevancia para una dada interacción droga-PFAM, esta racional se utiliza para dar mayor o menor peso a los vínculos de la red basada en inferencia. Experimentamos con poner un punto de corte sobre estos pesos, eliminando toda contribución a la red debajo de cierto umbral. EL umbral elegido es un hiperparámetro,

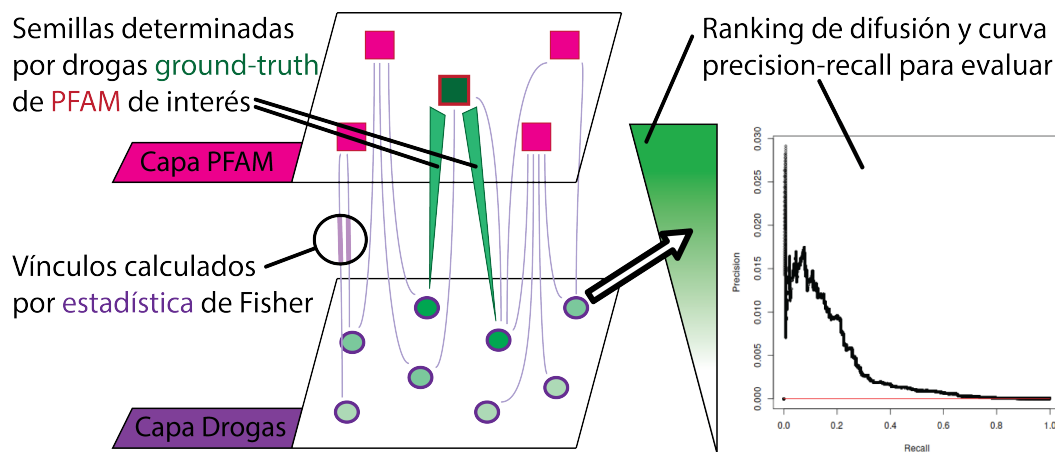


Figura 4.3: Esquema del procedimiento utilizado para priorizar nuevas interacciones físicas posibles entre un dado ligando y los sitios contenidos en un grupo PFAM presente en una proteína.

seleccionado del conjunto $\{0, 0,25, 0,5, 0,8, 0,9, 0,95\}$.

Además consideramos distintas constantes de supervisión $\alpha \in \{0,1, 0,3, 0,5, 0,7, 0,9\}$ para evaluar el balance óptimo entre la información expresada por la topología de nuestra red inferida (insupervisión) y la que obtenemos desde la información física BioLiP (supervisión). El impacto de dicha constante se puede recordar como parte de las ecuaciones 4.27 y 4.1.

Como se mencionó, separamos las interacciones físicamente motivadas BioLiP-PFAM, que conforman nuestro conjunto de datos confiables, o “ground truth”, en un conjunto de entrenamiento, validación y evaluación. Esta división tendrá proporciones 20/40/40. El conjunto de difusión servirá como semilla para establecer rankings de nodos, el de entrenamiento para ponderar el rendimiento de la difusión en red, y el de evaluación se reservará para cuando ya se hayan decidido los hiperparámetros óptimos. El procedimiento entero de difusión, entrenamiento reiterado y evaluación se realizará repitiendo esta partición de forma aleatoria diez veces, para generar una validación de mayor confianza.

Tanto en el entrenamiento como la validación, el rendimiento del algoritmo y los parámetros seleccionados se evalúa según dos características: su curva ROC y su curva Precisión-Recuperación (PR). Ambas fueron utilizadas en la sección 3.3, para caracterizar cómo los vínculos presentes en nuestra red de conexiones ligando-PFAM estadísticamente inferidas logra recuperar las interacciones entre ligandos y PFAM que, según BioLiP, poseen una fuerte validación fisicoquímica al haber sido observadas en una estructura cristalina molécula-proteína. El objetivo en esta sección es analizar si al incorporar un proceso de difusión de información sobre la red, se logra superar el rendimiento antes obtenido simplemente a partir de consideraciones sobre la estructura estática de la red. Con ese propósito, a lo largo del entrenamiento se buscará la confi-

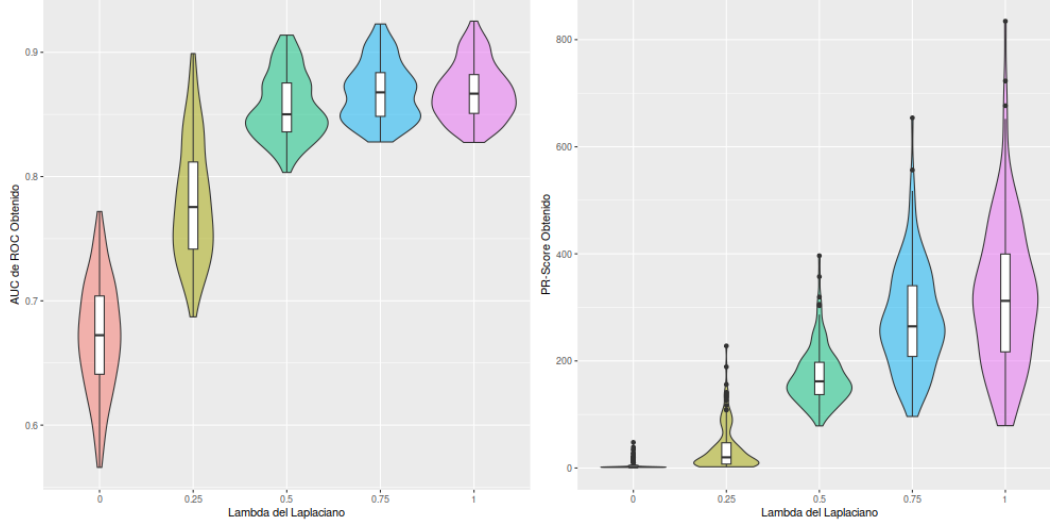


Figura 4.4: Distribución de puntajes **(izquierda)** AUC de la curva ROC y **(derecha)** AUC normalizado de la curva Precision-Recall para distintas instancias de difusión en red con distintos conjuntos de parámetros. Se visualiza una distribución diferente para aquellas difusiones realizadas con ajuste de p-valor de Holm, compensando FDR (“false discovery rate”), y sin ajustar. El violin de color da una idea de la distribución suavizada de valores, mientras que el boxplot contenido ofrece de forma concreta la mediana y cuartiles.

guración de parámetros que maximice el área debajo de la curva Precision-Recall, en comparación al área base según un filtro aleatorio. Se utiliza esta métrica y no el AUC de la ROC, porque para conjuntos de datos desbalanceados (se cuenta con muchos más negativos que positivos) el indicador más relevante es la recuperación precisa de datos conocidos. Este cociente entre el AUC de la PR y su AUC esperado en el caso nulo se denominará “PR-Score”.

Para ahondar en el proceso detrás de la etapa de entrenamiento de hiperparámetros, en las figuras 4.4, 4.5 y 4.6 se puede visualizar a gran escala el efecto de optar por distintos valores para cada parámetro. En la primera de las tres podemos observar la distribución de puntajes de entrenamiento obtenidos para cinco valores distintos del λ del Laplaciano. Este parámetro penaliza las conectividades entre nodos asimétricamente en base al grado del nodo fuente o del nodo destino: $\lambda = 0$ equivale al Laplaciano “HeatS” utilizado por Zhou et al[78], que penaliza de forma exclusiva según los nodos destino, mientras que en $\lambda = 1$ se obtiene el opuesto, denominado “ProbS”, que penaliza solo según el grado de los nodos fuente.

