UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Anotación Funcional de Proteínas basada en Representación Relacional en el entorno de la Biología de Sistemas

Autor: Beatriz García Jiménez

Directores: Dra. Araceli Sanchis de Miguel
Dr. Alfonso Valencia Herrera

DEPARTAMENTO DE INFORMÁTICA

Leganés, Junio 2012

ANOTACIÓN FUNCIONAL DE PROTEÍNAS BASADA EN REPRESENTACIÓN RELACIONAL EN EL ENTORNO DE LA BIOLOGÍA DE SISTEMAS

AUTOR
BEATRIZ GARCÍA JIMÉNEZ

DIRECTORES
ARACELI SANCHIS DE MIGUEL
Departamento de Informática
Universidad Carlos III de Madrid (UC3M)

ALFONSO VALENCIA HERRERA
Programa de Biología Estructural y Biocomputación
Centro Nacional de Investigaciones Oncológicas (CNIO)

TESIS DOCTORAL

ANOTACIÓN FUNCIONAL DE PROTEÍNAS
BASADA EN REPRESENTACIÓN RELACIONAL
EN EL ENTORNO DE LA BIOLOGÍA DE SISTEMAS

Autor: Beatriz García Jiménez

Directores: Dra. Araceli Sanchis de Miguel
Dr. Alfonso Valencia Herrera

Tribunal Calificador		Firma
Presidente:	.....	.....
Vocal:	.....	.....
Vocal:	.....	.....
Vocal:	.....	.....
Secretario:	.....	.....

Calificación: .....

Leganés, ... de ..... de 2012

A mis padres

"Las grandes obras son hechas no con la fuerza,
sino con la perseverancia."

Samuel Johnson

Agradecimientos

Gracias a mis padres. Por dármelo todo a cambio de nada. Por dejarme llegar hasta el final, aislada, sin preguntar ni compartir tiempo con vosotros, quizá demasiado tiempo. A mi hermana María, por facilitarme la vida del día a día, por los ánimos, por las horas que no te he dado. Al resto de mi familia.

A Araceli y Alfonso, por llevar hasta el final una tesis en bioinformática dirigida en la distancia. Araceli, gracias por dejarme elegir libremente mi tema de tesis, por tu confianza y apoyo incondicional, por tus ánimos y tus consejos; y por recordarme siempre que lo mejor es enemigo de lo bueno. Alfonso, gracias por permitirme trabajar contigo y por dirigir mi tesis, por todo lo que he aprendido de ti sobre biología computacional y sobre investigación, por resolver siempre mis dudas en algún momento, y por plantearme retos continuos para llegar a finalizar esta complicada tesis con significado biológico real, a pesar de las múltiples dificultades.

A Tirso, gracias por ayudarme a finalizar parte de este trabajo. Realmente me ha encantado trabajar contigo a diario buscando la interpretación biológica de las proteínas. A David, Iakes y Edu, por las largas reuniones siempre desesperanzadoras, de las que finalmente empecé a aprender a investigar en biología molecular. Gracias de verdad David.

A M.Paz, con la que he pasado más horas que con nadie estos años. Por escucharme, por ayudarme, por aconsejarme, por enseñarme, por animarme, por informarme, por protegerme,... Estamos unidas por más de lo que pensamos.

A Lourdes y Sandra, gracias por seguir ahí. Lourdes, gracias por tu admiración y creencia infinita en mi trabajo y en mí; acepta una fuerza e inteligencia que no está enfrente, sino dentro de ti.

A Anaïs Baudot y Enrico Glaab, por sus interesantes y esclarecedores comentarios. A los desarrolladores de las herramientas ACE y CLUS por su importante ayuda, especialmente Daan Fierens, Jan Struyf y Leander Schietgat. A Giuliano Armano, por todos los contactos que hice durante mi estancia. A Ross King por sus consejos.

A Agapito, por las sugerencias, por estar ahí puntualmente a lo largo de la tesis. A Ricardo, por la primera publicación de mi tesis.

A Paula, por dejarme compartir despacho contigo y por tus pragmáticos consejos al final de la tesis. A Jose, por tu oportuna ayuda en la recta final y el optimismo que emanas. A Germán, por estimar y respetar mi trabajo siempre. A Juan, quien me iba a decir que serías mi primer compañero de docencia, del que aprendí. A Jorge, por las conversaciones necesarias. Al resto de compañeros de CAOS, de SCALAB y de la UC3M, que alguna vez me animaron o aconsejaron.

A mi grupo de biólogos y bioinformáticos del Centro Nacional de Investigaciones Oncológicas: los que aún están y los que estaban en mis inicios en el Centro Nacional de Biotecnología, especialmente a José María, José Manuel, Ángela y Almudena (gracias por acoger a la *chica de los seminarios*). A todos los que alguna vez me escucharon y ayudaron, gracias por vuestras ideas, consejos y resolución de dudas completamente desinteresada.

A pilates y baile, por empezar la tesis conmigo y acompañarme todo este tiempo. A Leganés Camina que se unió después. Gracias a todos por los inevitables ratos de dispersión, necesarios para continuar con la tesis. A Agustín, mi amigo incremental, gracias por tu realismo. A Tania, gracias por tus ánimos y tus relatos. A Afluentes, gracias por los interesantes, instructivos e inolvidables momentos sobre la vida; Geraldina, gracias por ser tú, por haber compartido parte de tu apreciado tiempo conmigo.

A mi nueva familia de allá, que me acogen sin apenas conocerme. Por apoyarme en la preparación de la presentación de esta tesis en plena Navidad.

Gracias a todos mis amigos y conocidos que me han dado ánimos y se han interesado por el desarrollo de esta tesis.

Gracias, gracias, gracias Tomás. Por todo lo que sabes que has representado en esta tesis, y lo que significas en mi vida. Por tu apoyo, crítica y ayuda fundamental en la tesis. Por las largas reflexiones. Por reavivarme la pasión por la curiosidad, por el conocimiento y por la vida. Por estar siempre ahí en espera activa. Por todo lo que no te puedo agradecer en sólo unas líneas. Por supuesto, por los múltiples detalles también. Por fin, el momento que tanto hemos esperado de *"después de la tesis"* ya está aquí...

Resumen

La anotación funcional es un tema de investigación abierto e importante en Biología Molecular. El problema de definir función a nivel de terminología es complicado, puesto que la función ocupa muchos niveles para una misma proteína y no existe un criterio unificado. Ante estas dificultades, la forma de determinar la función de una proteína es anotarla con distintos términos en diferentes vocabularios.

Las proteínas desarrollan su función en cooperación con otras proteínas formando complejos. Estas interacciones se representan en una red, formada por interacciones que han sido demostradas experimentalmente entre proteínas. Analizar y utilizar la red de interacciones es una tarea de interés debido al gran número de asociaciones existentes, y a las múltiples formas en que una proteína puede influir en la función de otras. Por lo tanto, esta tesis se centra en la predicción de anotación funcional basada en redes.

Es evidente que este complejo escenario no puede afrontarse sin el uso de herramientas computacionales. De hecho existe una actividad considerable en el área de Biología Computacional dedicada específicamente a este tema. Esta tesis es parte de este esfuerzo en la aplicación de métodos computacionales a problemas biológicos en el área de Biología de Sistemas. Esta aproximación puede enmarcarse en este contexto de la Biología de Sistemas, puesto que no se analiza la función de forma aislada para cada molécula, sino a nivel de sistema, teniendo en cuenta todas las relaciones existentes entre genes y proteínas conectados a distintos niveles. Para aprovechar todas estas relaciones biológicas, y mantener su semántica estructural, esta tesis plantea usar Representación Relacional, por ser un dominio particularmente apropiado para ello. A partir de dicha representación se aplican múltiples transformaciones y técnicas de Inteligencia Artificial para extraer conocimiento de las proteínas relacionadas, y proponer nuevas funciones a través de la predicción de asociaciones funcionales entre proteínas.

La propuesta general de esta tesis es la caracterización de función de proteínas y genes basándose en información de redes, a través de la Representación Relacional y el Aprendizaje Automático. En concreto, partiendo de una representación relacional para anotación funcional, se busca el diseño computacional necesario para resolver dos problemas concretos, diferentes e interesantes en Biología. Uno es la predicción de asociaciones funcionales entre pares de proteínas en *E.coli*, y el otro la extensión de rutas biológicas en humanos. Ambos se evalúan en términos computacionales y de interpretación biológica. También se proponen nuevas anotaciones funcionales de proteínas a ser verificadas experimentalmente. Además, se exploran diversos enfoques en la representación del conocimiento y en las técnicas de aprendizaje, proponiendo estrategias concretas para resolver otros problemas bioinformáticos, especialmente influenciados por la información relacional y el aprendizaje multi-clase y multi-etiqueta.

Abstract

Functional annotation is an open and interesting research topic in Molecular Biology. Determining a function in terminology terms is a hard task, due to lack of unified criterion and also because a function takes up many levels for the same protein. Given this difficulties, the way to determine a protein function is to annotate it with several terms from different vocabularies.

Proteins carry out their function together with other proteins, being part of protein complexes. These interactions are represented in a network of experimentally verified protein-protein interactions. Analyzing and using the interaction network is task of interest due to the great number of associations, and to the multiple ways in which a protein could influence in the function of others. Therefore, this thesis focuses in the prediction of functional annotation based on networks.

It's apparent that this complex scenario couldn't be faced without computational techniques. In fact, in Computational Biology, there is a considerable activity specially devoted to this topic. This thesis is part of this effort for applying computational methods to biological problems in the Systems Biology area. This approximation can belong to the Systems Biology context, because it does not analyze function in an isolated way for each molecule, but at system level, taking into account all the relations among genes and proteins linked at different levels. To take advantages of all these biological relations, and to preserve their structured semantics, this thesis suggests to use Relational Representation, since in particular it is suitable for the concerning domain. Over such representation, multiple transformations and Artificial Intelligence techniques are applied to retrieve implicit knowledge from the related proteins, and to propose new functions through the prediction of functional associations between proteins.

The main proposal of this thesis is to characterize the function of proteins and genes based on networks, through Relational Representation and Machine Learning. Specially, from a relational representation specific to functional annotation, we look for the computational design needed to solve two specific, biological interesting and different problems. The former consists of predicting functional association between pair of proteins in *E.coli*, and the latter comprises expanding pathways in humans. We perform an assessment in computational and biological interpretation terms. Besides, we propose new putative protein functional annotations to be experimentally verified. In addition, the thesis investigates diverse approaches to knowledge representation and learning techniques, suggesting specific strategies to tackle other biological problems, specially where relational data or multi-class and multi-label targets are present.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En la visión de abajo a arriba de la **Biología Molecular** clásica se analizan las moléculas de forma independiente, con el objetivo de identificar los responsables biológicos de todos y cada uno de los procesos que suceden continuamente en un organismo vivo. Este interés radica en que se trata de un conocimiento esencial, tanto para comprender el funcionamiento interno de todas las especies, como para poder diseñar fármacos que solventen las enfermedades y los desórdenes metabólicos, interviniendo sobre los genes y proteínas responsables.

El origen de la **Bioinformática** se puede situar incluso antes de los 70 y de la acumulación de datos [Ouzounis and Valencia, 2003], por la necesidad de utilizar procedimientos computacionales para analizar los diversos, complejos y heterogéneos datos biológicos. No obstante, el uso de la Inteligencia Artificial en la Bioinformática se ha vuelto más relevante en los últimos años por la ingente cantidad de datos biológicos, que aparecen y que crecen exponencialmente, procedentes de la experimentación a gran escala con las denominadas tecnologías de generación masiva automatizada de resultados (del inglés, *high-throughput technologies*), que actualmente también permiten la construcción de grandes redes complejas. Todos estos datos se pueden manejar manualmente uno a uno por científicos experimentales, como se ha hecho tradicionalmente. Pero el uso de la informática resulta imprescindible, no sólo para automatizar procesos, sino también para facilitar el análisis y extracción de conocimiento de toda esta información de carácter biológico.

Existen diversas definiciones de Bioinformática y Biología Computacional, a veces considerándolos sinónimos y otras no [Baldi and Brunak, 2001], dependiendo de la cantidad de implicaciones biológicas que conlleve el problema frente a un simple uso de habilidades técnicas. Aunque, desde un punto de vista informático, frecuentemente se simplifica su definición a la aplicación de técnicas computacionales a datos biológicos, en esta tesis se considera otra de las definiciones más amplia. La definición de Bioinformática o Biología Computacional que se maneja en esta tesis es la de un campo multidisciplinar que utiliza técnicas computacionales, matemáticas y estadísticas para tratar problemas de Biología Molecular, afrontando temas de investigación científica, sin olvidar todas sus cuestiones intrínsecas, teóricas y experimentales.

Dentro de la Bioinformática o la Biología Computacional existen múltiples áreas genéricas de investigación, que a muy alto nivel incluyen: localización de genes (regiones que codifican proteínas), determinación de sitios de ensamblaje alternativo, localización celular, predicción de estructura secundaria y tridimensional, determinación de sitios funcionales, predicción de anotación de función de genomas y proteomas, identificación de familias y dominios de proteínas, construcción y predicción de redes de interacción de proteínas y de redes de

regulación génica, entre otros. Dentro de cada una de estas áreas hay muchas divisiones posibles en problemas concretos, porque cada una representa un amplio conjunto de conceptos.

Desde un punto de vista global, el mundo de las **redes** [Newman, 2010] está revolucionando múltiples ámbitos de la vida actual, abarcando desde las redes de ordenadores y comunicaciones, Internet (red de documentos digitales hiper-enlazados), las redes formadas por artículos científicos referenciados entre sí, las redes profesionales, las redes sociales, ..., hasta las redes biológicas a diferentes niveles. Todas ellas comparten una estructura formada por nodos (entidades: ordenador, página web, documento, persona, molécula o célula, etc.), y arcos (interacciones entre pares de entidades: conexión eléctrica, enlace, referencia, amistad, reacción bioquímica o conexión funcional, respectivamente, etc.) [Leskovec, 2008]. Muchas de estas redes complejas de diferente naturaleza comparten ciertas propiedades universales por ser redes libres de escala [Barabasi and Bonabeau, 2003]: la existencia de algunos nodos populares, con cientos o miles de interacciones frente a las dos o tres de la mayoría de nodos, robustez frente a fallos puntuales, pero vulnerabilidad a ataques coordinados. Por ejemplo, enfermedades como el cáncer o Internet son sistemas que comparten el hecho de que un fallo en un nodo clave (un gen o un servidor) puede desencadenar un problema grave [Solé, 2009]. Por otro lado, el análisis de la red completa proporciona nueva información en forma de propiedades emergentes de la red, que no se podrían extraer de los elementos individuales que la componen. De forma que el estudio en conjunto, de estas redes complejas y sus propiedades, proporciona una nueva visión opuesta al análisis de las entidades individuales de forma aislada, que promete una mayor comprensión de los sistemas de redes, y nuevas aplicaciones en todos los ámbitos donde se encuentran; desde la prevención de un ataque informático hasta la de una enfermedad.

En Biología Molecular existen muchas redes, porque por todas partes hay interacciones y asociaciones entre las moléculas, de diversos tipos y a distintos niveles. Las interacciones proteína-proteína relacionan pares de proteínas a nivel físico, y los complejos de proteínas relacionan grupos de más de dos proteínas también físicamente, formando extensas y complicadas redes de interacción en un organismo. Otros ejemplos de relaciones entre proteínas o genes son las rutas biológicas (del inglés, *pathways*), que asocian funcionalmente proteínas que participan en distintos pasos de una cadena de reacciones, como por ejemplo el transporte a través de la membrana celular. Existe una larga lista de posibles relaciones biológicas que permiten definir grupos y redes de genes o proteínas, caracterizados por participar en un mismo proceso o por compartir un valor común para un criterio dado. Algunos ejemplos más son: genes en la misma red de regulación génica, proteínas con la misma localización celular, proteínas de la misma familia o con un tipo de dominio compartido, o genes con datos fenotípicos comunes (mismo tejido o implicación en una enfermedad).