Una mirada a la figura 4.4 muestra una tendencia clara que favorece al parámetro $\lambda = 1$, tanto en el rendimiento de la Precision-Recall como de la ROC. Como veremos, este resultará el parámetro más fácil de decidir, que deja en evidencia lo relevante de la elección de Laplaciano, y la importancia de los desarrollos como el realizado por Zhou[78], o las consideraciones tomadas en la sección 4.1.

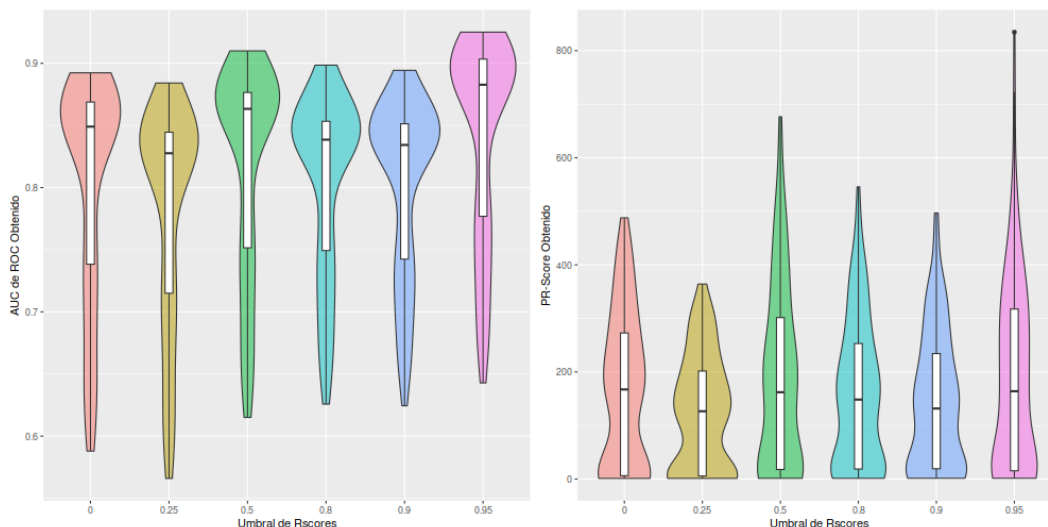


Figura 4.5: Distribución de puntajes (**izquierda**) AUC de la curva ROC y (**derecha**) AUC normalizado de la curva Precision-Recall para distintas instancias de difusión en red con distintos conjuntos de parámetros. Esta vez se compara el rendimiento cuando el Laplaciano se construye a partir de una adyacencia con distintos umbrales de puntaje de relevancia respecto a las asociaciones droga-PFAM inferidas con el test de Fisher de la sección 3.3.

La siguiente distribución de puntajes la podemos visualizar según las distintas elecciones de umbral para el puntaje de relevancia de las interacciones ligando-PFAM que conforman la red sobre la cual se difunde. Este umbral cumple el rol de “podar” la red, seleccionando la evidencia más fuerte, con el criterio de que con esas interacciones se logrará una topología de red que favorecerá el aprendizaje. En la figura 4.5 vemos una tendencia no tan clara como en el caso del anterior parámetro, y una distribución de rendimiento más alto con un umbral exigente de 0,95, o con un valor mediano de 0,5, donde se equipara la construcción de una topología de red con información diversa (bajo umbral) e información confiable (alto umbral). Estos dos valores óptimos quedarán más claros cuando se repasen los modelos finales que ajusta el algoritmo.

Finalmente, en la figura 4.6 notamos, para tomar la decisión de la constante α de supervisión, una tendencia más clara para el AUC de la curva ROC, y un poco menos clara para el puntaje que más nos interesa, el PR-Score. Notamos un ligero empate entre el rendimiento obtenido para una constante de supervisión 0,7 o 0,9, donde la primera presenta el máximo más marcado, pero la segunda una mediana más alta y una varianza más reducida. Veremos en la decisión final de parámetros que ambos pueden resultar seleccionados, dependiendo de la elección de conjunto de entrenamiento y validación, y que esto no tiene un gran efecto sobre los puntajes de validación. Lo que esto nos indica es que la preferencia, en este conjunto de datos, es hacia una difusión más cercana a los nodos semilla de donde proviene la información inicial. Esto es interesante cuando se combina con el resultado de la figura 4.4, donde se observó que para la construcción

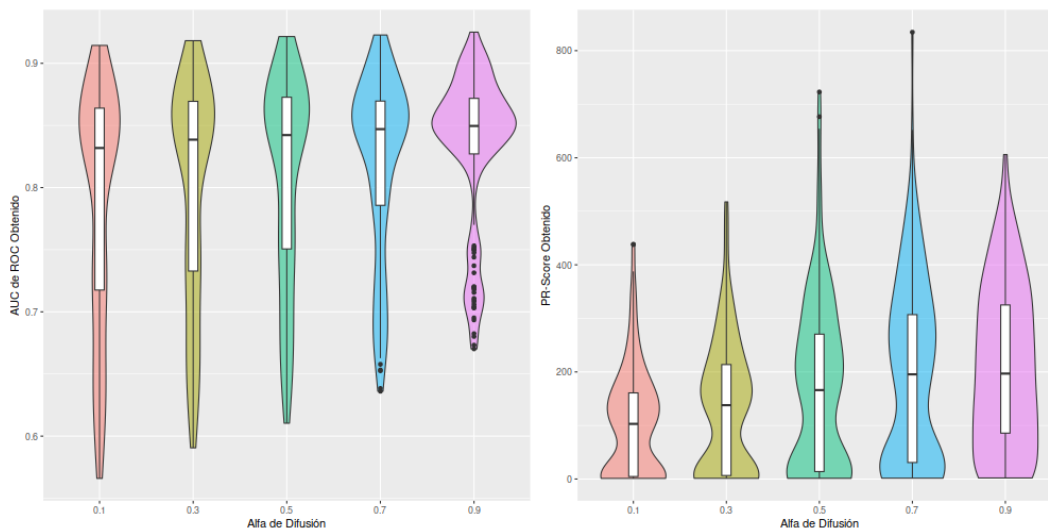


Figura 4.6: **(Izquierda)** AUC de la curva ROC y **(derecha)** AUC normalizado de la curva Precision-Recall para distintas instancias de difusión en red con distintos conjuntos de parámetros. En esta instancia se observan las distribuciones de puntajes para distintas constantes de supervisión del algoritmo semi-supervisado de difusión.

del Laplaciano se prefiere una penalización de grado respecto a los nodos fuente.

A partir del anterior análisis ya somos capaces de decidir por los parámetros óptimos. Realizamos entonces una última difusión en red, utilizando como semilla a los conjuntos de entrenamiento y validación combinados, y evaluamos la calidad del ranking obtenido utilizando el conjunto de evaluación, que hasta el momento no se había utilizado. Vemos los resultados en la figura 4.7, para las diez iteraciones distintas de validación cruzada, cada una en un color distinto. Por el lado de la curva ROC observamos una mejora muy marcada respecto a lo obtenido en la figura 3.14 del capítulo 3. Por el lado de la curva PR, a primera vista puede parecer que se logra un rendimiento similar, por las magnitudes de precisión. Sin embargo, al tomar en cuenta el caso base, vemos que la precisión bajo un filtro aleatorio sería mucho más pequeña que en la evaluación del capítulo 3, por lo que nuestro resultado obtenido es mucho más notorio.