Hace poco más de una década surge la **Biología de Sistemas** [Kitano, 2000], con la idea de modelar circuitos, que permitan definir la estructura, dinámica y simulación de los sistemas biológicos. Posteriormente también se añade a esta disciplina el estudio de las redes biológicas. Es un área que estudia los organismos vivos a un nivel de sistema, desde arriba hacia abajo y no a la inversa, como se hacía hasta entonces. Trata de afrontar las limitaciones de la visión de abajo a arriba, siguiendo el lema *'el todo es más que la suma de las partes'*, con muchos aspectos a analizar: topología de la red, propiedades emergentes, cómo construir o extender la red de interacciones, su influencia en la anotación funcional, etc.

La estructura de la red de interacción es muy valiosa y ha sido ampliamente estudiada [Rojas et al., 2006], por contener mucho conocimiento sobre las colaboraciones que se producen entre proteínas para llevar a cabo una determinada función. Generalmente no existe un único gen o producto genético responsable de una función o proceso biológico en el

organismo, sino que dicho proceso es el resultado de la combinación de la acción de varios genes o productos genéticos. De ahí la relevancia de las relaciones entre genes y proteínas, cuyo uso se pretende incentivar en esta tesis doctoral.

Para el análisis automático con técnicas computacionales y de Inteligencia Artificial de los datos biológicos, que se encuentran altamente relacionados en redes de interacciones, resulta adecuado partir de una **Representación Relacional** [Dzeroski and Lavrac, 2001]. Es la forma natural de representar las interacciones, asociaciones funcionales y las propiedades correspondientes a los nodos o elementos biológicos, de múltiples y variadas fuentes. Una Representación Relacional aprovecha el conocimiento implícito en todas las relaciones biológicas existentes, evitando perder su estructura y semántica, como le sucede a muchas otras técnicas. Para la extracción automática de conocimiento, se ha extendido exitosamente el Aprendizaje Automático [Mitchell, 1997] dentro del área de Inteligencia Artificial. En particular, el Aprendizaje Automático Relacional [Dzeroski and Lavrac, 2001] permite preservar fácilmente el origen estructurado de los elementos biológicos y la semántica de sus relaciones, al contrario que el Aprendizaje Automático Proposicional clásico, la metodología más frecuentemente utilizada, que tiene que plasmar toda la información en una única tabla atributo-valor [Dzeroski, 2003].

Ante el todavía vigente reto de la Biología Molecular de caracterizar funcionalmente el genoma y el proteoma, la **predicción de anotación funcional** (en la que se centra esta tesis) es un área de investigación de interés biológico muy amplia. Esta predicción podría ser útil para ordenar y/o seleccionar el subconjunto de las asignaciones de función más probables o relevantes, para ser verificadas en un laboratorio experimental. Así, se pueden reducir en gran medida los costes de los experimentos *in-vivo* de determinación de función, restringidos a unos cuantos frente a todas las posibles anotaciones funcionales.

Por lo tanto, en esta tesis se quiere conocer cómo afecta el contexto relacional y su uso a la predicción de anotación funcional del genoma y el proteoma, intentando explotar esta gran cantidad de relaciones de diversa índole que ocurren (de forma estable o transitoria) entre genes y proteínas, de forma binaria o en grupos. La propuesta de tesis es estudiar el área de predicción de anotación funcional centrado en la Biología de Sistemas, a través de la Representación Relacional y el Aprendizaje Automático. Se pretende usar redes biológicas o predecir parte de ellas, considerando una interacción y una asociación funcional como un tipo de anotación; o también se puede considerar una extensión de la red, con interacciones o asociaciones funcionales punto a punto o grupales.

La anotación funcional se puede descomponer en múltiples dominios o problemas concretos, como son la predicción de función molecular, la predicción de proceso biológico, la predicción de fenotipo, la predicción de localización celular, la predicción de participación en una ruta metabólica, etc. Algunos de estos dominios se abordan durante el desarrollo de esta tesis doctoral, pero desde una perspectiva común de predicción de anotación funcional con/en redes de interacción.

Este documento contiene un proyecto de tesis doctoral, estructurado de la forma que se detalla a continuación. En el capítulo 2 se describe el estado del arte, tanto del contexto biológico de anotación funcional y la Biología de Sistemas, como del Aprendizaje Automático, especialmente Relacional, principal enfoque computacional utilizado en esta tesis. En el capítulo 3 se exponen los objetivos de la tesis. En el capítulo 4 se describe la metodología de evaluación propuesta para este trabajo. Los capítulos 5 al 8 presentan las aportaciones principales de esta tesis, describiendo una propuesta de representación relacional genérica de datos biológicos para anotación funcional (capítulo 5), la resolución de dos problemas

diferentes que combinan Biología de Sistemas, Representación Relacional y Anotación Funcional (capítulos 6 y 7), y la exploración de diferentes enfoques de Aprendizaje Automático en problemas bioinformáticos (capítulo 8). En el capítulo 6 el problema afrontado es la predicción de asociaciones funcionales entre pares de proteínas en *E.coli* y en el capítulo 7, la extensión de rutas biológicas en humanos. Posteriormente, el capítulo 9 presenta un análisis y discusión de la anotación funcional con datos relacionales, derivada de los problemas afrontados en los capítulos previos. Por último, los capítulos 10 y 11 exponen las conclusiones generales y líneas futuras de investigación.

Capítulo 2

Estado del Arte

En este capítulo se presenta una introducción general a las áreas principales dentro de las que se desarrolla este trabajo bioinformático. Las dos primeras secciones describen el punto de vista computacional, introduciendo el Aprendizaje Automático Proposicional y principalmente describiendo el Aprendizaje Automático Relacional. En la tercera sección se combina el aspecto computacional con el biológico, revisando sus aplicaciones en Biología y los retos que presenta. Mientras que la sección cuarta resume el contexto biológico (Biología de Sistemas y Anotación Funcional). Finalmente se incluye una sección de discusión, con una recopilación de los aspectos relevantes de estudio en esta tesis y los dos problemas concretos que se afrontan.

2.1. Aprendizaje Automático Proposicional

2.1.1. Introducción al Aprendizaje Automático

El Aprendizaje Automático (AA) es una disciplina científica, dentro del campo de la Inteligencia Artificial, preocupada por mejorar el comportamiento de un sistema al realizar una tarea, mediante la adquisición de conocimiento por medio de la experiencia [Mitchell, 1997]. El AA frecuentemente se relaciona con el Análisis o Minería de Datos (del inglés, *Data Mining*), o la extracción de patrones a partir de los datos conocidos.

Según su relación con el entorno, los sistemas de AA se pueden clasificar en supervisados y no supervisados [Borrajo et al., 2006]. En el AA supervisado se aprende a partir de un conjunto de ejemplos etiquetados previamente por una fuente externa. Mientras que en el AA no supervisado, el conjunto de entradas no está etiquetado. El aprendizaje por refuerzo se encuentra a medio camino de las anteriores, aprendiendo a decidir la acción más adecuada ante una situación determinada, por medio de una serie de refuerzos externos.

Otra clasificación del AA según el tipo de razonamiento [Borrajo et al., 2006] divide los sistemas en inductivos, deductivos, abductivos o analógicos. En el AA inductivo, se parte de las observaciones existentes para expandir el conocimiento que las generó; en el AA deductivo se derivan nuevas reglas lógicas a partir del conocimiento existente; en el AA abductivo, al contrario, partiendo del conocimiento y las consecuencias, se infieren las posibles premisas que causaron éstas; y el AA analógico es una combinación del inductivo junto con el deductivo.

Una clasificación alternativa que depende más del tipo de tarea a resolver por el AA [Borrajo et al., 2006], entre las que se distinguen tres tipos básicos:

- **Clasificación**: se puede definir como "la búsqueda de una relación de correspondencia entre las observaciones y las clases"[Borrajo et al., 2006]. Se emplea cuando a cada

instancia se le debe asociar una clase o categoría, dando lugar a grupos mutuamente excluyentes; con una clase binaria (sí/no) o con diferentes valores nominales. Un ejemplo típico de esta tarea de aprendizaje es la concesión o no de un crédito bancario, en función de los datos del cliente. La **regresión** es un caso particular de la tarea de clasificación, en el que la salida no son clases discretas, sino valores numéricos en un rango continuo. Por ejemplo, predecir la cantidad de energía que se va a necesitar en una ciudad, para evitar apagones imprevistos. También son casos típicos de esta tarea la predicción de series temporales [Mitchell, 1997].

- **Caracterización**: también conocida como extracción de reglas de asociación. Lo que se quiere obtener en este caso son relaciones entre combinaciones de atributos, sin que el atributo clase tenga una importancia relevante, pudiendo existir o no. Así, se reconoce cómo la ocurrencia de un suceso (un valor de un atributo o un conjunto de ellos) puede generar la aparición de otros. La **extracción de patrones frecuentes** se puede considerar una subtarea del proceso de caracterización, ya que dichos patrones se deben buscar, para posteriormente establecer reglas de asociación entre ellos. El ejemplo clásico de la tarea de caracterización es el análisis de la cesta de la compra, que permite determinar qué productos se compran generalmente juntos, para emplearlos en las estrategias de distribución o promoción de productos de los supermercados [Agrawal and Srikant, 1994].

- **Agrupamiento (del inglés, *Clustering*)**: en este caso no hay una clase, sino que se pretenden obtener grupos de elementos, los cuales abarquen instancias que tengan gran similitud entre sí y muchas diferencias con los de otros grupos. Se obtiene generalmente un prototipo por grupo. Un ejemplo clásico es la segmentación de personas en distintos grupos, bien empleados de una empresa, o bien clientes sobre los que hacer propaganda personalizada [MacQueen, 1967; Mitchell, 1997].

Uno de los enfoques más extendidos y en el que se centra gran parte de esta tesis es el AA supervisado inductivo aplicado a la tarea de clasificación o regresión. Es decir, se parte de un conjunto de ejemplos etiquetados (supervisado) a partir de los que extraer el conocimiento existente (inductivo) que ha permitido asignar las clases a dichos ejemplos, para poder predecir la clase de futuros ejemplos (clasificación).

2.1.2. Algoritmos Proposicionales de Clasificación y Regresión

Cabe destacar que en bioinformática se utilizan con frecuencia y con reconocido éxito técnicas sub-simbólicas como las redes de neuronas artificiales (del inglés, *Artificial Neural Networks, ANN*) [Rumelhart and McClelland, 1986] [Jensen et al., 2002a; Rost and Sander, 1994] y las máquinas de vector de soporte (del inglés, *Support Vector Machine, SVM*) [Vapnik, 1998] [Lee et al., 2009; Re and Valentini, 2009].

Sin embargo, en esta tesis se presta mayor atención a la aplicación de técnicas simbólicas [Borrajo et al., 2006], como son los árboles y reglas de decisión, ampliamente usados en bioinformática [Che et al., 2011], por ser su modelo de salida más fácilmente interpretable y útil para obtener justificaciones de la predicción en términos biológicos y para facilitar el estudio experimental posterior.

Árboles y Reglas de Decisión

Se trata de algoritmos de clasificación cuyo modelo de salida está representado en un árbol de decisión o un conjunto de reglas equivalente. Para definir sucesivamente los nodos y ramas del árbol se realiza una elección sucesiva del atributo que más discrimina del grupo según la medida mínima de entropía. El autor principal de esta técnica es Quinlan, que diseñó inicialmente el algoritmo ID3 [Quinlan, 1986], y posteriormente su evolución en C4.5 [Quinlan, 1993]. C4.5 incluye mejoras (como por ejemplo la poda) que resuelven ciertos problemas que presenta ID3 frente al ruido y la sobre-adaptación.

Los árboles o reglas de regresión son una variante utilizada cuando el atributo clase es continuo, o también se puede considerar su uso si se requiere una salida numérica en vez de nominal. CART [Breiman et al., 1984] es el algoritmo original de árboles de regresión, conteniendo valores numéricos en las hojas del árbol. La selección de atributos maximiza la reducción esperada en varianza o en desviación absoluta. M5 [Quinlan, 1993] es una variante de CART que tiene en cada hoja un modelo lineal construido con regresión clásica en función de los valores de los atributos seleccionados.

Por otro lado, existe una variante *C4.5Multi-label* [Clare, 2003] que permite la clasificación de una instancia en más de una clase a la vez.

CLUS [Blockeel et al., 1998] es un sistema de inducción de reglas y árbol de decisión que implementa el marco de agrupación de predicción. Este marco unifica la agrupación no supervisada y el modelado de predicción, permitiendo una extensión a predicciones más complejas, como aprendizaje multi-tarea o clasificación multi-etiqueta. En este enfoque, un árbol de decisión se ve como una jerarquía de agrupaciones (o *clusters*). Ahí, el nodo superior corresponde a una agrupación que contiene todos los datos, que recursivamente se dividen en agrupaciones más pequeñas conforme se desciende en el árbol, maximizando la reducción de la varianza intra-cluster. Permite construir modelos de clasificación y de regresión. Dependiendo de la definición de distancia entre ejemplos, CLUS puede construir un árbol que prediga sobre varios atributos de salida a la vez, incluso permitiendo la existencia de una jerarquía entre ellos. De la misma forma, CLUS también permite realizar clasificación multi-etiqueta. Su flexibilidad y su previo uso en otras aplicaciones de anotación funcional [Vens et al., 2008] hace que CLUS sea uno de los principales algoritmos seleccionados en esta tesis para la fase experimental.

Aprendizaje Basado en Instancias

Como contraposición a los sistemas sub-simbólicos, las técnicas de aprendizaje vago o basado en instancias y los algoritmos bayesianos, también se pueden clasificar dentro del aprendizaje simbólico.

El aprendizaje basado en instancias (del inglés, *Instance-Based Learning, IBL*) [Aha et al., 1991] consiste en almacenar las instancias del conjunto de entrenamiento (todas o un subconjunto), sin construir ningún modelo concreto. Para clasificar los nuevos ejemplares se toma el valor de la clase del o de los ejemplos más próximos almacenados. Un algoritmo clásico es el de los k-vecinos más próximos (del inglés, *K-Nearest Neighbour, KNN*) [Mitchell, 1997].

Clasificadores Bayesianos

Los métodos de clasificación bayesianos se basan en el teorema de Bayes, utilizando estimaciones de probabilidades de pertenecer a una clase según el valor de cada atributo en el conjunto de entrenamiento. El clasificador bayesiano clásico es *Naive Bayes* [John and Langley, 1995], aunque existen algoritmos mejorados como las Redes Bayesianas (*BayesNet*)

[Friedman et al., 1997; Bouckaert, 2004], o AODE [Webb et al., 2005], siendo este último de especial interés en esta tesis.

- **AODE**

  AODE (del inglés, *Averaged One-Dependence Estimators*) [Webb et al., 2005] es un algoritmo de aprendizaje bayesiano o basado en probabilidades condicionadas. AODE calcula la media de un pequeño conjunto de modelos alternativos bayesianos simples, que tienen una asunción de independencia más débil que el clásico *Naive Bayes* (NB) [John and Langley, 1995]. De esta forma se evita el sesgo que provoca el incumplimiento de la independencia total de atributos en NB, mejorando los resultados de predicción y manteniendo la eficiencia computacional, con sólo un pequeño incremento en la varianza.

  Otros algoritmos como LBR (del inglés, *Lazy Bayes Rules*) [Zheng and Webb, 2000] y SP-TAN (del inglés, *Super Parent Tree Augmented Naive bayes*) [Keogh, 1999] han demostrado mejoras sobre los resultados de NB, aliviando el aumento de error por independencia de atributos. Sin embargo, ambos enfoques son muy costosos computacionalmente, LBR en tiempo de clasificación y SP-TAN durante el entrenamiento. AODE es una técnica novedosa que exige una menor asunción de independencia entre atributos que NB, y alcanza resultados comparables a los de LBR y SP-TAN sin incurrir en elevados costes computacionales, lo cual es deseable ante grandes conjuntos de datos, como los que se manejan en Biología Molecular.

  AODE se inspira en la noción de estimadores de dependencia n [Sahami, 1996]. Un estimador de dependencia n es similar a NB salvo que cada atributo depende como máximo de otros 'n' atributos, aparte de la clase. NB es un estimador de dependencia $n = 0$, a diferencia de TAN (del inglés, *Tree Augmented Naive bayes*) [Friedman et al., 1997], que es un estimador de dependencia $n = 1$. Estimadores de mayor dependencia típicamente tienen menor sesgo, pero mayor varianza y coste computacional que NB.

  AODE evita este aumento de la complejidad computacional porque no realiza una selección del mejor modelo. En su lugar, calcula la media de todos los modelos plausibles en los que todos los atributos dependen sólo de la clase y de un atributo más, común a todos los modelos, (sólo $n = 1$ para mantener la eficiencia). De esta forma, además se disminuye la varianza frente a LBR y SP-TAN (a costa de un cierto incremento del sesgo). Esto es así porque, al seleccionar sólo un modelo, se escoge el que se adapta mejor a los datos de entrenamiento, pero si éstos cambiaran, las variaciones en los resultados serían mayores, incrementando así la varianza.

2.1.3. Algoritmos Proposicionales de Caracterización

Los algoritmos de caracterización extraen reglas de asociación. Una regla de asociación es una expresión $X \Rightarrow Y$ donde X e Y son conjuntos de elementos [Agrawal et al., 1993]. El significado intuitivo de estas reglas indica que un conjunto de datos que cumple X, también tiende a cumplir Y. Un conjunto de elementos (X o Y) es una serie de atributos binarios con valor verdadero. Esta tarea de AA extrae reglas semejantes a las de clasificación, pero en las que en la parte derecha de la regla puede aparecer cualquier atributo que no sea la clase; así se pueden obtener relaciones entre atributos, no sólo entre los atributos y la clase. Los algoritmos de caracterización tratan de extraer todas las reglas de asociación que satisfagan un cierto nivel frecuencia y confianza.

El algoritmo más representativo es APRIORI [Agrawal et al., 1996]. Se divide en dos pasos: primero, calcular las frecuencias de los atributos y combinaciones de atributos; segundo, elegir las reglas más frecuentes y que superen unos umbrales mínimos de aceptación (confianza y soporte o frecuencia). En la búsqueda de patrones frecuentes, se hace una búsqueda en anchura por niveles, yendo de la generación de patrones más generales a más específicos. Se hace una iteración por niveles, generando especializaciones de los patrones que ya son frecuentes en el nivel anterior, podando los infrecuentes. Finalmente, se calculan las reglas de asociación, descomponiendo en dos partes los patrones frecuentes, y verificando una confianza mínima de que se cumpla el patrón completo frente a sólo la parte izquierda de la regla.

La búsqueda por niveles sólo necesita recorrer la base de datos $k+1$ veces, siendo k el número de niveles de longitud o profundidad de los patrones, porque todos los candidatos de un nivel se evalúan en una sola pasada. Esta característica permite reducir las dimensiones y complejidad, y ser aplicado a grandes conjuntos de datos, lo cual es muy importante en dominios de Biología Molecular, como los afrontados en esta tesis.

Aplicado al contexto de esta tesis, de anotación funcional del proteoma, esta tarea de AA permite obtener relaciones (patrones frecuentes o reglas de asociación) entre los genes o proteínas de un conjunto con una misma anotación, a partir de los atributos simples. Por ejemplo, se pueden buscar reglas de asociación entre los elementos de una ruta metabólica particular, o los que tienen asociada una enfermedad, o los que tienen un perfil de expresión similar.

2.2. Aprendizaje Automático Relacional

2.2.1. Definición

Mientras que la mayoría de enfoques en Aprendizaje Automático clásico (del inglés, *Machine Learning*) y Análisis de Datos (del inglés, *Data Mining*) buscan patrones en una única tabla de datos, el Aprendizaje Automático Relacional (AAR) o Aprendizaje Relacional [Dzeroski and Lavrac, 2001; Raedt, 2008] busca patrones que involucran múltiples relaciones de una base de datos relacional. La entrada de los algoritmos de AAR generalmente son varias tablas, no sólo una. Para enfatizar este hecho, frecuentemente se denomina Aprendizaje Automático Multi-Relacional [Dzeroski, 2003] [1].

En el ámbito de este trabajo los términos ingleses *'Relational Data Mining'*, *'Logical Learning'* y *'Relational Learning'* se unifican en el término Aprendizaje Automático Relacional (o su simplificación Aprendizaje Relacional). Además de que la diferencia entre estos conceptos es muy difusa, en este trabajo se considera que, según algunas tendencias recientes [Raedt, 2008], lo importante es combinar dos disciplinas científicas de la Inteligencia Artificial, como son el Aprendizaje Automático y la Representación del Conocimiento. Es decir, estudiar el Aprendizaje Automático y el Análisis de Datos, pero con una representación más expresiva, como es la relacional.

Las dos características diferenciadoras del AAR son un lenguaje de representación más expresivo y el uso de conocimiento del dominio. En el AAR el lenguaje de representación del conocimiento debe permitir representar los datos distribuidos en varias tablas y con todas sus relaciones. Para ello, el enfoque que se ha utilizado mayoritariamente se basa en la programación lógica inductiva (del inglés, *Inductive Logic Programming, ILP* [Muggleton, 1991], proveniente de la intersección entre el Aprendizaje Automático y la Programación

[1] En este documento se utilizan indistintamente los términos 'relacional'y 'multi-relacional'.

Lógica (del inglés, *Logic Programming* [Lloyd, 1987]). En el Aprendizaje Automático atributo-valor o Aprendizaje Automático Proposicional (AAP) se aprende de una sola tabla. Su denominación proviene del lenguaje de representación utilizado, que es la lógica proposicional o de orden cero. Mientras que en el AAR el lenguaje de representación mayoritario es un subconjunto de la lógica de primer orden o lógica de predicados. Esta lógica incluye predicados y variables no presentes en la proposicional, haciendo esta representación más expresiva. Así, el término *relación* en una base de datos relacional se corresponde con el término *predicado* en la representación en lógica de predicados, y los *'atributos de una relación'* con los *'argumentos de un predicado'*, respectivamente.

Respecto al conocimiento del dominio, dado que la entrada a los algoritmos de AAR está expresada en lógica de predicados, además de la información propia de los ejemplos, se puede incluir otra información de contexto en nuevos predicados lógicos.

Partiendo de una base de datos relacional, los datos se pueden transformar a representación proposicional para aplicar una técnica de Aprendizaje Automático atributo-valor. Transformar a representación proposicional [Dzeroski and Lavrac, 2001] consiste en integrar los datos de varias tablas con sus relaciones en una única tabla, mediante uniones y agregados, generalmente implicando una pérdida de información o semántica, que se evita con el AAR.

Los modelos de representación de los patrones aprendidos a partir de una tabla se han extendido para múltiples tablas. Por ejemplo, existen reglas de asociación relacionales, árboles de decisión relacionales y reglas de clasificación relacionales, entre otros. Igualmente, los algoritmos que generan estos modelos se han generalizado para ser aplicables a datos relacionales, manteniendo el algoritmo proposicional original como un caso particular. Así, existen algoritmos para inducir:

- *Reglas de decisión relacionales*: FOIL [Quinlan and Mostow, 1990] e ICL [Raedt and Laer, 1995], extendidos del algoritmo proposicional CN2 [Clark and Niblett, 1989].
- *Árboles de decisión relacionales*: S-CART [Kramer, 1996], extendido de CART [Breiman et al., 1984]; y TILDE [Blockeel and Raedt, 1998], extendido de C4.5 [Quinlan, 1993].
- *Patrones frecuentes y reglas de asociación relacionales*: WARMR [Dehaspe and Raedt, 1997], extendido de APRIORI [Agrawal et al., 1996].
- *Aprendizaje basado en distancias relacionales*: RIBL [Emde and Wettschereck, 1996], extendido de KNN [Mitchell, 1997].
- *Aprendizaje por refuerzo relacional*: RRL [Dzeroski et al., 2001] extendido del método *Q-learning* [Kaelbling et al., 1996].

Este tesis se centra principalmente en los árboles de decisión relacionales. Éstos se pueden construir en un solo paso con TILDE [Blockeel and Raedt, 1998], o en dos pasos, a través de la extracción de patrones frecuentes con WARMR y el uso de un árbol de decisión proposicional, como el ya descrito CLUS, con cláusulas lógicas en los nodos, para que sea relacional, como se explica en más detalle dentro del apartado 2.2.5.

2.2.2. Ventajas frente a Aprendizaje Automático Proposicional

El Aprendizaje Automático Multi-Relacional presenta las siguientes ventajas frente al enfoque proposicional o atributo-valor, siendo muchas de ellas relevantes en un dominio biológico:

- *Conservación de la semántica del dominio:*

  La representación del conocimiento en el AAR es más expresiva, permitiendo incluir directamente información estructurada (por ejemplo, redes o grafos de interacción entre proteínas), mediante relaciones entre entidades. Además, se puede preservar la representación de la información no relacional, con simples atributos numéricos o nominales.

- *Decremento en el número de valores desconocidos e inexistentes:*

  El AAR elimina la gran cantidad de valores nulos o celdas vacías que se generan en la tabla única del AAP, correspondientes a valores desconocidos o inexistentes de alguna propiedad (hecho muy frecuente en los dominios biológicos). En AAR los atributos con valores desconocidos no generan una tupla en la sub-tabla dividida correspondiente. Por ejemplo, si una proteína no posee un dominio transmembrana (un tipo de anotación particular), en AAP se tendría un atributo reservado para dicha propiedad, que para esa proteína estaría vacío; mientras que en AAR se definiría el dominio transmembrana en una tabla independiente, asociada a las demás por el identificador de la proteína, sin incluir tupla alguna para dicha proteína sin dominio transmembrana. Cabe destacar que esta ventaja existe bajo el 'supuesto del mundo cerrado' con el que se trabaja en lógica de predicados (típica representación relacional), en la que se asume que todo lo no indicado explícitamente es falso. Aunque por la incertidumbre intrínseca de los datos biológicos no se puede asegurar dicho supuesto, en la práctica los métodos computacionales aplicados en Biología Molecular tienen que restringir su ámbito de trabajo a los datos realmente disponibles.

- *Interpretación más sencilla:*

  Debido a la representación en lógica de predicados, el AAR permite definir en el sesgo del lenguaje la estructura y tipo de predicados a incluir en el modelo de representación aprendido. Dicho sesgo se puede diseñar en función del tipo de salida deseado, para justificar la clasificación hecha utilizando términos útiles para los científicos que interpreten el modelo, como por ejemplo los biólogos.

- *Almacenamiento eficiente de atributos multi-valuados:*

  En la teoría de bases de datos relacionales, un atributo multi-valuado es aquel que puede tomar más de un valor simultáneamente para el mismo atributo, con una cantidad indeterminada a priori de valores diferentes [de Miguel Castaño et al., 1999]. En AAP no se puede representar, al menos de manera eficiente, porque se necesitaría un atributo booleano para cada uno de los posibles valores, que en la mayoría de los casos tomaría valor falso, incrementando de forma inútil el tamaño de la tabla de datos. Sin embargo, en AAR se puede separar dicho atributo en una nueva tabla o predicado independiente, compartiendo el atributo del identificador principal con la tabla original, e incluyendo tantas tuplas como sean necesarias para un mismo identificador, con un valor diferente del atributo multi-valuado. En dominios biológicos es muy frecuente la presencia de atributos multi-valuados, porque un gen o producto genético suele tener asociadas N anotaciones diferentes de un mismo tipo, por ejemplo más de un término Gene Ontology [Ashburner et al., 2000].

- *Mejora en el almacenamiento y gestión del conocimiento del dominio:*

En AAR los datos se organizan en módulos o tablas independientes, según las relaciones definidas. Esta característica hace más fácil trabajar con muchos datos, manipulando diferentes predicados lógicos (añadir, eliminar, mezclar *N* tablas), y diversas fuentes de datos (incluyendo una nueva relación o tabla para cada fuente; como rutas, minería de textos, perfiles de expresión de genes u homología, entre otros). No hay que limitarse a una tabla única y enorme, como en el AAP, con miles de atributos, siendo muchos redundantes, con la dificultad que implica tratar un número de atributos muy elevado, incluso mayor que el número de ejemplos. En AAP los datos pueden llegar a ser inmanejables y los algoritmos tener un coste computacional excesivo, más aún si se parte de un conjunto de datos muy amplio, como ocurre siempre en los dominios de biología molecular.

En los dominios biológicos hay mucha información relacional, debido a la estructura intrínseca de las moléculas y a la importancia de la similitud entre diferentes secuencias y estructuras, de la misma o diferentes especies. Dichos datos relacionales son aún más relevantes en el dominio de anotación funcional que se propone afrontar en este trabajo. Ya que las relaciones entre los diferentes genes y proteínas son fundamentales si se quiere explicar por qué llevan a cabo una función juntos. Por ejemplo, a bajo nivel, las moléculas (nodos) tienen enlaces químicos (como en la estructura de un grafo). Además, a alto nivel, hay redes de interacción compuestas de conexiones funcionales (enlaces) entre proteínas (nodos); o asociaciones de ortología entre un par de genes de diferentes especies con alta similitud entre sus secuencias de nucleótidos, lo que es una evidencia de una posible asociación funcional.

Por todas estas razones, se considera que el Aprendizaje Automático Relacional es más adecuado para el dominio de anotación funcional que el enfoque proposicional clásico.

2.2.3. Representación en Lógica de Predicados

En primer lugar, se introduce el lenguaje de representación de la mayoría de técnicas de AAR. Éstas se han desarrollado principalmente en el área de la Programación Lógica Inductiva [Muggleton, 1991]. Así, el lenguaje de representación son programas lógicos, un subconjunto de la lógica de primer orden, también llamada lógica de predicados. Prolog [Bratko, 2001] es el lenguaje de representación para las entradas y salidas. A continuación se presenta la terminología básica usada en programación lógica, siguiendo un orden de los elementos más específicos a los más generales, como hace Vens [Vens, 2007].

- Elementos básicos: *constantes* y *variables*. Siguiendo la convención del lenguaje Prolog [Bratko, 2001], las variables se nombran empezando con mayúscula.
- Un *término* es una constante, una variable o una *función* (por ejemplo: *f(X,Y,z)*, siendo *X* e *Y* variables y z una constante).
- Un *predicado* o *átomo* es un símbolo (que identifica el predicado) seguido por una tupla de términos entre paréntesis. Por ejemplo, *nombre_predicado(Var1,constante,Var2)*. De forma resumida, cuando no interesa el contenido de cada argumento, sino sólo su cantidad, se puede representar como *nombre_predicado/N*, siendo *N* el número de argumentos.
- Un *literal* es un predicado o la negación de un predicado (representado como ¬).
- Una *cláusula* es una disyunción de literales, por ejemplo $h_1 \wedge h_2 \wedge ... \wedge h_j \ \neg c_1 \wedge \neg c_2 \wedge ... \wedge \neg c_k$. Una cláusula se suele escribir como una implicación de la forma

$h_1 \wedge h_2 \wedge ... \wedge h_j \leftarrow c_1 \wedge c_2 \wedge ... \wedge c_k$, donde $h_1 \wedge h_2 \wedge ... \wedge h_j$ (cabeza de la cláusula) es la parte de hipótesis o conclusión, y $c_1 \wedge c_2 \wedge ... \wedge c_k$ (cuerpo de la cláusula) es la parte de condición. Generalmente, $\vee$ y $\wedge$ se sustituyen por comas. En una cláusula, todas las variables de los literales están cuantificadas universalmente.