Como resumen, se puede observar los resultados de las diez validaciones cruzadas en el cuadro 4.1. Mientras que en la capítulo 3 se obtenía un AUC siete veces mayor al esperable por azar, luego de aplicar nuestro algoritmo de difusión y ajuste de hiperparámetros obtenemos AUCs cientos de veces mayores al obtenible por un filtro aleatorio, o sea, entre uno y dos órdenes de magnitud mayores.

La reducción en el puntaje esperable por el caso base se debe a que se está priorizando sobre una cantidad mayor de datos: todas las interacciones posibles en la red. Por azar, se esperaría seleccionar una interacción “correcta” (según el conjunto de validación) con una probabilidad del orden de 10^{-5} , mientras que, al principio del ranking, se las obtiene entre un 2% y 6% de las veces. Esto marca una mejora importante, ya que,

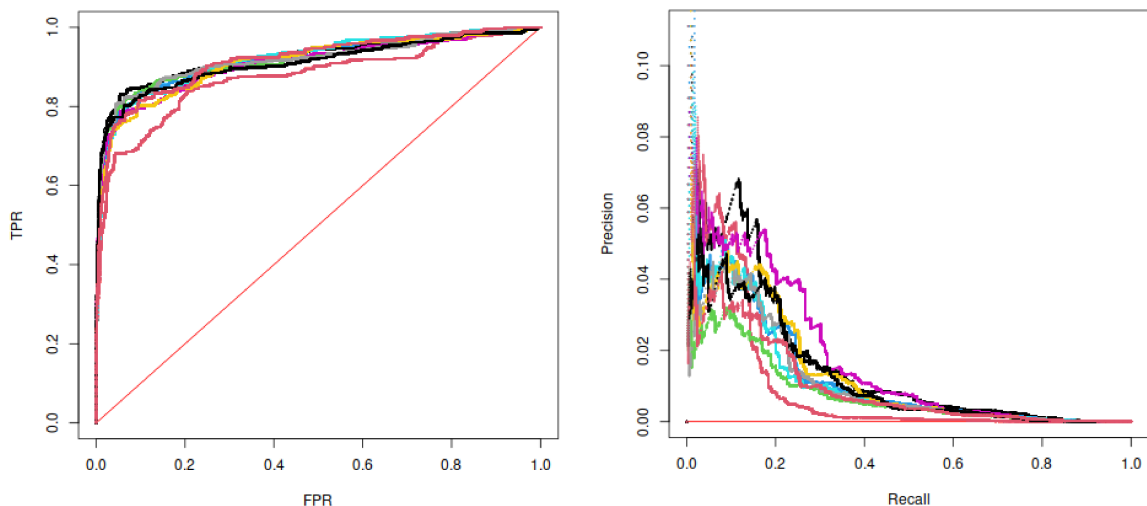


Figura 4.7: **(Izquierda)** Curva ROC y **(derecha)** curva Precision-Recall (precisión-recuperación) para diez rankings distintos de interacciones droga-PFAM predichas desde diez difusiones en red, optimizadas independientemente desde distintas particiones entrenamiento-validación. Se compara contra el conjunto de evaluación de interacciones “ground truth”, luego de utilizar las particiones de entrenamiento y validación del mismo para obtener los parámetros más convenientes.

a la hora de proponer posibles sitios donde un ligando podría interactuar, la prioridad es limitar las recomendaciones a un mínimo, con la mayor confianza posible.

Este fue un ejemplo simple pero efectivo acerca de los beneficios de una priorización utilizando difusión en red. Combinando una topología de red con datos de mayor confianza en un algoritmo semi-supervisado, fuimos capaces de enriquecer un conjunto de entrenamiento separado de los datos físicos BioLiP, y recuperar con confianza un conjunto de evaluación separado.

Esto nos muestra que, en el caso de la información con buena densidad, en redes chicas, es muy factible no solo realizar una difusión en red para obtener mejores predicciones, sino también encontrar el conjunto de parámetros óptimo para este procedimiento. En la sección siguiente mostraremos un caso opuesto: cuando la densidad de información es baja, y la red sobre la cual se prioriza es mucho más grande, debemos encarar de forma creativa tanto al desafío del mayor aprovechamiento de la información, como las limitaciones de la complejidad computacional.

4.4. Red híbrida para datos droga-target

En esta sección analizaremos el efecto al combinar información de fuentes diversas. La sección previa buscaba optimizar la recuperación de vínculos físicos PFAM-droga, mientras que a continuación nuestro objetivo será recuperar vínculos de nuestra base

λ	Umbral R-Score	α	ROC Entrenamiento	PR-Score Entrenamiento	ROC Validación	PR-Score Validación
1	0.95	0.7	0.91	834.6	0.92	246.1
1	0.5	0.1	0.9	644.25	0.87	383.13
1	0.5	0.7	0.91	692.89	0.92	212.36
1	0.5	0.9	0.88	478.42	0.92	205.37
1	0.5	0.9	0.89	525.93	0.92	159.82
1	0.95	0.7	0.92	877.59	0.91	271.36
1	0.5	0.9	0.89	604.3	0.91	221.2
1	0.5	0.7	0.87	467.58	0.92	281.93
0.75	0.95	0.3	0.89	551.3	0.91	328.66
1	0.5	0.9	0.9	501.92	0.92	142.41

Cuadro 4.1: Resumen de diez iteraciones de validación cruzada para el procedimiento de ajuste de hiperparámetros sobre algoritmo de difusión en red para la búsqueda de nuevas interacciones entre ligandos y sitios correspondientes a grupos PFAM.

de datos curada droga-proteína, como vimos en la sección 3.1.3 del capítulo 3.

En este caso, recomendaremos vínculos nuevos posicionándonos en una dada proteína y buscando nuevas drogas con las cuales podría interactuar, según la topología de nuestras redes. Nuestras semillas consistirán, para cada target, en el conjunto de drogas que sabemos que interactúan con ella (excluyendo conjuntos de validación y evaluación, claro). Desde esas semillas difundiremos sobre redes de similaridad exclusivamente droga-droga.

Una de estas redes es la de similaridad química de Tanimoto, explorada en detalle en el capítulo 3. La segunda será una red de asociación entre drogas por un scaffold de primera capa compartido: es decir, ambas drogas se componen de un mismo esqueleto, con distintas estructuras periféricas asociadas. Finalmente, la última red de asociación droga-droga es la única de las tres que no se basa en un criterio estructural químico. En cambio se utiliza el índice de Jaccard visto en el capítulo 3 para establecer una adyacencia según con cuántas de las mismas proteínas un par de drogas interactúa.

Con estas tres medidas distintas armaremos tres redes con los mismos nodos, pero distintos vínculos entre ellos. Los tres combinados darán lugar a nuestro Laplaciano, a través de una adyacencia híbrida por combinación lineal (de peso normalizado):

$$\overline{\overline{A}}^{(hyb)} = a_{Tan} \overline{\overline{A}}^{(Tan)} + a_{sca} \overline{\overline{A}}^{(sca)} + a_{int} \overline{\overline{A}}^{(int)} \quad (4.30)$$

o sea, una suma pesada entre la similaridad de Tanimoto, la asociación por scaffolds, y la medida por interacciones compartidas con proteínas.

La pregunta será, entonces, qué combinación es la óptima. Pero en las redes de la capa química se encuentran alrededor de siete millones de nodos. A priori es difícil saber si la combinación óptima para una región de esta red enorme será la misma que en otra

parte.