- Una conjunción de cláusulas se denomina *teoría clausal*.
- Según el número de literales en la cabeza y el cuerpo de una cláusula, se distinguen formas especiales de cláusula:
  - Una *cláusula de Horn* es aquella que tiene como máximo 1 literal en la cabeza.
  - Una *cláusula determinada* tiene exactamente 1 literal en la cabeza.
  - Un *patrón* o *query* es una cláusula con ningún literal en la cabeza.
  - Un *hecho* es una cláusula determinada sin literales en el cuerpo. Normalmente la flecha de implicación se omite para los hechos.
- Una *cláusula programa* es una cláusula de la forma $h \leftarrow l_1, l_2, ..., l_m$, donde *h* es un predicado y $l_1, l_2, ..., l_m$ son literales.
- La *definición de predicado* es un conjunto de cláusulas programa con el mismo nombre de predicado y número de argumentos en todas sus cabezas.
- Finalmente, un *programa lógico* es un conjunto de definiciones de predicados.

Un predicado de programación lógica se corresponde con una relación en una base de datos relacional, los argumentos de un predicado se corresponden con los atributos de una relación, y las vistas de una base de datos se pueden representar en forma de clausulas. Así, un programa lógico que representa la información de una base de datos relacional se compone de dos partes. Por un lado, la parte que se define por extensión, formada por una enumeración de hechos que representan los datos y relaciones de todas las tuplas de la base de datos. Por otro, la parte que sigue una definición intensiva, compuesta por las cláusulas que representan las vistas de la base de datos. Con ILP también se pueden añadir cláusulas adicionales con conocimiento experto del dominio que facilite y mejore el aprendizaje (por ejemplo, representar la estructura interna o relaciones entre proteínas y genes). La combinación de la parte extensiva e intensiva se denomina conocimiento base.

La programación lógica es deductiva, porque sólo puede utilizar el conocimiento del dominio para extraer predicados que siempre son ciertos. Sin embargo, la programación lógica *inductiva* (ILP) utiliza inferencia inductiva, porque partiendo de un conjunto de ejemplos (en forma de predicados lógicos), se añade conocimiento del dominio (parte típicamente deductiva), siendo así capaz de encontrar regularidades o hipótesis (que pueden ser ciertas o no) a ser aplicadas sobre nuevos ejemplos.

Es muy importante destacar que la mayoría de sistemas de aprendizaje automático basados en ILP (como los usados en este trabajo) aprenden un programa lógico. Es decir, un conjunto de cláusulas programa, o una relación de salida en función de otras relaciones dadas como conocimiento del dominio.

A partir de este apartado el resto se refieren a herramientas de AAR, que prácticamente en su totalidad usan lógica de predicados.

2.2.4. Transformación a Proposicional

La transformación a proposicional se define como un cambio de representación, de relacional a proposicional (o atributo-valor) [Dzeroski and Lavrac, 2001]. Este proceso involucra la construcción de características a partir del conocimiento del dominio y de las propiedades estructurales reflejadas en una representación relacional. Esta aproximación permite aplicar algoritmos de AAP aunque los datos sean estructurados y tengan relaciones.

Siempre que no existan relaciones entre los datos, se puede aplicar directamente AAP. Cuando aparecen relaciones, se puede elegir entre aplicar directamente AAR, o transformar a proposicional la representación para poder aplicar AAP. El aprendizaje con transformación a representación proposicional es una opción práctica que permite aprovechar los progresos en AAP, que está más desarrollado y sus avances también son mayores frente al aprendizaje relacional puro, menos investigado. La desventaja es que no siempre se puede aplicar, o al menos sin perder semántica en los datos, siempre y cuando la estructura sea compleja, y no sea simplificable a una representación basada en el individuo (o ejemplo). Lo cual sucede cuando hay recursividad o estructuras complejas en los individuos, y sólo se puede aplicar aprendizaje relacional puro. Ante una estructura relacional que no requiera estrictamente una representación en lógica de primer orden (sin términos estructurados ni recursividad), se podría aplicar AAR o una combinación de relacional y proposicional, y seleccionar la que proporcione mejores resultados, dado que una vez que se tiene definida la estructura relacional, cualquiera de los dos enfoques es fácilmente aplicable.

A la hora de transformar los datos a representación proposicional existen múltiples opciones [Dzeroski and Lavrac, 2001]. Generalmente se utiliza aprendizaje relacional para llevar a cabo la transformación a proposicional. Dado que la mayoría de sistemas de AAR están basados en programación lógica inductiva (ILP), como se ha comentado previamente, los atributos que se generan son conjunciones de literales, que se convierten en atributos booleanos en la representación proposicional. Inevitablemente se pierde semántica al transformar a proposicional una representación relacional, porque no se pueden generar todas las características derivadas de la aplicación de todas las instancias de relaciones existentes en el conjunto de datos. Las variantes de los métodos para transformar a proposicional dependen del sesgo del lenguaje, que determina qué subconjunto de atributos se derivan de todo el conocimiento relacional existente, dado que la derivación de todos los atributos posibles es inviable, debido al crecimiento exponencial con el tamaño del conjunto de datos y del número de relaciones entre elementos. El sesgo es necesario para reducir el número de hipótesis candidatas. Se divide en sesgo del lenguaje (que determina el espacio de hipótesis) y sesgo de la búsqueda (que restringe el espacio de búsqueda de todas las posibles hipótesis). La búsqueda puede extraer todas las combinaciones restringidas a un subconjunto de variables [Lavrac et al., 1991]; realizar una selección probabilística de las combinaciones que más discriminan [Kramer et al., 1997]; o seleccionar combinaciones por frecuencia [Dehaspe and Raedt, 1997], principalmente. También existen métodos para transformar a representación proposicional de propósito específico [Dzeroski and Lavrac, 2001]. Otra opción para transformar parte de la información relacional a proposicional sería calcular agregados que codifiquen de forma implícita el conocimiento de las relaciones.

Además, se debe tener en cuenta que no siempre es mejor obtener tantos atributos proposicionales como sea posible a partir de la representación relacional. No sólo porque decrementa la eficiencia del método de aprendizaje proposicional debido al tamaño, sino porque puede incluir características irrelevantes. Por lo tanto, también se pueden desarrollar múltiples enfoques para tratar este aspecto dentro de la construcción de características mediante

transformación a representación proposicional.

Hay que diferenciar este enfoque de extracción de características por combinación de propiedades sencillas relacionadas (transformación a proposicional), frente a una selección de características entre cientos o miles de atributos definidos originalmente en el conjunto de datos.

2.2.5. Herramientas de Aprendizaje Automático Relacional

En esta sección se introducen brevemente diferentes herramientas de AAR. La mayoría de ellas son sistemas de aprendizaje basados en lógica de primer orden.

FOIL, Progol y Aleph

FOIL [Quinlan and Mostow, 1990] es el primer sistema de aprendizaje de reglas con lógica de primer orden. Dichas reglas deben describir un conjunto de ejemplos positivos, de acuerdo al conocimiento del dominio, y no describir ningún ejemplo negativo. En el algoritmo, se van añadiendo progresivamente literales seleccionados heurísticamente, hasta que se cubren todos los ejemplos. Otra herramienta de AAR que usa lógica inductiva es Progol [Muggleton, 1995]. Tanto FOIL como Progol aprenden reglas, no árboles de decisión explícitamente. Aleph [Srinivasan, 2007] es un sistema ILP posterior, de aprendizaje de conceptos relacionados. Desde su versión inicial, ha evolucionado de forma que actualmente, dependiendo de la configuración elegida por el usuario, puede realizar la misma funcionalidad que otros sistemas lógicos como por ejemplo Progol, FOIL, TILDE o WARMR.

Árboles de Decisión Relacionales: TILDE

La inducción de árboles de decisión (del inglés, *Top-Down Induction Decision Trees, TDIDT*) [Quinlan, 1986] es una de las técnicas más conocidas en Aprendizaje Automático. Utiliza la estrategia divide y vencerás, que es la más popular en el aprendizaje atributo-valor. Debido a las diferencias en el formato de representación entre las cláusulas de la lógica inductiva y la estructura de los árboles de decisión, la estrategia *divide y vencerás* no se usa frecuentemente en Aprendizaje en Lógica de Primer Orden. En su lugar, se emplea la estrategia de *cobertura*, generalmente para inducir reglas (como hacen otros sistemas precursores como FOIL [Quinlan and Mostow, 1990] y Progol [Muggleton, 1995]) en vez de árboles. Al generalizar la estrategia de inducción de árboles de decisión proposicionales utilizando un enfoque de ILP como es el aprendizaje por interpretación [Raedt, 1997], se obtienen árboles de decisión relacionales, con un algoritmo (TILDE [Blockeel and Raedt, 1998]) que extiende el clásico C4.5 [Quinlan, 1993] (con sus heurísticas de selección de atributos y sus criterios de poda), como muestra la figura 2.1.

TILDE [Blockeel and Raedt, 1998; Blockeel, 1998] es un sistema de Aprendizaje Automático Relacional cuya salida es un árbol de decisión relacional o árbol de decisión en lógica de primer orden. Está incluido en la herramienta ACE-ilProlog [Blockeel et al., 2000, 2006a]. Un árbol de decisión relacional [Blockeel and Raedt, 1998] es como un árbol de decisión proposicional pero con conjunciones de literales (cláusulas en lógica de primer orden) en los nodos, en lugar de comparaciones de valores de atributos. En los árboles de decisión proposicionales se comprueba el valor de los atributos (con operadores de igualdad o menor o mayor). Mientras que en los árboles de decisión con lógica de primer orden, se comprueba la existencia o no (verdadero o falso) de una conjunción de literales.

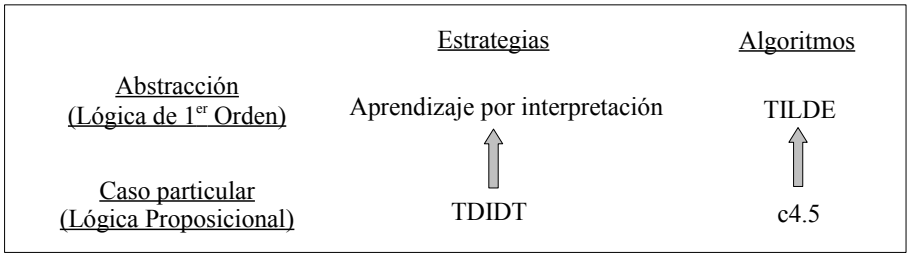


Figura 2.1: Esquema de la relación entre árboles de decisión en lógica proposicional y en lógica de primer orden.

Al construir un árbol relacional existe una diferencia en el proceso de refinamiento (elección de cláusulas de la condición de un nodo): usar el operador de subsunción-θ (del inglés, θ-subsumption) [Plotkin, 1970]. Se trata de hacer una abstracción, eligiendo predicados que incluyen a los existentes, que son un supra-conjunto. Así, se van clasificando ejemplos por capas, unas más generales sobre otras más específicas. Se comprueba qué literal divide mejor los ejemplos (más homogéneamente), calculando la mínima entropía (C4.5) o el ratio de ganancia (TILDE), que son medidas similares. Ese literal se añade a la cláusula del nodo que se está refinando.

El resto de características son iguales que en un árbol de decisión proposicional binario:

- Utiliza la estrategia divide y vencerás.
- Se selecciona la rama sucesora izquierda del árbol si se cumple la evaluación del nodo o la derecha en caso negativo.
- Una variable que aparece en la rama positiva puede volver a evaluarse posteriormente en la misma rama, pero no en la negativa, que asume que no existe.
- Para evaluar un nuevo ejemplo se recorren todas las ramas que cumplan la condición de los nodos intermedios hasta llegar a una hoja, que le asigne la clase mayoritaria; o la media, si es un árbol de regresión.
- Un árbol de decisión relacional puede transformarse en una lista de reglas de decisión (y también en un programa Prolog).

TILDE permite construir un árbol de decisión cuyos nodos contengan tanto predicados lógicos como datos no relacionales (comparaciones con atributos numéricos).

A continuación, se enuncian otros posibles usos del algoritmo TILDE y versiones relacionales similares:

- **Predicción múltiple** [Blockeel et al., 1999]: para predecir varios atributos de salida a la vez.
- **Uso de funciones agregadas** [Blockeel and Dzeroski, 1999]: permite añadir a los nodos comparaciones con el valor de salida del agregado (media, moda, mínimo, máximo, etc.). Pueden ser útiles para dar relevancia a propiedades relativas a los N elementos de un grupo.

- **Ranking** [Todorovski et al., 2002]: si se usa un conjunto de clasificadores, permite determinar cuál de ellos es mejor para cada rama del árbol, así como el orden entre los demás, para cada clase asignada en una hoja del árbol.
- **Multi-clasificación jerárquica** [Struyf et al., 2005]: para predecir un conjunto de clases (organizadas en una jerarquía) para un ejemplo dado.

ACE (del inglés, *A Combined Engine*) [Blockeel et al., 2000, 2006a] es un sistema de minería de datos con una interfaz común para varios algoritmos de Aprendizaje Automático Relacional, incluyendo TILDE [Blockeel and Raedt, 1998], WARMR [Dehaspe and Raedt, 1997], ICL [Raedt and Laer, 1995] y RRL [Dzeroski et al., 2001].

Patrones Frecuentes y Reglas de Asociación: PolyFARM y WARMR

WARMR [Dehaspe and Raedt, 1997] es un sistema para extraer reglas de asociación en lógica de primer orden. También está incluido en la herramienta ACE. El objetivo del algoritmo de WARMR es encontrar todos los patrones de ocurrencia frecuente siguiendo unas restricciones dadas. Es un método basado en niveles, similar al algoritmo APRIORI [Agrawal et al., 1996]. El algoritmo realiza una búsqueda en anchura en el espacio de patrones, ordenada por la generalidad de los patrones. La poda está basada en la relación entre la especificidad y la frecuencia: si un patrón no es frecuente, entonces ninguna de sus especializaciones pueden serlo. Así, este método de aprendizaje es rápido y eficiente en bases de datos grandes.

Si se compara WARMR con TILDE, este último conserva la eficiencia de la técnica de construcción de los árboles de decisión, que trabajan con un fragmento de datos cada vez más pequeño [Clare, 2003]. Por otro lado, TILDE tiene un sesgo al ir construyendo particiones, de forma que, si una decisión es perjudicial, ésta se propaga al resto de ramas inferiores del árbol. Por su parte, las reglas de asociación no tienen este sesgo, y se pueden aplicar a bases de datos más grandes, con una menor complejidad computacional.

En el dominio de anotación funcional en el que se centra este trabajo, las reglas de asociación se podrían usar como un paso de pre-procesamiento ante una base de datos inmanejable para un árbol de decisión. Así, primero, se extraen las relaciones más importantes, y segundo, éstas se usan como atributos booleanos de un árbol de decisión, como se ha hecho en otras aplicaciones bioinformáticas [Clare et al., 2006]. La salida de WARMR también se podría usar para obtener descripciones que caractericen un grupo de genes o proteínas dado, en los términos deseados, definiendo los predicados lógicos de entrada en función de ellos.

PolyFARM [Clare and King, 2003] es un sistema de extracción de reglas de asociación, semejante a WARMR, que además divide el procesamiento entre varias máquinas de ejecución paralela. Está implementado en Haskell. Su objetivo original fue acelerar la ejecución que proporcionaba WARMR, en el que está basado. WARMR inicialmente era demasiado lento como para poder ser aplicado a un gran conjunto de datos, como los procedentes de la biología molecular. Posteriormente, mejoró notablemente, al implementarse la nueva versión de la herramienta ACE (en la que están incluidos WARMR y TILDE), que usa desde entonces el motor de búsqueda *ilProlog*, más eficiente.

Aprendizaje Híbrido: Relacional y Proposicional

Se puede seleccionar el método de aprendizaje a utilizar en función de los datos: proposicional, relacional o híbrido. En esta tesis, la denominación Aprendizaje Híbrido se refiere a la combinación de aprendizaje relacional y proposicional. En el aprendizaje híbrido,

el aprendizaje relacional se utiliza para transformar los datos relacionales a representación proposicional (ver sección 2.2.4).

Entre las diversas opciones de aprendizaje híbrido, existe una muy aplicada en dominios biológicos (ver sección 2.3.1) semejantes al que se afronta en esta tesis. Se trata de transformar a representación proposicional mediante la extracción de patrones frecuentes con un algoritmo de generación de reglas de asociación (PolyFARM [Clare and King, 2003] o WARMR [Dehaspe and Raedt, 1997]), de forma que cada uno de los patrones sea un atributo booleano proposicional, que toma valor 1 si la conjunción de literales se cumple en el ejemplo independiente, o 0 en caso contrario. Posteriormente, se aplica un árbol de decisión proposicional (C4.5 [Quinlan, 1993] o CLUS [Blockeel et al., 1998]).

Este aprendizaje híbrido que combina extracción de patrones frecuentes con un árbol de decisión proposicional se utiliza en el método de predicción DMP (del inglés, *Data Mining Prediction*) [King et al., 2000b], que combina los algoritmos PolyFARM y C4.5. DMP se ha aplicado ampliamente en la predicción funcional de diferentes especies (como *E.coli*, *M.Tuberculosis*, *Arabidopsis Thaliana*) desde hace algo más de una década [King et al., 2000a, 2001; Clare et al., 2006], incluso verificando sus predicciones experimentalmente [King et al., 2004b].

Posteriormente, a partir de DMP (ver evolución en sección 2.3.1) se ha aplicado una nueva combinación de algoritmos más actuales y flexibles: WARMR y CLUS, también aplicada con éxito en otros dominios de predicción funcional genómica, incluso multi-etiqueta y jerárquica [Blockeel et al., 2006b; Vens et al., 2008]. Estos enfoques son especialmente útiles en Biología Molecular, porque los genes o proteínas están involucrados en varias funciones (multi-etiqueta), y las funciones en biología se organizan con frecuencia en niveles (jerárquico).

Por lo tanto, se decide utilizar esta última combinación de aprendizaje híbrido en el capítulo 7 de esta tesis.

Otros sistemas de Aprendizaje Relacional

- **Aprendizaje Relacional con Probabilidad**

  Cuando hay incertidumbre, ruido y/o valores desconocidos en los datos, puede resultar adecuado utilizar un enfoque estocástico. Relacionado con esta idea, ha surgido una nueva área de investigación denominada Aprendizaje Relacional Estadístico (del inglés, *Statistical Relational Learning*) [Raedt and Kersting, 2003]. Esta área se define como una intersección entre el razonamiento probabilístico, la lógica de primer orden y el Aprendizaje Automático. La probabilidad frecuentemente viene representada en forma de redes bayesianas, modelos ocultos de Markov o gramáticas estocásticas. Existen múltiples sinónimos o enfoques estrechamente relacionados, como por ejemplo: Modelos Probabilísticos Relacionales (del inglés, *Relational Probability Model*), Programación Lógica Inductiva Probabilística (del inglés, *Probabilistic ILP*), Aprendizaje Lógico Probabilístico (del inglés, *Probabilistic Logic Learning*), Programación Lógica Estocástica (del inglés, *Stochastic Logic Programming*), etc.

  A continuación se mencionan algunos ejemplos de herramientas de Aprendizaje Relacional Estadístico o Probabilístico. El algoritmo FAM (del inglés, *Failure Adjusted Maximization*) [Cussens, 2001; Chen et al., 2008] asigna probabilidades a posteriori a un árbol de decisión completamente construido. Es una versión de un algoritmo de estimación clásico de Esperanza-Maximización (EM) [Dempster et al., 1977]. Otro ejemplo es Profile (del inglés, *Probabilistic First-order LEarning*), un conjunto de

diferentes herramientas desarrolladas en Java, para Aprendizaje Relacional Estadístico y Programación Lógica Inductiva Probabilística. Incluye programas con versiones relacionales probabilísticas para árboles de decisión (Tilde-CRF [Gutmann and Kersting, 2006]), redes bayesianas (nFOIL [Landwehr et al., 2005] y Balios [Kersting and Dick, 2004]) y modelos ocultos de Markov (Xanthos [Kersting et al., 2006]). Un caso más es el paquete de software Alchemy [Domingos et al., 2006], basado en una representación lógica de Markov. También cabe mencionar otros sistemas como ProbLog [Raedt et al., 2007], una extensión probabilística del lenguaje Prolog, o PRISM [Sato and Kameya, 2001] (del inglés, *PRogramming In Statistical Modeling*), un lenguaje de programación general para el modelado simbólico-estadístico.

- **Enfoque Relacional no Basado en Lógica de Predicados: Weka Relacional**

  Como extensión de la herramienta *Weka* [Witten and Frank, 2005] de Aprendizaje Proposicional, surge *Weka Relacional* [Woznica, 2006] para Aprendizaje Automático Relacional. Lo más relevante es su limitación a muy pocos algoritmos, muchos menos que *Weka* proposicional, así como su representación relacional no basada en la lógica de predicados. Simplemente permite que los datos de entrada estén distribuidos en varias tablas relacionadas entre sí, en lugar de una sola.

  El principal enfoque de aprendizaje que proporciona *Weka Relacional* es el basado en instancias (extensión del algoritmo *KNN* [Mitchell, 1997]). Cabe mencionar que también presenta una extensión relacional de las máquinas de vector de soporte (*SVM* [Vapnik, 1998]). Aunque la herramienta implementa diferentes medidas de distancia, el aprendizaje basado en instancias requiere definir una métrica de distancia entre cada par de ejemplos. Entre atributos numéricos es sencillo establecer una distancia. Pero entre atributos nominales, como son la mayoría de las anotaciones que aparecen en los dominios biológicos como los afrontados en este trabajo, no es nada trivial establecer una distancia. Además, existe la complejidad añadida de definir distancias entre atributos de diferentes tablas.

  Otra desventaja de *Weka Relacional* frente a un sistema de AAR con una representación en lógica de predicados es la carga computacional (en memoria y tiempo). Primero, porque *Weka Relacional* está implementado en lenguaje Java, que consume muchos recursos de memoria, cuya ocupación aumenta notablemente en aprendizaje relacional, por las múltiples relaciones entre diversas tablas. Segundo, porque en el enfoque basado en instancias hay que calcular las distancias *todos contra todos* los elementos (relacionados o no), restringiendo la cantidad de instancias que el sistema puede manejar. Ambas limitaciones no son oportunas en dominios biológicos que trabajan con bases de datos muy grandes, con muchos elementos.

- **Aprendizaje Multi-Instancia**

  Por otro lado, en el límite entre el Aprendizaje Proposicional y el Aprendizaje Relacional, se encuentra el Aprendizaje Multi-Instancia [Dietterich et al., 1997]. Éste se originó al afrontar un caso práctico de descubrimiento de fármacos.

  La idea principal en este tipo de aprendizaje es que cada ejemplo es un conjunto de tuplas, de número indeterminado y que puede ser diferente para cada ejemplo. Es decir, la clase se asocia a una bolsa de instancias, en vez de tener asignada una clase a cada instancia particular. Dentro del conjunto de instancias, unas pertenecen realmente a la

clase y otras no. De forma que si una de las instancias cumple la condición o condiciones de clasificación, se le asigna la clase positiva a todo el conjunto.

Para este tipo de aprendizaje se puede usar una representación atributo-valor o relacional. El enfoque del Aprendizaje Multi-Instancia se podría aplicar a la predicción de función de listas de genes, aunque incluye algunos sesgos [Blockeel et al., 2005] que no lo hacen estrictamente válido para aprender árboles de decisión como TILDE .

2.3. Aplicación del Aprendizaje Automático a Biología Molecular

2.3.1. Biología Molecular como Campo del AA

Además de ser un interesante área de aplicación, la Biología Molecular es un campo abierto para el Aprendizaje Automático [Larrañaga et al., 2006; Tarca et al., 2007; Inza et al., 2010].

La figura 2.2 muestra una perspectiva de todas las tareas biológicas en las que se aplica el AA, poniendo de manifiesto que la aplicación de AA a Biología es un área bien establecida, desde hace tiempo y que continua vigente.

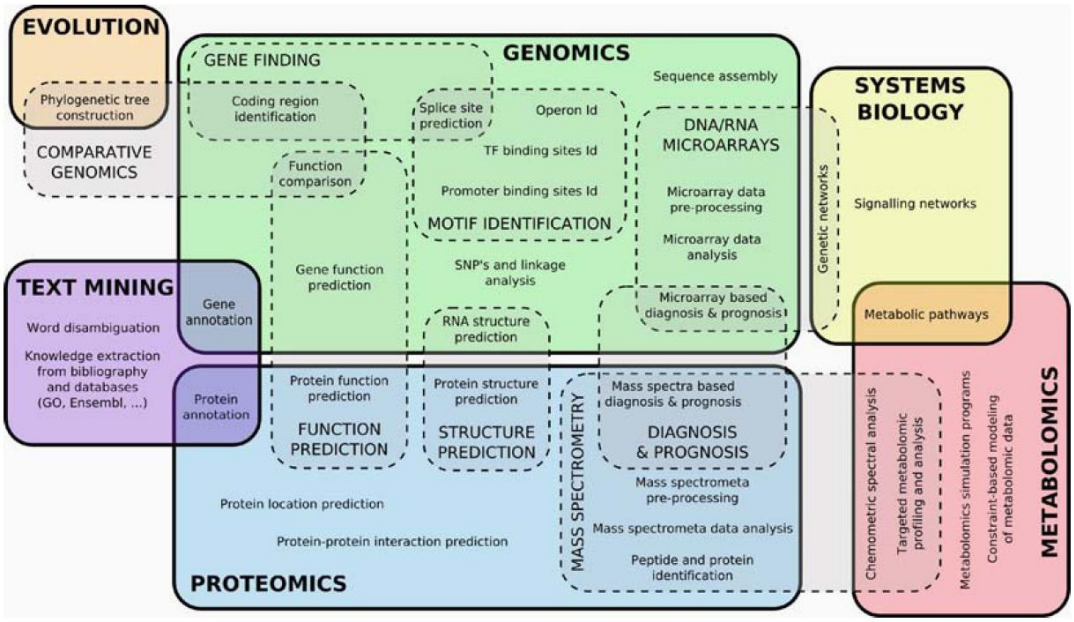


Figura 2.2: Esquema general de las aplicaciones de Aprendizaje Automático en Biología Molecular. Fuente: [Inza et al., 2010].

Según un estudio reciente [Jensen and Bateman, 2011] los métodos de Aprendizaje Automático supervisado más usados en Biología son las redes de neuronas, las máquinas de vector de soporte, los modelos de Markov y los árboles de decisión, con la aparición creciente de los bosques de árboles aleatorios (del inglés, *random forests*).

Para un repaso bibliográfico exhaustivo, se pueden consultar las revisiones de aplicación a Biología de Aprendizaje Automático (y otros sistemas de inteligencia computacional) [Larrañaga et al., 2006; Fogel, 2008]. A continuación se presentan algunos ejemplos de uso de diferentes técnicas del AA para distintos dominios de la Biología Molecular:

- La definición de sitios de iniciación de la traducción en los genes de *E.coli* [Stormo et al., 1982] fue uno de los primeros usos de las redes de neuronas en bioinformática.

- La localización de genes en la secuencia de ADN (es decir, regiones que codifican proteínas) es una de las aplicaciones más importantes del AA en este área [Mathe et al., 2002], usando árboles de decisión o clasificadores bayesianos, entre otros; incluso combinando diferentes fuentes de información [Allen et al., 2004], como también se hace durante esta tesis (ver capítulo 6).

- Para el reconocimiento de regiones promotoras de la transcripción se han utilizado redes de neuronas, modelos de Markov, computación evolutiva y métodos del vecino más cercano [Fogel, 2008].

- La detección de sitios de ensamblaje alternativo se ha realizado con modelos ocultos de Markov [Cawley and Pachter, 2003].

- Para la predicción de genes involucrados en enfermedades genéticas se han utilizado árboles de decisión [López-Bigas and Ouzounis, 2004].

- En la predicción de los efectos fenotípicos de polimorfismos de nucleótidos aislados (del inglés, *single nucleotide polymorphisms, SNP*) no sinónimos se han comparado los resultados de máquinas de vector de soporte y bosques de árboles aleatorios, usando información estructural y evolutiva [Bao and Cui, 2005].

- A partir de datos de análisis de ADN, se usa el algoritmo C4.5 para extraer reglas que proporcionen conocimiento comprensible por el humano [Sebban et al., 2002], con el mismo objetivo que se utilizan las reglas de decisión en esta tesis (ver capítulo 7).

- Para la predicción de la estructura secundaria de proteínas se ha empleado el método del vecino más cercano [Yi and Lander, 1993].

- Para la predicción de estructura terciaria de las proteínas y clasificación en familias de proteínas se han utilizado redes bayesianas [Raval et al., 2002].

- Para el reconocimiento de patrones partiendo de datos de microarrays se encuentran múltiples técnicas de aprendizaje supervisado [Valafar, 2002; Xu et al., 2007a] y de agrupamiento [Sheng et al., 2005], consiguiendo principalmente subconjuntos de genes (perfiles de expresión génica) implicados en la diagnosis de cáncer.

- Para predecir si la respuesta reguladora de un gen es alta o baja se utiliza un conjunto de árboles de decisión [Middendorf et al., 2004].

- *PSORT II* [Horton and Nakai, 1997] realiza localización sub-celular usando árboles de decisión inicialmente, y k-vecinos más cercanos en una versión posterior.

- *ProtFun* [Jensen et al., 2002b] realiza anotación funcional con combinaciones de redes de neuronas.

- Se realiza modelado de redes de regulación génica, por ejemplo, con redes de neuronas y optimización por enjambre de partículas [Xu et al., 2007b].

- Para la reconstrucción de árboles filogenéticos se suelen aplicar otros sistemas inteligentes, como es la computación evolutiva, por ejemplo los algoritmos genéticos [Matsuda, 1995] o las colonias de hormigas [Catanzaro et al., 2007].
- También hay que mencionar la presencia de sistemas de inteligencia computacional en software comercial [Fogel, 2008], especialmente para el descubrimiento de fármacos [Thomsen, 2007].

Finalmente, cabe destacar la importancia de la anotación funcional (caracterizar funcionalmente genes y proteínas), tanto en genómica como en proteómica, ya que es el objetivo más importante de la Biología Molecular. La anotación funcional es el área en la que se centra esta tesis. Para más detalles sobre anotación funcional, consultar la sección 2.4.1 y el anexo B.

Tras este breve repaso se puede concluir que el AA es una técnica de pasada y continua aplicación para resolver problemas en múltiples dominios de la Biología Molecular, justificando así la metodología computacional elegida en esta tesis para afrontar varios problemas biológicos.