Queriendo resolver esta duda, centramos esta sección alrededor de un método modular para la búsqueda de parámetros óptimos. Separaremos la red de Tanimoto según una partición de Louvain, que como hemos visto resulta en una estructura altamente modular con paralelos en los scaffold moleculares. Luego separaremos las dos otras redes según los mismos grupos de nodos, y para cada una de ellos realizaremos repetidas difusiones en red, variando los parámetros de balance de la red híbrida (a_{Tan} , a_{sca} , a_{int}) y el parámetro de supervisión-insupervisión α . De esta forma nos aseguraremos de extraer la información óptima de cada región de la red.

Separamos los datos de interacción en entrenamiento, validación y evaluación en una proporción de 70/20/10, muy distinta a la partición de la sección anterior por un problema muy particular: en aquel caso, los datos de interacción habían sido aglomerados según grupo PFAM, pero, en este caso, dicha información sigue esparcida entre diversas proteínas. La densidad de estos datos es mucho más baja, que lleva a una mayor necesidad de aprovechar los datos en el entrenamiento.

Como estrategia para el aglomeramiento de datos proponemos un segundo elemento para nuestro método. En la etapa de entrenamiento de una comunidad, tanto las semillas como el conjunto de validación para recuperar serán constituidas por todas las drogas activas en la comunidad de interés. Esta decisión se toma bajo la hipótesis de que la recuperación del conjunto activo de un target, o del de todos los targets que se asocian con una dada comunidad química, tendrán comportamientos similares, debido al paralelo fuerte entre la información de interacciones mutuas y las asociaciones de similaridad química, como vimos en el capítulo 3. Esto se traduciría en la organización de la red por comunidades, ofreciendo un comportamiento similar frente a un dado conjunto de parámetros óptimos encontrados.

Se seleccionan los parámetros óptimos puntuando cada ranking obtenido contra el conjunto de validación, acorde a una métrica comparativa llamada Bayesian Personalized Ranking, diseñada por Rendle et al[82]. En vez de colocar un umbral y puntuar dicho umbral, esta medida se fija la ubicación en el ranking de los positivos del conjunto de evaluación, y la compara con aquellas entradas que no representan positivos. Cuando un positivo se encuentra más alto que un negativo, suma puntaje; en el caso contrario, lo resta; y, para los equivalentes, no cambia el resultado. La diferencia en preferencia (o sea, cuánto más alto está uno que el otro) se pondera según una función sigmoide. El resultado de este método es que (a) se coloca el peso del entrenamiento sobre una medida local, comparativa en regiones del ranking, a diferencia del AUC de una ROC o PR, que evalúa la escala global de una curva (incluso en el caso de una ROC restringida a una primera parte del ranking); y (b) las entradas no positivas, que no implican una falta de interacción con el target de interés sino un desconocimiento al respecto, no son tratadas como si el ranking hubiese encontrado información incorrecta, sino simplemente información no preferida en comparación a los positivos ya conocidos.

Luego de entrenar de esta manera a cada comunidad por separado, en cada una se

difundirá con los parámetros óptimos, utilizando como semilla a los conjuntos de entrenamiento y validación, esta vez sí de forma separada por cada target. Así se propondrá un ranking separado de drogas candidato por target y por comunidad de la red. Estos rankings tendrán que ser consolidados, y la forma más eficiente encontrada fue una normalización simple, en vez del filtro de disparidad explicado en la sección 4.1.6.

Una vez consolidados los rankings por target, se procedió a puntuarlos según el conjunto de evaluación siempre excluido hasta aquel momento. Sin embargo, ya no se empleó una puntuación global sobre el ranking como en la sección 4.3, con las curvas ROC y PR: esta nueva base de datos, y sus recomendaciones, se basan en información esparsa, y su objetivo es llegar a las recomendaciones más puntuales posibles para los objetivos prácticos. Por esa razón se define una medida de puntuación según lo primero que se reportaría en el ranking. Como ilustra la figura 4.9, para el conjunto de evaluación consignaremos el número de aciertos dentro del top-5 (los primeros cinco elementos), top-10, top-20 y top-50 del ranking generado.

El segundo método contra el cual comparar fue la aplicación de una red pura de Tanimoto. Nuestro primer análisis de los resultados fue examinar cuántos targets recibieron un mejor rendimiento (según métricas top-5 a top-50) con la red híbrida en comparación a la pura Tanimoto, y viceversa. Los resultados, visibles en la figura 4.10, demuestran una tendencia particular. Nuestro método de división y entrenamiento resulta en más targets beneficiados en la primera parte del ranking, a expensa de la última parte. Este es un resultado muy favorable: en estas aplicaciones con poca densidad de información, donde los experimentos pueden resultar costosos o inconvenientes, lo que más se desea es una recomendación puntual.

Esta tendencia también se refleja en el promedio de rendimiento por target según las distintas medidas. Como vemos en la figura 4.11, el método híbrido posee una ventaja apreciable en las primeras cinco recomendaciones del ranking, mientras que el segundo mejor método, Tanimoto, acorta la distancia en el top-10 y finalmente lo supera en el top-20. La tercera mejor red es consistentemente la pura por interacciones droga-target en común, y la última es siempre la medida por esqueletos “scaffold” en común. Esto no significa que estas redes no conformen una parte del aprendizaje híbrido: simplemente no serán preferidas por su cuenta, y tendrán un protagonismo menor a la red de Tanimoto.

Este resultado indica un rendimiento inesperado de parte de nuestro método: el entrenar por comunidades, y realizar el entrenamiento con todas las semillas allí activas, resulta capaz de particularizar bien las recomendaciones cuando, en la etapa de evaluación, se difunde desde las semillas específicas a cada target por separado.

No solo eso, sino que se privilegia a la primera parte del ranking, algo que no se planteó como objetivo puntual al comenzar este trabajo, que sin embargo se considera de gran valor. Esto se logró sin entrenar buscándolo: el puntaje de entrenamiento, como mencionamos, era una medida comparativa dentro de un mismo ranking, mientras que la medida de top 5-50 se reservó exclusivamente para la evaluación. Resulta entonces que la medida BPR, que jamás considera explícitamente el concentrar los resultados

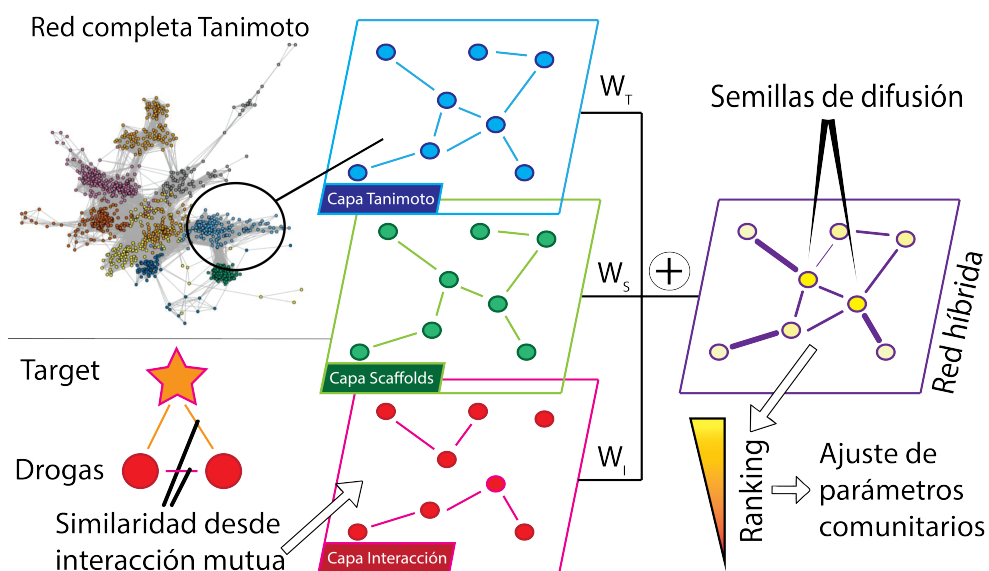


Figura 4.8: Esquema del procedimiento para la construcción de redes híbridas ajustadas por cluster. La red de Tanimoto se estructura en comunidades por un algoritmo de Louvain. Se separan las comunidades para ejecutar el proceso de aprendizaje, separando las semillas allí contenidas en entrenamiento, validación y evaluación. Se prueban distintas combinaciones de red por similaridad de Tanimoto, scaffolds y por targets en común. A partir del puntaje de validación de cada intento se establecen los hiperparámetros óptimos para la red híbrida, y el conjunto de evaluación permanece intacto para la etapa de evaluación del algoritmo.