Aplicaciones del AAR

Desde sus inicios, el AAR ha sido aplicado ampliamente y con éxito a dominios biológicos y químicos, siendo trabajos de referencia los de King, Muggleton, Page y Dzeroski, entre otros, como se detalla a continuación. La primera aplicación fue la resolución del problema de la mutagénesis en 1996 [Srinivasan et al., 1996]. También fue de los primeros problemas que se resuelven con AAR, porque el AAP (en concreto, regresión lineal múltiple) no puede solucionarlo. Otro ejemplo inicial son los errores cometidos por C4.5 que se corrigen con AAR en el dominio de la toxicología, en 1999 [Srinivasan et al., 1999]. En el trabajo de Page y Craven [Page and Craven, 2003] se puede consultar un análisis estructurado y detallado de los tipos de problemas que se pueden afrontar con el Aprendizaje Automático Multi-Relacional en el dominio bioinformático, junto con los datos de distinta naturaleza que se pueden incluir. Incluso, el reto de la conferencia de Programación Lógica Inductiva del año 2005 (ILP-2005) [Kramer and Pfahringer, 2005] estuvo centrado en un dominio biológico, como es la clasificación funcional de los genes de la bacteria *Saccharomyces cerevisiae*. A continuación, se presentan algunos ejemplos concretos especialmente relacionados con el área que se trata en este trabajo: la anotación funcional de genes y proteínas.

En la tesis doctoral de Clare [Clare, 2003] se aplica el AAR con éxito a la predicción de función de genes individuales de *Saccharomyces cerevisiae (yeast)*, generando diversos trabajos derivados con los mismos conjuntos de datos [Struyf et al., 2005; Blockeel et al., 2006b], o aplicando la misma técnica a otras especies [Clare et al., 2006]. El trabajo de Clare se centra en aplicar el método DMP (del inglés, *Data Mining Prediction*), un sistema de aprendizaje híbrido, que combina un algoritmo de inducción de reglas de asociación relacionales (PolyFARM), y posteriormente un árbol de decisión proposicional, C4.5.

Posteriormente se incluye una ampliación para predicción multi-etiqueta aplicada también a genómica funcional, de forma que el árbol predice un vector de clases booleanas, en lugar de una sola clase. Este enfoque evoluciona para predecir clases en varios niveles de una jerarquía (clasificación multi-clase jerárquica) [Clare, 2003]. Aquí se asigna un coste mayor a errores de clasificación en niveles superiores de la jerarquía frente al coste en niveles inferiores, donde es más fácil equivocarse y, por tanto, se debe penalizar menos. Para ello se modifica tanto la herramienta que obtiene reglas de asociación (PolyFARM), como la que construye el árbol de decisión proposicional (C4.5). Paralelamente se desarrolla otro sistema de clasificación

jerárquica y multi-etiqueta basado en el sistema CLUS de *clustering* de predicción, que se aplica sobre el mismo conjunto de datos para predecir funciones individuales de *yeast* [Blockeel et al., 2006b; Vens et al., 2008].

Las técnicas de AAR también se han aplicado a otros dominios biológicos similares. Algunos ejemplos recientes son la predicción de interacción proteína-proteína [Tran et al., 2005], la extracción de grupos de genes a partir de datos de microarrays [Trajkovski et al., 2008], o la clasificación de lugares de enlace en hexosas [Nassif et al., 2009].

En áreas afines, últimamente se aplica el Aprendizaje Relacional o/y la Programación Lógica Inductiva a dominios dentro de la Biología de Sistemas. Valencia ha predicho la biodegradación de nuevos componentes químicos por la acción de microorganismos, utilizando el conocimiento relacional implícito de la estructura química a nivel atómico, transformado a una representación proposicional [Gómez et al., 2007]. Dzeroski ha aplicado TILDE al modelado ecológico, para simular y evaluar modificaciones genéticas en algunos cultivos agrícolas [Ivanovska et al., 2008]. Por su parte, Muggleton se centra en aplicaciones industriales. Por ejemplo, la mejora de las variedades de plantas de cultivo, a través de la predicción de la maduración del tomate y su calidad, o la identificación de componentes metabólicos clave en un tumor de hígado [Muggleton et al., 2010].

Por su parte, hay que destacar el proyecto multidisciplinar *Robot Scientist* de King [King et al., 2004a, 2009], de gran relevancia internacional. Auna las ciencias de la computación y la microbiología para extraer automáticamente conocimiento de experimentos *in-vivo*, a lo largo de varias iteraciones de procesamiento en laboratorio y en computador. Se trata de la primera implementación física de un laboratorio microbiológico controlado automáticamente por hipótesis derivadas de técnicas de Aprendizaje Automático.

2.3.2. Retos del AA en Bioinformática

Existe una larga lista de retos para el AA asociados al dominio y contexto biológico que hacen, aunque difícil, más interesante el desarrollo del trabajo de esta tesis. Las principales son:

- Manejo de un gran número de datos, que obligan a automatizar absolutamente todos los procesos.
- Enormes cantidades de información procedentes de una sola fuente de datos. Se requieren muchos recursos computacionales, impidiendo incluso la aplicación de ciertos métodos estándar.
- Ruido intrínseco en los datos [Baldi and Brunak, 2001]. Las razones de dicho ruido, aparte de un error inicial experimental, pueden ser: un proceso erróneo de carga en las bases de datos públicas, una interpretación incorrecta de los experimentos, una revisión de diferentes elementos por distintos supervisores, o la unión de información cargada en las bases de datos por diversas personas. Este ruido repercute en la inexactitud y falta de completitud de los datos utilizados para el entrenamiento del Aprendizaje Automático. Tanto en la asignación de clases, que siempre incluye falsos positivos y falsos negativos, como en los atributos, al utilizar los valores conocidos hasta el momento, que en un futuro pueden modificarse (por ejemplo, corrigiendo errores o añadiendo anotaciones desconocidas). De forma que en un esquema ideal de clasificación, las predicciones positivas estarían repartidas entre los verdaderos y los falsos positivos; mientras que en problemas de clasificación en Biología, al estar basados

con frecuencia en conocimiento incompleto, las predicciones positivas se dividen con un porcentaje mínimo para verdaderos positivos, otro similar para falsos positivos, y la mayor parte para desconocidos o *'falsos falsos positivos'* [Yu et al., 2008]. Así, hay que tener presente que dicho ruido y falta de completitud afecta al rendimiento de los resultados obtenidos en un proceso de AA, al depender en gran medida de la calidad de los datos de entrenamiento [Jansen and Gerstein, 2004].

- Múltiples identificadores biológicos diferentes para un mismo elemento, lo cual exige un proceso de mapeo continuo [Huang et al., 2007], así como una posible fuente adicional de ruido.
- Interrelación compleja de todos los elementos biológicos. Se conoce que la mayoría de las anotaciones funcionales (utilizadas frecuentemente como datos de entrada de la predicción) proceden de transferencia por homología [Rost et al., 2003] (ver una descripción de la homología en la sección B.3.1). Por lo tanto, se deben analizar siempre cuidadosamente dichas interrelaciones, para no sesgar el proceso de aprendizaje al incluir instancias muy similares para entrenar y para evaluar a la vez. Por ejemplo, muchos genes y proteínas pueden ser miembros de una misma familia por similitud en secuencia o estructura, incluso en diferentes especies.
- Redundancia en las bases de datos. Los mismos genes y proteínas pueden estar en distintos repositorios, hasta con distintos valores en sus propiedades. Incluso las secuencias pueden ser diferentes, debido a un proceso de secuenciación distinto o a la transcripción alternativa. Así, determinar el valor más adecuado es un problema añadido.
- Clasificación multi-clase. En Biología, frecuentemente no se afrontan problemas en los que una clasificación binaria sea suficiente, sino que se necesita realizar una selección entre N posibles valores o funciones [García-Pedrajas and de Haro García, 2008].
- Clasificación multi-etiqueta. Cada gen o proteína generalmente está involucrado en más de una función, requiriendo que se pueda asignar más de un valor de la clase a cada ejemplo [Tsoumakas and Katakis, 2007]. La multi-funcionalidad es un reto importante y poco afrontado hasta hoy en la anotación funcional [Juncker et al., 2009].
- Clases des-balanceadas. En Bioinformática la cantidad de ejemplos positivos (con frecuencia, procedentes de fuentes experimentales) siempre es mucho menor que la de ejemplos negativos. En el dominio de anotación funcional, que también es multi-clase, el número de instancias para cada una de las N posibles clases es, además, muy diferente entre sí. Este des-balanceo de clases en predicción de función puede venir dado porque unas funciones son más comunes (transporte y enlace) en la célula que otras (funciones relacionadas con ácidos grasos y metabolismo de los fosfolípidos); y también porque las anotaciones están sesgadas y limitadas a tipos de proteínas sobre las que se han hecho más estudios y análisis [Al-Shahib et al., 2005].
- Definición de clases imprecisa, no fiable y a veces a distintos niveles (por ejemplo: en una jerarquía), o incluso con clase desconocida para algunos ejemplos.
- Múltiples fuentes de información a integrar en un solo esquema de representación del conocimiento. Este reto también implica decidir si tomar todos los datos disponibles o un subconjunto que cumpla unos criterios, si tomar los datos pre-procesados u originales, entre otras cuestiones.

- Valores desconocidos (del inglés, *missing values*). Pueden ser debidos a la pérdida de datos por un problema particular de manejo de datos [Inza et al., 2010]. Pero los valores desconocidos más relevantes, o más complicados de gestionar, surgen del hecho de la inexistencia de anotaciones de todos los tipos de información para todos los genes o proteínas, con una carga semántica asociada de carácter biológico, que desaconseja su gestión con los métodos estándar. La representación del conocimiento debe cubrir este tipo de casos.

- Lista o grupo de elementos biológicos. Con frecuencia se utiliza como unidad el concepto de lista o grupo de genes/proteínas, en lugar de considerar un gen o producto genético independiente. Con lo que hay que asociar a cada grupo sus propiedades, intrínsecas y agregadas.

- Falta de estandarización de los métodos de predicción de anotación de función. Dificultad en cómo y con qué comparar, y qué medida de evaluación utilizar.

En conclusión, existen múltiples habilidades que las técnicas de AA deben desarrollar y evaluar para la resolución de problemas biológicos.

2.4. Anotación Funcional con Información de Redes

2.4.1. Anotación Funcional

El objeto de estudio de esta tesis es caracterizar la función de genes y proteínas basándose en información de redes, a través de la representación Relacional y el Aprendizaje Automático.

Como se ha comentado previamente, la anotación funcional es el problema fundamental a resolver en Biología Molecular, es decir, definir qué tarea/s se encarga de desarrollar cada proteína y gen en un organismo, para conocer dónde actuar en caso de mal-función o enfermedad.

En Biología, la función de un gen o proteína no es un término definido explícitamente, sino que la función es un fenómeno complejo que se asocia al gen o proteína mediante muchos niveles solapados y entrelazados. Forma parte de la anotación funcional de una proteína, por ejemplo: identificar si está involucrada en un proceso biológico, una red de regulación, con qué moléculas interacciona, cuál es su localización celular o su función molecular; o asignarle su perfil de expresión o su fenotipo (tejido o asociación a enfermedad, entre otros). Así, se puede usar la noción generalizada de que "función es todo lo que le pasa a y a través de una proteína"[Rost et al., 2003]. Por lo tanto, la forma de definir la función de un gen o proteína (es decir, anotar función) es asignar distintos términos, en distintos vocabularios, y distintos niveles de función (ver debajo apartado 'Niveles de Anotación'). El anexo B presenta más información sobre la anotación funcional en Biología.

Desde el punto de vista del Aprendizaje Automático, la predicción de anotación funcional se puede relacionar fundamentalmente con las aplicaciones habituales de clasificación o etiquetado. Aunque con ciertas exigencias y condiciones restrictivas, como algunas de las expuestas en la sección 2.3.2.

Las funciones celulares son casi siempre el resultado de la acción coordinada de varias proteínas, interaccionando en complejos, asociadas en rutas o redes de proteínas [Bader et al., 2008]. Así, conocer la red de interacción es una tarea esencial para comprender y explicar la dinámica de todos los procesos biológicos; debido al gran número de interacciones y

asociaciones funcionales existentes entre proteínas, sus condiciones variadas de aparición, y a las múltiples formas en que una proteína puede influir en la función de otras.

Por lo tanto, en la anotación funcional es muy importante tener presente la influencia de las interacciones, asociaciones funcionales y los sistemas que forman, desde el enfoque de la Biología de Sistemas, descrito en la sección 2.4.2 y siguientes.

Niveles de Anotación

La función de un gen o proteína se puede definir a distintos niveles, dependiendo de los aspectos bioquímicos, fisiológicos o fenotípicos considerados [Friedberg, 2006]. Además, la anotación puede ser individual o para un grupo de elementos.

El mismo compuesto biológico puede tener asignada una función diferente en cada nivel, incluso varias en el mismo nivel, porque su función tiene consecuencias desde el nivel subcelular hasta el nivel de organismo completo. Por ejemplo, una proteína quinasa, en un aspecto bioquímico (función molecular, individual), su función es fosforilar un grupo hidroxilo; en un aspecto fisiológico (proceso biológico, grupo), la quinasa forma parte de una ruta de señalización, donde fosforila y es fosforilada a la vez; y en un aspecto fenotípico, una mutación en la quinasa podría causar una enfermedad.

Además, una misma proteína puede tener más de una función dependiendo de diversos eventos, como modificaciones puntuales, interacciones con otras proteínas, formación de complejos y participación en una ruta biológica, entre otros.

También hay que destacar que la composición de dominios de una proteína puede hacer que realice distintas funciones, dependiendo de la parte de la proteína considerada. Las proteínas generalmente están compuestas por uno o más dominios, que son regiones funcional y estructuralmente independientes, dentro de una misma secuencia.

Por lo tanto, al hablar de función biológica, hay que especificar el nivel de función del que se trata. Así, al usar o desarrollar un método de predicción de función, unos enfoques serán más adecuados que otros, dependiendo del propósito específico (el nivel de función que se busca), eligiendo un vocabulario adecuado para ello, y teniendo en cuenta las relaciones semánticas entre los datos usados para anotar.

2.4.2. Biología Molecular, Redes y Biología de Sistemas

En su definición original, la Biología de Sistemas es un nuevo campo de la Biología cuyo objetivo es desarrollar la comprensión de los sistemas biológicos a nivel de sistema [Kitano, 2001; Ideker et al., 2001]. Actualmente, la Biología de Sistemas Molecular es una disciplina integradora que busca explicar las propiedades y comportamiento de los sistemas biológicos complejos en términos de sus componentes moleculares y sus interacciones [Aebersold, 2005; Likic et al., 2010].

Tradicionalmente los biólogos moleculares estudian individualmente los genes y proteínas, sus interacciones y su influencia en las moléculas relacionadas. Recientemente, los progresos tecnológicos de experimentación a gran escala han permitido y creado la necesidad de construir automáticamente grandes modelos de redes de interacción a partir de los datos experimentales [Bader et al., 2008]. De esta forma se hace posible el estudio del comportamiento y las propiedades del sistema biológico completo, como indica la Biología de Sistemas.

Este reciente interés por el estudio de los sistemas completos proviene de la idea subyacente de que hay propiedades de la red que no se explican por una combinación sencilla de los componentes. En otras palabras, como *"el total es más que la suma de las partes"*, se requiere

el estudio de la red como un todo. Esta nueva visión de análisis 'de arriba hacia abajo', trata de complementar todos los resultados obtenidos con la clásica visión 'de abajo a arriba', que estudia los genes y proteínas como entidades aisladas en la Biología Molecular, limitada por la complejidad de los sistemas vivos [Pazos et al., 2003].

La tarea central de la Biología de Sistemas es (a) reunir información de forma exhaustiva sobre los elementos individuales del sistema a distintos niveles e (b) integrar estos datos para generar modelos de predicción del sistema [Ideker et al., 2001].

La Biología de Sistemas no sólo incluye el diseño, la construcción de la red y el análisis de las propiedades emergentes, sino que también considera la simulación y control de la dinámica de la red a partir del modelo construido [Kitano, 2002].

2.4.3. Asociaciones Funcionales e Interacciones en Biología Molecular

Como definición general, se puede decir que un par de proteínas asociadas funcionalmente significa que están relacionadas por la función que realizan. La asociación funcional puede tener diversos grados de fortaleza, desde establecer una conexión física hasta pertenecer a puntos distantes en una misma ruta de señalización. Si dicha asociación funcional implica contacto físico, entonces se denomina interacción [2].

Para caracterizar funcionalmente el proteoma, la información que aporta el interactoma (conjunto de interacciones y asociaciones funcionales) es fundamental. El gran número de posibles asociaciones entre proteínas, la variedad de condiciones ambientales y estados celulares en los que estas asociaciones pueden reorganizarse, y las múltiples formas en que una proteína puede influir en la función de otras, requieren el desarrollo de enfoques experimentales y computacionales para analizar y predecir asociaciones funcionales entre proteínas como parte de su actividad en el interactoma. Una parte considerable de la diversidad y complejidad biológica está codificada en las interacciones y asociaciones funcionales entre moléculas de mayor nivel que los genes, como son las proteínas [Rojas et al., 2006]. Por tanto, el conocimiento de las redes de proteínas es crítico para poder comprender, explicar y regular la dinámica de casi todos los procesos biológicos de los sistemas vivos, tales como: transducción de señal (como la replicación y traducción de genes), metabolismo (como la síntesis y uso del ATP), arquitectura celular (como la construcción de estructura del cito-esqueleto) y transferencia de información.

Establecer la estructura de todas las interacciones y asociaciones funcionales entre las proteínas de una célula viva, incluyendo las variaciones temporales (según el estado celular) y espaciales (según el compartimento o zona de la célula), es un problema muy complejo y sujeto a errores experimentales. En la sección B.4 aparece una breve explicación sobre dicha obtención experimental de las interacciones.

Para construir la red específica de un sistema biológico, se parte de las interacciones y asociaciones funcionales detectadas experimentalmente o mediante métodos de predicción (para una descripción de estos procedimientos, ver las secciones B.4 y B.5).

[2] En esta tesis se considera que las asociaciones funcionales incluyen a las interacciones, como un caso particular. No obstante, muchas veces se utilizan ambas denominaciones, para insistir en la existencia de interacciones *físicas* dentro de las asociaciones funcionales. Por otro lado, desde el punto de vista de representación computacional, en esta tesis, todas las asociaciones funcionales, incluidas las interacciones, se denominan 'relaciones', donde también se incluyen las relaciones entre un gen y las proteínas que genera su expresión.

Tipos de Asociación Funcional e Interacción

Existen múltiples tipos de asociaciones funcionales entre genes, proteínas y otros productos genéticos. Las asociaciones funcionales pueden ser transitorias o estables, débiles o fuertes, entre pares o entre grupos, a nivel físico o a nivel funcional, etc. Es importante recordar que cuando existe contacto físico se denominan interacciones.

A parte de las interacciones entre proteínas, se pueden considerar interacciones entre regiones específicas de una molécula, o incluso entre residuos concretos. Las interacciones pueden ser por pares, como las interacciones proteína-proteína; o entre más de dos elementos, como en los complejos. Así, las interacciones proteína-proteínas se pueden considerar interacciones físicas directas, y las de complejos, interacciones físicas indirectas. Esto se debe a que un complejo es un grupo de unas pocas proteínas con un alto grado de conexión entre ellas, pero sin estar cada proteína del complejo en contacto directo con todas las demás del mismo. No obstante, en las bases de datos se almacenan como pares de todas contra todas las proteínas del complejo, ya que los procesos experimentales de determinación de complejos (ver sección B.4.1) generalmente no permiten diferenciar las interacciones físicas directas de las indirectas dentro del complejo.

A partir de los pares de asociaciones funcionales e interacciones proteína-proteína se pueden construir redes de proteínas. Por lo tanto, las redes incluyen las mismas proteínas que las interacciones o asociaciones funcionales por pares, pudiendo desempeñar muchas de ellas diversas funciones, dependiendo de con quién y a qué nivel se asocien.

Si se tiene un grupo de muchas proteínas asociadas funcionalmente a nivel de sistema se habla de **ruta biológica**. De manera formal, una ruta biológica o proceso biológico (del inglés, *pathway*) es una recopilación abstracta de unas decenas de proteínas y otros compuestos, implicados en la realización de una misma función a nivel de sistema, generalmente organizados en cascadas de interacciones [Cary et al., 2005; Ooi et al., 2010]. Existen rutas o procesos metabólicos, de regulación y de señalización, como por ejemplo la degradación de compuestos, la replicación del ADN, la infección por el virus de la gripe, o la regulación de la hormona del tiroides.

Finalmente, desde una perspectiva amplia a nivel funcional, cualquier tipo de anotación compartida entre proteínas puede representar una relación entre las mismas. Como por ejemplo proteínas con una localización celular compartida, genes con un nivel de expresión similar, proteínas de la misma familia o con un tipo de dominio compartido, genes con datos fenotípicos comunes (mismo tejido o implicación en una enfermedad), etc. Para una lista más detallada consultar la sección 5.1 o [Lee et al., 2007].

Bases de Datos de Interacciones y Asociaciones Funcionales

Aunque las bases de datos más desarrolladas y mantenidas son las de los elementos individuales de la red, como son los genes y las proteínas principalmente, también se deben almacenar las interacciones y asociaciones funcionales entre dichas moléculas.

Existen múltiples bases de datos de interacciones y asociaciones funcionales [Klingstrom and Plewczynski, 2011], con distintas características, en formato, contenido y localización. Los datos que se pueden encontrar en las diversas bases de datos son variados, conteniendo interacciones experimentales (de pequeña y gran escala) o predicción de interacciones y asociaciones funcionales por múltiples métodos computacionales. Hay muchos grupos experimentales con sus datos accesibles a través de una página web estática, aunque se han desarrollado interfaces genéricas, que permiten el acceso a múltiples repositorios distintos

desde un mismo punto, a través de servicios web, como los de la tecnología BioMOBY [Wilkinson and Links, 2002].

Las principales bases de datos de interacciones proteína-proteína son DIP [Salwinski et al., 2004], BIND [Alfarano et al., 2005], IntAct [Hermjakob et al., 2004b], MINT [Chatraryamontri et al., 2007] y, especialmente, BioGrid [Stark et al., 2006], que es el repositorio general de interacciones más amplio, agrupando otras bases de datos más pequeñas. Dichas bases de datos, entre otras, forman un consorcio internacional para armonizar las bases de datos de interacciones y facilitar el intercambio de información sobre interacciones, denominado IMEx (del inglés, *International Molecular Interaction Exchange consortium*). También existe un formato estándar para intercambiar datos de interacciones moleculares en XML, llamado PSI-MI (del inglés, *Proteomics Standards Initiative - Molecular Interaction*) [Hermjakob et al., 2004a], así como un estándar de mínima información necesaria para divulgar una interacción molecular experimental, MIMIx (del inglés, *Minimum Information required for reporting a Molecular Interaction experiment*) [Orchard et al., 2007].

Sobre las asociaciones funcionales, centrándose en rutas biológicas, las bases de datos más representativas son Reactome [Matthews et al., 2009], KEGG [Kanehisa and Goto, 2000] y MetaCyc [Caspi et al., 2010]. *Reactome* [Matthews et al., 2009] es una de las principales bases de datos de rutas, está verificada por expertos y centradas en el humano. Está organizada como una jerarquía. Es la que se utiliza en esta tesis para estudiar rutas biológicas. Al igual que BioGrid en las interacciones, *Pathway Commons* [Cerami et al., 2011] trata de agrupar diversas bases de datos de asociaciones funcionales de rutas biológicas.

También cabe destacar que existen bases de datos especializadas sólo en una especie concreta, cuando se dispone de datos particulares basados en estudios específicos para esa especie, quizá no conocidos para otras. Dichas bases de datos suelen contener datos más fiables que los procedentes de un estudio general para distintas especies, por tener en cuenta peculiaridades de la especie concreta. Un ejemplo de base de datos especializada es EcoCyc [Keseler et al., 2005] para la procariota *E.coli*, con un uso relevante en esta tesis.

Por último, la creciente importancia de la Biología de Sistemas, y la cantidad de investigadores dedicados a su estudio, ha dado lugar a la necesidad del intercambio de información formal, creando un lenguaje específico para ello denominado SBML (del inglés, *Systems Biology Markup Language*) [Hucka et al., 2004], así como una notación gráfica común, SBGN (del inglés, *Systems Biology Graphical Notation*) [Novere et al., 2009].

2.4.4. Redes en Sistemas Complejos y en Biología

La información biológica tiene varias características importantes [Ideker et al., 2001]:

- Interviene en múltiples niveles jerárquicos de organización.
- Se organiza en redes complejas.
- Son redes de información robustas, de forma que muchas perturbaciones pequeñas apenas afectan a la red.
- Existen nodos *clave*, cuya modificación podría ocasionar graves efectos en la red, tratándose de dianas importantes para la comprensión y manipulación del sistema.

La mayoría de estas propiedades de los datos biológicos coinciden con las de las denominadas redes libres de escala [Barabasi and Bonabeau, 2003]. Las redes libres de escala son redes complejas caracterizadas por la distribución de la cantidad de conexiones entre los

nodos, que sigue la ley de potencias: la probabilidad de que un nodo se conecte a otros k nodos es proporcional a $1/k^n$. Esta distribución define una función continua decreciente en el número de enlaces, describiendo una red con unos pocos nodos muy conectados (nodos *clave* o *hubs*), y la mayoría de nodos con pocos enlaces. Esta distribución es muy diferente a la de Poisson (con forma de campana, con una cantidad de enlaces entre nodos distribuidos homogéneamente) que siguen las redes aleatorias, las cuales han sido el modelo de referencia de todas las redes complejas, como las biológicas, durante mucho tiempo.

Las propiedades más importantes que caracterizan las redes libres de escala son:

- Existencia de algunos nodos 'populares', con cientos o miles de conexiones a otros, mientras que el resto sólo unas pocas, pareciendo que la red no tiene escala.
- Son robustas frente a fallos puntuales o accidentales.
- Son vulnerables a ataques específicos, por ejemplo, sobre un nodo *clave*.

Muchos sistemas actuales cumplen el modelo de las redes libres de escala: Internet, las colaboraciones científicas, las relaciones de amistad, la red comercial (en una tecnología específica, unas pocas empresas punteras tienen muchas más conexiones que el resto), etc. y, por supuesto, las redes biológicas a distintos niveles (como las redes de regulación de proteínas o el metabolismo celular).

No obstante, aunque todas estas redes comparten las propiedades mencionadas, las interacciones y asociaciones funcionales biológicas generalmente no pueden ser modeladas por completo simplemente con los principios de las redes libres de escala, por su mayor complejidad en diversos sentidos. Entre otros, las interacciones biológicas no son estables, ni en tiempo ni espacialmente; aunque las redes biológicas que representan dichos fenómenos de interacción sí son estables, pudiendo incluir todas las interacciones y asociaciones funcionales conocidas, aunque no sucedan simultáneamente. También, las redes biológicas son multi-función, y son muy variadas entre sí, por ejemplo, necesitando un análisis de propiedades particular para cada red. Las redes biológicas que se pueden conseguir matemática y computacionalmente hasta ahora son simplificaciones o limitaciones de la realidad biológica, que generalmente no alcanzan a representar toda su complejidad [Noble, 2006; Likic et al., 2010].

2.4.5. Aproximaciones al Estudio de las Redes Biológicas

Las aproximaciones computacionales para el estudio de las redes biológicas son diversas, siempre con el objetivo de analizar la información que contienen y de generar hipótesis [Junker and Schreiber, 2008]. Se basan principalmente en la teoría de grafos [Biggs et al., 1986; Newman, 2010], que se puede aplicar al estudio de las moléculas biológicas, por encontrarse estructuradas en una red [Huber et al., 2007].

Desde el punto de vista de la Inteligencia Artificial, en concreto del análisis de datos (del inglés, *data mining*), el uso de la teoría de grafos para extraer el conocimiento contenido en una red, se denomina minería de grafos (del inglés, *graph mining*) [Chakrabarti and Faloutsos, 2006]. Este área se aplica tanto en Biología como en redes sociales o en Internet, lo cual se conoce como minería de la web (del inglés, *web mining*). En esta sección se presenta brevemente el uso de la teoría de grafos, pero centrado principalmente en su aplicación a redes biológicas.

La representación de las redes complejas como grafos ha permitido su análisis sistemático usando los conceptos teóricos de grafos, para sugerir nuevas hipótesis sobre la topología y la función de las redes biológicas [Aittokallio and Schwikowski, 2006].

Una vez que se construye la red, los enfoques de análisis fundamentales son la caracterización de la topología y la localización de módulos funcionales. Adicionalmente, se pueden realizar comparaciones de redes biológicas entre diferentes especies.

Construcción de Redes Biológicas

El diseño de las redes biológicas es un paso previo al estudio de las mismas con la teoría de grafos. La información de entrada son las interacciones y asociaciones funcionales entre genes o proteínas determinadas experimentalmente, por ejemplo, capturadas en experimentos de expresión génica o secuenciación a gran escala, en auge en los últimos años [Armañanzas et al., 2012]. La construcción de redes se lleva a cabo con distintos métodos [Cho et al., 2007] (sistemáticos o a medida, de Aprendizaje Automático o no), entre los que destacan el uso de las redes bayesianas [Markowetz and Spang, 2007]. Las redes diseñadas pueden ser dirigidas o no dirigidas, sin o con pesos asociados a los arcos. Por ejemplo, una red de regulación de la transcripción se puede modelar como un grafo dirigido con pesos, donde el peso de los arcos representa el grado del efecto regulador de un factor de transcripción (nodo origen) a sus genes regulados (nodos destino) [Aittokallio and Schwikowski, 2006].

Cabe destacar que los datos experimentales disponibles suelen ser insuficientes para la construcción de la red completa, además de poder ser datos ruidosos [Aittokallio and Schwikowski, 2006]. Estos hechos provocan dificultades en el estudio de las redes biológicas por ser incompletas, debido a dicha falta de conocimiento, especialmente en las redes de interacciones proteína-proteína. Por lo tanto, para afrontar estos retos, son necesarios enfoques que refinen o completen las redes que han sido previamente definidas, limitando el diseño a redes más pequeñas o integrando información adicional, respectivamente. Este último enfoque de extensión de redes es el adoptado en las propuestas de esta tesis, siguiendo la línea de diversos métodos computacionales, que integran múltiples y heterogéneas fuentes de datos para la predicción de redes [Marc, 2005].

Caracterización de la Topología

Una vez que se tiene una red definida, el nivel de análisis más general consiste en caracterizar la estructura global de la red, utilizando sus propiedades topológicas cuantitativas [Assenov et al., 2008]. Algunos de los parámetros topológicos calculados con más frecuencia para describir una red son el grado de conectividad de los nodos de la red, el coeficiente de agrupamiento, el diámetro, la densidad y la heterogeneidad de la red. Con estas propiedades globales se puede determinar si la red sigue una topología libre de escala, de alta frecuencia en las redes biológicas.

De este estudio global cuantitativo sólo se obtiene un conocimiento limitado de la red, que generalmente se complementa con un análisis local de conectividad entre pares de nodos. Del análisis local, aplicando algoritmos convencionales de la teoría de grafos, se pueden extraer propiedades locales, como son la cantidad y complejidad de los *sub-grafos* contenidos en la red, la longitud del *camino* más corto entre pares de nodos conectados indirectamente o la presencia de *nodos centrales* o esenciales en la red. Estas propiedades aportan conocimiento relevante, dado que la centralidad de un nodo puede identificar posibles dianas de fármacos, o

la redundancia de caminos entre nodos puede explicar la robustez de ciertos procesos celulares [Albert, 2005].

Localización de Módulos Funcionales

Para tratar la complejidad de las redes grandes, se lleva a cabo una descomposición en grupos de moléculas asociadas funcionalmente. Cada módulo se puede comparar con otros datos del genoma a gran escala para generar hipótesis funcionales sobre la sub-red localizada.