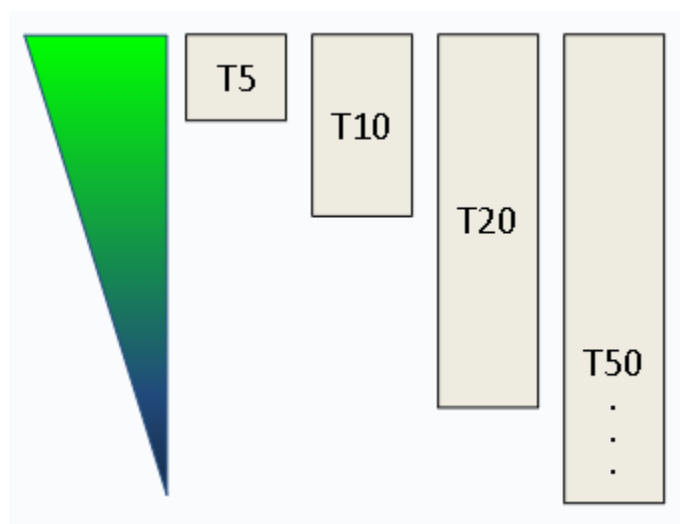


Figura 4.9: Ilustración de las métricas de evaluación del top 5-50. El triángulo verde representa el ranking en orden decreciente de puntaje (campo escalar difundido).

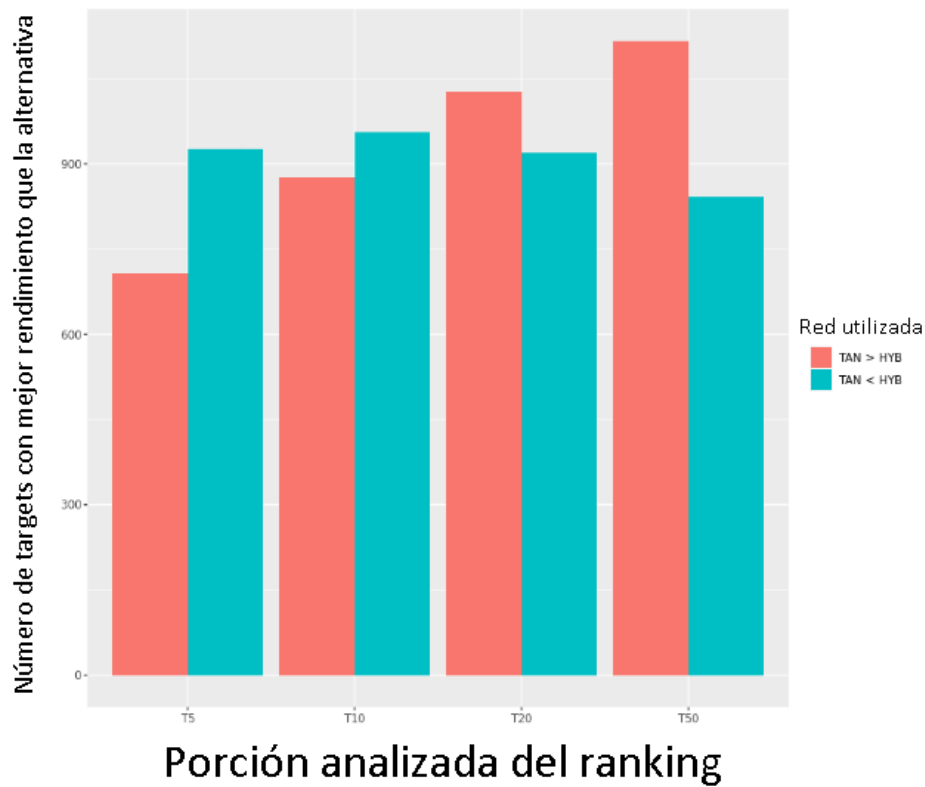


Figura 4.10: Evaluación del método de redes híbridas en comparación a la mejor alternativa (red pura de Tanimoto). Las columnas rojas muestran cuántos targets vieron un mejor rendimiento de evaluación utilizando la red de Tanimoto, mientras que en celeste se ven las que se vieron más beneficiadas por el procedimiento híbrido. Se nota un beneficio en la primera parte del ranking, y a medida que se baja en orden decreciente de puntaje, la red de Tanimoto comienza a sacar ventaja. Este resultado indica que el método de red híbrida da prioridad a un ranking más acotado, algo sumamente útil a la hora de proponer nuevos candidatos de droga para ser evaluados experimentalmente.

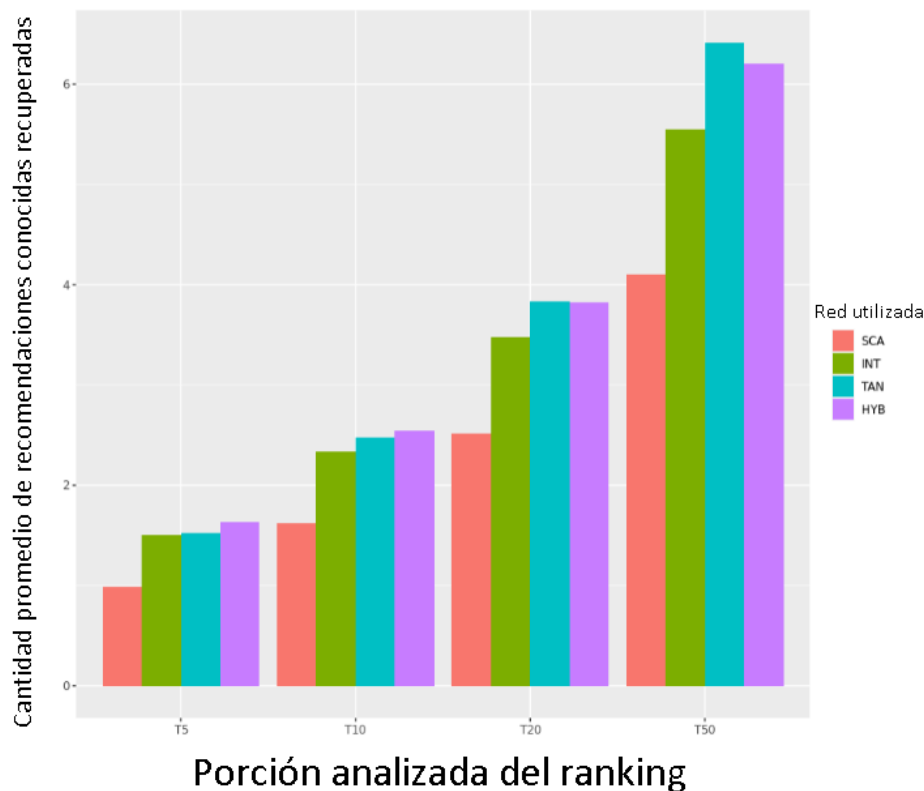


Figura 4.11: Evaluación de los tres métodos de red pura (rojo, verde, azul) comparado con el método de redes híbridas. Cada barra indica el promedio de recomendaciones conocidas que el algoritmo logra recuperar, en una dada parte del ranking, al utilizar una de las redes. Al igual que vimos en la figura 4.10, el uso de la red de Tanimoto es el segundo mejor método, y el procedimiento híbrido logra una ventaja apreciable en la primera parte del ranking.

precisos en lo más alto del ranking, logra ese efecto implícitamente al implementarlo dentro de nuestro método.

4.5. Discusión

En este capítulo cubrimos los elementos fundamentales de la difusión en red y su aplicación para el problema de la recomendación de nuevas interacciones o asociaciones en la bioinformática. No solo examinamos la mecánica interna de la priorización, sino que también examinamos los desafíos de un ajuste de hiperparámetros (sección 4.3) y cómo el mismo se complejiza al abarcar una red más grande, con menor densidad de información de supervisión (sección 4.4).

El fundamento de este capítulo fue una base sólida en términos metodológicos para el aprendizaje automatizado. El operador Laplaciano para la difusión en red tiene una

historia extensa, toma varias formas distintas y se utiliza en una enorme diversidad de aplicaciones [77][78][83][84][85][86][87]. Fue de gran interés para nosotros establecer un fundamento físico para este operador, y dar rigor a la analogía entre la difusión en grafo y la difusión física en el continuo. Aprendimos en el proceso que con exigir las relaciones de continuidad que conforman los elementos del Laplaciano físico, el gradiente y la divergencia, resulta para los grafos una familia versátil de operadores, que le aseguran a su usuario una conservación de la cantidad total de recurso de su vector de puntaje, además de un comportamiento convergente a un estado estacionario predecible.

Entre las características de este operador una de las más llamativas se refiere a su conjunto de autovalores, que le dan un carácter semidefinido. En la sección 4.2 exploramos los autoestados de un operador Laplaciano cuando se parte de una matriz de adyacencia fabricada a partir de la misma red, pero según distintos criterios de distancia. Corroboramos que entre todas se encuentra un comportamiento consistente, y que los autovectores reflejan la estructura de comunidad y conectividad presente en una red compleja. Esto es de importancia al campo amplio del análisis espectral en grafos [87][88][89], y corrobora que cualquier distancia/adyacencia que cumpla las propiedades que exigimos puede ser utilizado por tales métodos.

Nuestro enfoque principal de utilización de los Laplacianos fue la difusión en red, sin embargo. En la sección 4.3 recorrimos un ejemplo de ajuste de hiperparámetros, encontrando no solo el Laplaciano óptimo para el propósito de recomendar nuevas asociaciones físicas droga-PFAM, sino también encontrando el grado de confianza a tenerle a la red subyacente sobre la cual se realiza la difusión (cuantificado a través del umbral sobre el puntaje de relevancia y el grado de supervisión α).

En el proceso también encontramos que, al menos en aquel caso, es mejor no ajustar los p-valores de la familia de recomendaciones de un grupo PFAM antes de consolidarlo con las obtenidas para otros grupos. Esto se puede deber en gran parte a que el filtro de disparidad ya realiza un ajuste por el tamaño del conjunto de recomendaciones posibles, y toma en cuenta los saltos en sus puntajes, de forma que repetir un ajuste por método de Holm o FDR resulta excesivamente restrictivo. Esta hipótesis se ve fundamentada por el hecho de que Holm, el método más exigente, fue el peor puntuado, mientras que FDR estuvo en el punto medio entre el mismo y el no-ajuste. Esto nos indica que, a la hora de calcular p-valores para recomendar nuevas asociaciones, puede resultar mejor primero consolidar dichos p-valores y luego ajustar al conjunto entero, en vez de ajustar antes por familias separadas.

Finalmente abarcamos el caso más frecuente y exigente en términos de datos de interacción droga target. Consideramos redes enormes, conteniendo el orden de siete millones de drogas, con una densidad baja de información de interacciones. Enfrentados al desafío de esta baja densidad, sumado a la alta complejidad computacional de una red tan grande, diseñamos un enfoque de “dividir y conquistar”, donde las redes fueron partidas según comunidades, y el modelo óptimo fue entrenado para cada una individualmente, agregando el entero de la información de interacciones allí contenida.

El puntaje de entrenamiento fue una medida comparativa sin necesidad de umbrales, y sin las limitaciones de una curva ROC o PR, que para un ranking global con tan pocos datos pueden resultar inconsistentes. La medida BPR[82] resultó efectiva para el entrenamiento, ya que al realizar una difusión final y comparar con un conjunto de evaluación, se vio que los ranking resultantes privilegiaban tener información correcta al principio, a cambio de acertar menos en el rango más global.

Este fue un gran beneficio logrado no solo por el criterio de entrenamiento, sino por la posibilidad de entrenar de forma separada comunidades separadas. La posibilidad de poder enfocar en la parte alta del ranking las recomendaciones podría habilitar, en el futuro, a que se provea a los investigadores experimentales un conjunto más enfocado hacia los resultados potencialmente productivos.

4.6. Conclusión

Como parte de este último capítulo de la tesis, se buscó ilustrar algunos de los beneficios de las diversas formas de encarar el problema de la priorización en red. Este enfoque se apoyó en una fundación sólida de datos heterogéneos consolidados en el capítulo 3. Se ilustraron varias metodologías buscando aprovechar al máximo esta información diversa, con un enfoque consistente en la intuición física detrás del operador Laplaciano y las justificaciones de los criterios considerados a la hora de evaluar críticamente e integralmente los distintos métodos utilizados.

Capítulo 5

Conclusiones

A partir de las interacciones de a pares emerge un mundo de información, orden y estructura. Desde un análisis de las ocurrencias en común de nucleótidos en una cadena de ARN, hasta las redes de interacciones conocidas entre drogas y proteínas, hemos expuesto un panorama general de los resultados que pueden obtenerse cuando se considera no la identidad individual de un elemento, sino su relación con los demás miembros de un conjunto más grande.

El capítulo 2 nos enseñó un ejemplo de la concordancia que puede existir entre distintas especies del orden inferido al considerar interacciones de a pares entre sitios en una secuencia de splicing. Estas concordancias, y sus discrepancias en contraste, hicieron eco del historial evolutivo conocido de dichas especies, en especial al comparar los reinos eucariotas de plantas, hongos y animales. Además resultaron comportamientos en común a lo largo del conjunto entero de especies, revelando hechos acerca del comportamiento intrón-exón que reflejan conocimientos ya conocidos acerca de la dinámica del splicing y su historia evolutiva.

En el capítulo 3 cambiamos nuestro enfoque hacia los datos farmacológicos. Incorporamos en este nuevo análisis una gran variedad de tipos de interacción: conceptos de similaridad química entre drogas, interacciones curadas conocidas entre drogas y proteínas, asociaciones entre proteínas por composición y por linaje, y asociaciones entre drogas y sitios específicos de proteínas. Observamos una correspondencia notable entre la similaridad química y la información que se podía inferir de las interacciones droga-target. Dando un paso más, se catalogaron grupos PFAM de proteínas según la capacidad que uno puede tener para hallar drogas que interactúen sobre ellas. Este concepto de “drogabilidad” se particularizó a la preferencia que los targets de un grupo PFAM tengan por interactuar con drogas en específico, dando lugar a la “drogabilidad específica”. Dicho concepto demostró correlacionarse notablemente con la información que uno puede inferir de la medida de Tanimoto de similaridad química, y además demostró una capacidad sorprendente para predecir, en base solo a estadística, las subestructuras físicas de proteínas con las cuales una proteína interactuará.

Dicha predicción de estas estructuras fue perfeccionada en el capítulo 4. Primero se definió una familia de operadores Laplacianos que garantizaban las propiedades físicas que uno espera de la difusión, tratando a las relaciones de a pares ahora como caminos a transitar en un paisaje de red. Los estados ocupados por caminantes en este panorama se analizaron a continuación, dando una idea comprensiva del comportamiento de los autoestados del Laplaciano que indicaba una concordancia con la estructura modular de las redes. Finalmente se consideró una familia de Laplacianos y de métodos de aprendizaje para optimizar la predicción de interacciones físicas droga-target mencionada, logrando una mejora notable: esto demostró el efecto de combinar el paisaje de red con información preexistente.

Finalmente se llevó un enfoque de optimización al problema de priorizar por targets individuales. Aquí se aprovechó la modularidad del paisaje de red, razonando que en distintas regiones se obtendría un óptimo distinto para los parámetros. Este enfoque resultó exitoso, demostrando que, más allá de las interacciones de a pares aisladas, es de gran importancia el esquema de red a escala macroscópica y mesoscópica. Esto refuerza la idea de que el orden complejo construido a partir de las interacciones no solo revela nuevos conocimientos generales, sino que también se puede particularizar para problemas específicos de aprendizaje automatizado.

El panorama de redes es un campo amplio, con un número enorme de métodos como los descritos en esta tesis. Tanto en complejidad como en estructura, las interacciones construyen mundos que se pueden explorar de formas incontables. Entre nuestras conclusiones, es de suma importancia recordar lo siguiente: los sistemas contemplados no pueden ser reducidos a una suma de sus partes, no sin en el proceso perder una enorme riqueza de información. La bioquímica de sistemas, al menos en lo aquí demostrado, se presenta firmemente como una disciplina holística. A través de un trabajo interdisciplinario, y combinando información heterogénea, solo contemplando todas sus facetas disponibles se puede obtener una comprensión integral de sus problemáticas y resultados.

Bibliografía

- [1] User:Protein Chemist. File:peptide-bond-figure.png. <https://commons.wikimedia.org/wiki/File:PEPTIDE-BOND-FIGURE.png>, 2011. Accessed: 2023-11-12.
- [2] User:Jpablo cad. File:detalle de la membrana celular.svg. https://commons.wikimedia.org/wiki/File:Detalle_de_la_membrana_celular.svg, 2008. Accessed: 2023-11-12.
- [3] Mike Jones. File:cdmb.svg. <https://commons.wikimedia.org/wiki/File:Cdmb.svg>, 2006. Accessed: 2023-10-24.
- [4] User:j_ham3. File:network community structure.svg. https://commons.wikimedia.org/wiki/File:Network_Community_Structure.svg, 2011. Accessed: 2023-12-12.
- [5] Timothy W. Nilsen. The spliceosome: The most complex macromolecular machine in the cell?, dec 2003.
- [6] Gwendal Dujardin, Celina Lafaille, Manuel de la Mata, Luciano E. Marasco, Manuel J. Muñoz, Catherine Le Jossic-Corcos, Laurent Corcos, and Alberto R. Kornblihtt. How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Molecular Cell*, 54(4):683–690, may 2014.
- [7] Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang Dong Fu, and David L. Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes and Development*, 28(23):2663–2676, dec 2014.
- [8] Weijun Chen and Melissa J. Moore. Spliceosomes, mar 2015.
- [9] Anireddy S.N. Reddy, Yamile Marquez, Maria Kalyna, Andrea Barta, C W Reddy, and N. Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell*, 25:3657–3683, 2013.

- [10] Gil Ast. How did alternative splicing evolve? *Nature Reviews Genetics*, 5(10):773–782, 2004.
- [11] Igor B. Rogozin, Liran Carmel, Miklos Csuros, and Eugene V. Koonin. Origin and evolution of spliceosomal introns. *Biology Direct*, 7(1):1, 2012.
- [12] Xavier Roca, Adrian R. Krainer, and Ian C. Eperon. Pick one, but be quick: 5 splice sites and the problems of too many choices. *Genes and Development*, 27(2):129–144, 2013.
- [13] Olga V. Artemyeva-Isman and Andrew C.G. Porter. U5 snRNA Interactions With Exons Ensure Splicing Precision. *Frontiers in Genetics*, 12(July):1–33, 2021.
- [14] Mandy S. Wong, Justin B. Kinney, and Adrian R. Krainer. Quantitative Activity Profile and Context Dependence of All Human 5 Splice Sites. *Molecular Cell*, 71(6):1012–1026.e3, sep 2018.
- [15] R. Michael Stephens and Thomas D Schneider. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology*, 228(4):1124–1136, 1992.
- [16] T. A. Thanaraj and A. J. Robinson. Prediction of exact boundaries of exons. *Briefings in bioinformatics*, 1(4):343–356, 2000.
- [17] Ido Carmel, Saar Tal, Ida Vig, and Gil Ast. Comparative analysis detects dependencies among the 5 splice-site positions. *Rna*, 10(5):828–840, 2004.
- [18] Kentaro Sahashi, Akio Masuda, Tohru Matsuura, Jun Shinmi, Zhujun Zhang, Yasuhiro Takeshima, Masafumi Matsuo, Gen Sobue, and Kinji Ohno. In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites. *Nucleic Acids Research*, 35(18):5995–6003, sep 2007.
- [19] Stepan Denisov, Georgii Bazykin, Alexander Favorov, Andrey Mironov, and Mikhail Gelfand. Correlated evolution of nucleotide positions within splice sites in mammals. *PLoS ONE*, 10(12):1–24, 2015.
- [20] G. Yeo, S. Hoon, B. Venkatesh, and C. B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proceedings of the National Academy of Sciences*, 101(44):15700–15705, 2004.
- [21] William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions. *arXiv preprint*, (q-bio.QM):1–8, 2007.

- [22] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13):4786–4791, 2012.
- [23] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–12, apr 2006.
- [24] Yasser Roudi, Erik Aurell, and John a Hertz. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3(November):22, 2009.
- [25] Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9679–9684, jun 2011.
- [26] Gašper Tkačik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J. Berry, and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(3):P03011, mar 2013.
- [27] Andrea De Martino and Daniele De Martino. An introduction to the maximum entropy approach and its application to inference problems in biology, apr 2018.
- [28] Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Amos Maritan, and Nina V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 103(50):19033–19038, dec 2006.
- [29] F. Remacle, Nataly Kravchenko-Balasha, Alexander Levitzki, and R. D. Levine. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22):10324–10329, jun 2010.
- [30] Marc Santolini, Thierry Mora, and Vincent Hakim. A general pairwise interaction model provides an accurate description of in Vivo transcription factor binding sites. *PLoS ONE*, 9(6):1–14, 2014.
- [31] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold, sep 2005.

- [32] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–E1301, dec 2011.
- [33] Rocío Espada, R. Gonzalo Parra, Thierry Mora, Aleksandra M. Walczak, and Diego U. Ferreira. Inferring repeat-protein energetics from evolutionary information. *PLoS Computational Biology*, 13(6):1–16, 2017.
- [34] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How pairwise coevolutionary models capture the collective residue variability in proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, 2018.
- [35] Timetree.org. <http://www.timetree.org/search/goto-timetree>.
- [36] Stepan Denisov, Georgii Bazykin, Alexander Favorov, Andrey Mironov, and Mikhail Gelfand. Correlated Evolution of Nucleotide Positions within Splice Sites in Mammals.
- [37] Hiroaki Iwata and Osamu Gotoh. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics*, 12(1):45, 2011.
- [38] Schraga H. Schwartz, João Silva, David Burstein, Tal Pupko, Eduardo Eyra, and Gil Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, 18(1):88–103, 2008.
- [39] C. L. Fowlkes, E. B.; Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, (78(383)):553–42, 1983.
- [40] Wayne P. Maddison and Montgomery Slatkin. Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution*, (45):1184–1197, 1991.
- [41] Sankoff D. Minimal mutation trees of sequences. *Journal Applied Math*, (28(1)):35–42, 1975.
- [42] Jade Sales-Lee, Daniela S Perry, Bradley A Bowser, Jolene K Diedrich, Beiduo Rao, Irene Beusch, John R Yates, Scott W Roy, and Hiten D Madhani. Coupling of spliceosome complexity to intron diversity. *Current Biology*, 31(22):4898–4910, 2021.
- [43] Saurabh Chaudhary, Waqas Khokhar, Ibtissam Jabre, Anireddy S.N. Reddy, Lee J. Byrne, Cornelia M. Wilson, and Naeem H. Syed. Alternative splicing and protein diversity: Plants versus animals. *Frontiers in Plant Science*, 10(June):1–14, 2019.

- [44] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtkova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008.
- [45] Srikar Chamala, Guanqiao Feng, Carolina Chavarro, and W. Brad Barbazuk. Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Frontiers in Bioengineering and Biotechnology*, 3(MAR):33, mar 2015.
- [46] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, dec 2008.
- [47] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, dec 2012.
- [48] Julian Vosseberg and Berend Snel. Domestication of self-splicing introns during eukaryogenesis: the rise of the complex spliceosomal machinery. *Biology Direct*, 2017.
- [49] Raquel F. Carvalho, Dóra Szakonyi, Craig G. Simpson, Inês C.R. Barbosa, John W.S. Brown, Elena Baena-González, and Paula Duque. The arabidopsis SR45 splicing factor, a negative regulator of sugar signaling, modulates SNF1-related protein kinase 1 stability. *Plant Cell*, 28(8):1910–1925, aug 2016.
- [50] Tom Laloum, Guiomar Martín, and Paula Duque. Alternative Splicing Control of Abiotic Stress Responses. *Trends in Plant Science*, 23(2):140–150, 2018.
- [51] Jinbu Jia, Yanping Long, Hong Zhang, Zhuowen Li, Zhijian Liu, Yan Zhao, Dongdong Lu, Xianhao Jin, Xian Deng, Rui Xia, Xiaofeng Cao, and Jixian Zhai. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nature Plants*, 6(7):780–788, jul 2020.
- [52] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.
- [53] AP Dalke. chemfp-fast and portable fingerprint formats and tools. *Journal of cheminformatics*, 3(1):1–1, 2011.
- [54] David J Rogers and Taffee T Tanimoto. A computer program for classifying plants: The computer is programmed to simulate the taxonomic process of comparing each case with every other case. *Science*, 132(3434):1115–1118, 1960.

- [55] Ariel José Berenstein, María Paula Magariños, Ariel Chernomoretz, and Fernan Agueero. A multilayer network approach for guiding drug repositioning in neglected diseases. *PLoS neglected tropical diseases*, 10(1):e0004300, 2016.
- [56] Lionel Urán Landaburu, Ariel J Berenstein, Santiago Videla, Parag Maru, Dhannasekaran Shanmugam, Ariel Chernomoretz, and Fernán Agüero. Tdr targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. *Nucleic acids research*, 48(D1):D992–D1005, 2020.
- [57] User:Jü. File:amphetamine methylhexanamine similarity v.2.svg. https://commons.wikimedia.org/wiki/File:Amphetamine_%26_Methylhexanamine_similarity_V.2.svg, 2014. Accessed: 2024-1-24.
- [58] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne LeFebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.
- [59] Noel M O’Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.
- [60] Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.
- [61] Guy W Bemis and Mark A Murcko. Properties of known drugs. 2. side chains. *Journal of medicinal chemistry*, 42(25):5095–5099, 1999.
- [62] Oliver B Scott and AW Edith Chan. Scaffoldgraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics*, 36(12):3930–3931, 2020.
- [63] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.
- [64] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.
- [65] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.
- [66] Stijn Marinus Van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2000.

- [67] Thomas Shafee. File:ortholog paralog analog examples.svg. https://commons.wikimedia.org/wiki/File:Ortholog_paralog_analog_examples.svg, 2018. Accessed: 2024-1-24.
- [68] Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.
- [69] Ayan Acharya, Mukul Yadav, Mithilesh Nagpure, Sanathanalaxmi Kumaresan, and Sankar K. Guchhait. Molecular medicinal insights into scaffold hopping-based drug discovery success. *Drug Discovery Today*, 29(1):103845, 2024.
- [70] Hans-Joachim Böhm, Alexander Flohr, and Martin Stahl. Scaffold hopping. *Drug Discovery Today: Technologies*, 1(3):217–224, 2004.
- [71] Shuangjia Zheng, Zengrong Lei, Haitao Ai, Hongming Chen, Daiguo Deng, and Yuedong Yang. Deep scaffold hopping with multimodal transformer neural networks. *Journal of cheminformatics*, 13:1–15, 2021.
- [72] Xuhan Liu, Kai Ye, Herman WT van Vlijmen, Adriaan P IJzerman, and Gerard JP van Westen. Drugex v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. *Journal of Cheminformatics*, 15(1):24, 2023.
- [73] Hailin Chen, Zuping Zhang, and Jingpu Zhang. In silico drug repositioning based on the integration of chemical, genomic and pharmacological spaces. *BMC bioinformatics*, 22(1):1–12, 2021.
- [74] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, et al. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.
- [75] Lun Yang and Pankaj Agarwal. Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12):e28025, 2011.
- [76] Kayleigh D Rutherford, Gaston K Mazandu, and Nicola J Mulder. A systems-level analysis of drug–target–disease associations for drug repositioning. *Briefings in functional genomics*, 17(1):34–41, 2018.
- [77] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

- [78] Tao Zhou, Zoltán Kuscik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.
- [79] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488, 2009.
- [80] Riccardo Marcaccioli and Giacomo Livan. A pólya urn approach to information filtering in complex networks. *Nature communications*, 10(1):745, 2019.
- [81] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Inf. Process. Lett.*, 90:215–221, 2004.
- [82] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.
- [83] Russell Merris. Laplacian matrices of graphs: a survey. *Linear algebra and its applications*, 197:143–176, 1994.
- [84] Russell Merris. A survey of graph laplacians. *Linear and Multilinear Algebra*, 39(1-2):19–31, 1995.
- [85] Chen Gong, Tongliang Liu, Dacheng Tao, Keren Fu, Enmei Tu, and Jie Yang. Deformed graph laplacian for semisupervised learning. *IEEE transactions on neural networks and learning systems*, 26(10):2261–2274, 2015.
- [86] Guoji Fu, Peilin Zhao, and Yatao Bian. p -laplacian based graph neural networks. In *International Conference on Machine Learning*, pages 6878–6917. PMLR, 2022.
- [87] Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p -laplacian. In *Proceedings of the 26th annual international conference on machine learning*, pages 81–88, 2009.
- [88] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- [89] Daniel Spielman. Spectral graph theory. *Combinatorial scientific computing*, 18:18, 2012.