Existen varios tipos de módulos funcionales, dependiendo de cómo se describen. Los principales son dos: los patrones (del inglés, *motif*), conocidos en minería de grafos como sub-grafos frecuentes, y las agrupaciones (del inglés, *cluster*) [Aittokallio and Schwikowski, 2006].

Los patrones son sub-grafos que aparecen con una frecuencia significativamente mayor que la esperada por azar. Para detectarlos, se deben aplicar técnicas para calcular los sub-grafos presentes en una red, agruparlos en clases de sub-grafos isomorfos y determinar las clases más frecuentes de lo esperado en modelos computacionales aleatorios.

Una alternativa para la identificación de módulos funcionales es el descubrimiento de áreas densamente conectadas (agrupaciones), potencialmente involucradas en funciones celulares o complejos de proteínas comunes. El agrupamiento de grafos se basa en la asunción de que un grupo de nodos asociados funcionalmente es más probable que tengan una alta conectividad entre ellos y estén más separados del resto de la red. Además, los módulos no suelen ser componentes de la red disjuntos, sino que comparten nodos, enlaces e incluso funciones, siguiendo preferentemente una organización jerárquica, lo que se debe tener presente en el diseño de algoritmos de agrupación de redes [Barabasi and Oltvai, 2004]. Existen algoritmos que utilizan diferentes estrategias de búsqueda local, con distintas heurísticas basadas en propiedades de la red, otros basados en distancias, y otros supervisados [Aittokallio and Schwikowski, 2006].

2.5. Discusión y Problemas Biológicos Afrontados

A pesar de que se lleva trabajando desde hace bastantes años en Bioinformática, aún hay muchos problemas por resolver computacionalmente en Biología Molecular, en particular en anotación funcional. Aunque se disponga de la secuencia de una proteína o gen, mientras no se conozca cuál es su función o en qué procesos biológicos está implicado, la caracterización biológica no está resuelta. Una parte de los genomas secuenciados está anotada, acorde a métodos *ad-hoc* y técnicas de aplicación parcial, pero otra gran parte aún no lo está. De todos los genes conocidos, al menos el 50 % presentan una anotación funcional ambigua (desconocida, probable o tentativa) [Hawkins and Kihara, 2007]. Por tanto, la anotación funcional completa de genes y proteínas continúa siendo uno de los retos bioinformáticos más importante, y es en este área biológica en la que se centra esta tesis.

Además, según se presenta en la sección 2.4, en la última década, con la aparición de la Biología de Sistemas, la anotación funcional ya no sólo se considera de forma individualizada para cada gen o proteína del organismo. Sino que se debe tener presente una visión global del sistema biológico, con todas las asociaciones existentes entre los genes y proteínas conectados a distintos niveles en múltiples redes.

En esta tesis se relaciona la anotación funcional con la Biología de Sistemas desde dos puntos de vista complementarios. Primero, considerando las interacciones y asociaciones

funcionales de un gen o proteína (en pares o en grupos) como un tipo de anotación en sí misma. Segundo, usando dichas relaciones (parte de la red o completa) como fuente de información con la que anotar compuestos biológicos en otro vocabulario (como por ejemplo con las funciones de la red), incluso para anotar con un tipo de asociación diferente, a distinto nivel.

Para estudiar algunos aspectos de la anotación funcional encuadrada en la Biología de Sistemas, se escogen dos dominios específicos y de relevancia biológica:

1. El primer problema consiste en predecir asociaciones funcionales entre pares de proteínas de *E.coli*, en un sentido amplio, sin necesidad de que exista interacción física. Se puede considerar una forma de anotación, donde en lugar de asignar a una proteína un conjunto de términos en un vocabulario concreto, se le asignan las proteínas con las que se asocia funcionalmente. Igualmente se puede ver como una extensión de una red punto a punto.

2. A su vez, las interacciones por pares se pueden emplear como entrada para predecir anotación funcional en otro vocabulario, en combinación con otras propiedades simples, como sucede en el segundo problema seleccionado. En este caso, se anota funcionalmente una proteína humana con las rutas biológicas en las que participa, es decir, se quiere predecir la pertenencia de una proteína a una ruta (pertenencia a un grupo con una anotación común). También se puede considerar como predicción de la asociación funcional de una proteína con un grupo, o extensión de una red en grupo. Además, hay que destacar el interés por extender la red en distintas zonas, es decir, con proteínas diversas en su función molecular, acorde a la heterogeneidad molecular de las rutas originales.

En conclusión, se trata de dos problemas diferentes, pero con un objetivo común encuadrado en la Biología de Sistemas: extender redes, o lo que es lo mismo, anotar proteínas con interacciones o asociaciones funcionales.

Por otro lado, el Aprendizaje Automático (descrito en la sección 2.1) se presenta como una opción para automatizar y unificar los mecanismos de anotación de genomas y proteomas. En concreto, el Aprendizaje Relacional (sección 2.2) y, en especial, la Representación Relacional en la que se basa, podría ser una alternativa interesante para intentar generalizar la representación de los datos de origen biológico y todas sus relaciones, aprovechando toda la semántica del contexto biológico, intrínsecamente estructurado y con todos sus elementos exhaustivamente relacionados, tal y como se consideran en la Biología de Sistemas.

Para afrontar los dos problemas elegidos se propone definir una representación relacional para anotación de función, y posteriormente realizar las transformaciones necesarias en la representación del conocimiento, para resolver cada problema con las aproximaciones de Aprendizaje Automático más adecuadas. Se elige entre los enfoques y algoritmos proposicionales o relacionales presentados en las secciones 2.1 y 2.2, prefiriendo aquellos que cumplan las condiciones biológicas exigidas, como gestionar adecuadamente los valores desconocidos (como AODE), y afrontar con éxito una clasificación multi-clase y multi-etiqueta (como la combinación WARMR y CLUS). En ocasiones se menciona que los árboles de decisión no son la mejor técnica para predecir función [Al-Shahib et al., 2005], porque separan el espacio de búsqueda en rectángulos, y la naturaleza de los datos biológicos tiende a ser más compleja, similar a los hiper-planos de las máquinas de vector de soporte, que tienen en cuenta varios atributos a la vez. No obstante, los árboles de decisión permiten una mayor comprensión del modelo de conocimiento extraído. Además, se puede modificar la forma del espacio de búsqueda de los árboles, mediante un cambio en la representación de los datos, como se hace

en esta tesis al partir de una representación relacional, que se convierte en una proposicional, la cual incluye en cada atributo de entrada al árbol una combinación de los atributos originales.

El Aprendizaje Automático se plantea como una propuesta para integrar con relativa facilidad distintas fuentes de información biológica y, a través de la generalización, llegar a dos enfoques unificados de predicción de asociación funcional, en un caso puntual y en otro grupal.

Se trata de dos problemas de anotación que comparten la presencia de las relaciones biológicas como una parte fundamental de ambos, tanto en la información de entrada como en el objetivo de predicción de anotación; enfocados en el uso exclusivo de las secuencias y de pares de asociaciones funcionales o interacciones, sin propiedades complejas procedentes de la teoría de grafos. Pero estos dos problemas presentan diferencias interesantes por las que también se han seleccionado. En particular, divergen en los datos (origen y características), en las evidencias funcionales de origen biológico, y en el enfoque de aprendizaje requerido. Respecto a los datos, ambos problemas se diferencian principalmente en la especie de aplicación (procariota frente a eucariota), en el nivel de abstracción de la información de entrada (preprocesados frente a datos simples de la secuencia) y en la cantidad de relaciones diferentes definidas por proteína (quince frente a cientos de ellas). La base biológica del primer problema es la información evolutiva, mientras que en el segundo se trata de predecir función en ausencia de evidencias evolutivas (sin homología), basándose en propiedades simples de la secuencia. Sobre el aprendizaje, las diferencias fundamentales son la clasificación binaria asignando una clase a cada ejemplo, frente a la clasificación multi-clase y multi-etiqueta, ya comentadas.

Para concluir, y acorde a la lista de retos descrita en la sección 2.3.2, se puede ver que esta tesis pretende afrontar dos problemas de anotación complejos, con muchos aspectos biológicos a tener siempre presentes en la toma de cada decisión de análisis, diseño y resolución. La finalidad es llegar a una aportación válida y útil en el ámbito biológico, y no sólo una simple aplicación de un método computacional a datos reales, sin simplificar completa ni parcialmente su contexto, ni ignorar su semántica asociada.

Capítulo 3

Objetivos y Alcance

Teniendo en cuenta el estado de la cuestión descrito en el capítulo 2, y en particular las conclusiones presentadas en la sección 2.5, la **motivación** general de esta tesis es contribuir al análisis y extracción de conocimiento automático de la gran cantidad de datos que se generan a diario en Biología Molecular, cuyo ritmo de crecimiento supera los procedimientos clásicos, manuales y experimentales, limitados en tiempo y coste. En particular, se quiere conocer cómo afecta el contexto relacional y su uso a la anotación funcional del genoma y el proteoma. Ya que existen muchas relaciones entre compuestos, en vez de estudiar estos elementos individualmente, en Biología Molecular actualmente se enfocan los problemas desde las redes de interacción. Como ya se ha mencionado, desde un punto de vista biológico, se denomina Biología de Sistemas, y desde el computacional, se puede encuadrar en la Representación y el Aprendizaje Automático Relacional.

Así, la **propuesta de tesis** es estudiar el área de predicción de anotación funcional centrado en la Biología de Sistemas, a través de la representación relacional y el Aprendizaje Automático.

Alcance

- Afrontar dos casos específicos dentro del área de anotación funcional, que son anotar proteínas con asociaciones funcionales por pares y, utilizar propiedades simples para anotar proteínas con potenciales funciones similares en cadenas de reacciones biológicas.
- Analizar los resultados en términos de la potencial aplicación del Aprendizaje Automático a datos de Biología Molecular en el contexto de problemas de anotación y etiquetado.
- Entender el área de anotación funcional centrado en la Biología de Sistemas, sin pretender cubrir todos los aspectos, ni solucionar todos los problemas del área; limitándose a anotar proteínas con asociaciones funcionales, basándose en interacciones y asociaciones (por pares o en redes).
- Interpretar los resultados en términos biológicos, discutiendo las diferencias entre soluciones óptimas en términos biológicos y computacionales.
- Discutir la importancia de las consideraciones biológicas y detalles específicos del área de aplicación, guiando las decisiones principalmente por las restricciones o intereses biológicos, más que por los computacionales.

Objetivos

1. Diseñar un modelo de representación del conocimiento de un área específica de Biología Molecular con un enfoque relacional, que incluya genes y productos genéticos, con sus propiedades y relaciones, válido para la predicción de anotación funcional en la Biología de Sistemas.

2. Recopilar, integrar y procesar diferentes fuentes de información de relevancia biológica para la anotación funcional, construyendo así un conjunto de datos propio, prestando especial atención a la selección de datos actualizados y verificados por expertos.

3. Afrontar dos problemas reales y diferentes de predicción de anotación funcional en Biología de Sistemas, teniendo en cuenta todas las restricciones y la complejidad del entorno biológico correspondiente. En concreto, se propone predecir asociaciones funcionales entre pares de proteínas en *E.coli*, y extender rutas biológicas en Humanos con proteínas predichas por su implicación en las mismas.

4. Proponer, comparar y analizar distintas representaciones del conocimiento de datos de genómica y proteómica, evaluando su implicación en el proceso de Aprendizaje Automático para anotación funcional.

5. Analizar la relevancia del enfoque de la Biología de Sistemas en el Aprendizaje Relacional, es decir, de la importancia de las relaciones entre elementos biológicos, para predecir anotación funcional con aprendizaje automático.

6. Explorar, comprender y concluir los posibles enfoques a seguir en la aplicación de Aprendizaje Automático a la resolución de otros problemas de anotación funcional en Biología Molecular.

7. Usar combinaciones de componentes estándar de Inteligencia Artificial y Bioinformática (herramientas, algoritmos de aprendizaje, métodos de predicción, etc.).

Capítulo 4

Metodología de Evaluación

Este capítulo presenta y justifica el sistema elegido para la evaluación de la consecución de los objetivos de esta tesis doctoral.

Los objetivos se diferencian en dos grupos. Uno primero donde sólo se puede verificar la existencia de un modelo de representación de datos relacional para Biología Molecular, un conjunto de datos que cumplan con las consideraciones establecidas, y el uso de componentes estándar; abarcando los objetivos 1, 2 y 7, respectivamente. Y un segundo grupo o caso de evaluación, que requiere de un análisis experimental, comprendiendo los restantes objetivos, del 3 al 6.

En las siguientes secciones se exponen, primero, los criterios generales para evaluar la parte experimental de la tesis, y segundo, las medidas específicas utilizadas. Las medidas de evaluación se dividen en aquellas centradas en valorar el rendimiento de la clasificación y las que permiten interpretar y analizar los resultados. Se describen en más detalle las más relevantes durante la tesis (curva PR y ROC) o menos frecuentes (curva de coste, media-macro y media-micro y similitud semántica).

4.1. Enfoque de Evaluación Experimental

Es muy importante tener en cuenta que para la parte experimental se utiliza una evaluación distinta a las medidas clásicas de Aprendizaje Automático, ya que en este caso:

- No sólo es relevante comparar las nuevas propuestas con diferentes configuraciones y diferentes algoritmos de Aprendizaje Automático, sino con otros métodos no basados en inteligencia computacional, pero de referencia en el campo de la anotación funcional en bioinformática.
- No sólo se evalúa sobre ejemplos conocidos de un conjunto de test, sino sobre otros nuevos de los que se desconoce su anotación.
- No sólo se valora el rendimiento, sino también la interpretación y análisis de los resultados, y en gran medida. Por una parte, porque es necesario comprender lo que implica y significa biológicamente un conjunto de predicciones. Por otra parte, porque en esta tesis se quiere entender cómo y de qué aprenden realmente los sistemas de clasificación propuestos.
- Además se evalúan los sistemas definidos según la relevancia biológica de su aplicación a datos reales, estudiando casos concretos, para valorar la calidad y utilidad de dichos

modelos de clasificación.

Los resultados experimentales de esta tesis se evalúan fundamentalmente mediante la comparación de los sistemas propuestos con otros, para evaluar la bondad de los resultados en diferentes aspectos:

- Comparación con la aleatoriedad o valor por defecto: comprobar si se supera el mínimo admisible determinado por una predicción aleatoria o mayoritaria. En situaciones muy restringidas en Biología, basta con lograr ser ligeramente mejor que la aleatoriedad, aunque las medidas estándar estén lejos de su límite superior.

- Comparación con otras configuraciones del mismo algoritmo.

- Comparación con otros algoritmos de Aprendizaje Automático: sobre un mismo conjunto de proteínas o genes.

- Comparación con otras representaciones del conocimiento (relacional, proposicional, variantes y combinaciones).

- Comparación con otros métodos bioinformáticos: realizar una evaluación externa con algún método de predicción que resuelva la misma tarea de anotación funcional (sin necesidad del uso de Inteligencia Artificial), si existe y está disponible un método comparable.

Siempre que sea posible, las comparaciones se realizan sobre las mismas proteínas o genes, e incluso sobre los mismos conjuntos de entrenamiento y test.

Para todas las comparaciones, se usan medidas combinadas de evaluación de rendimiento y de interpretación de resultados. Se elige entre los dos grupos de medidas de evaluación que se presentan en la siguiente sección.

4.2. Medidas de Evaluación

Las medidas utilizadas para evaluar los resultados de esta tesis se pueden descomponer en dos grupos, según su finalidad:

1. Evaluación del rendimiento en la clasificación. Se refiere a las medidas clásicas de Aprendizaje Automático.

2. Interpretación y análisis de las predicciones, principalmente en un contexto biológico. Se trata de medidas adicionales, que varían según el aspecto concreto que se quiere analizar.

Dependiendo del resultado a evaluar, se elige un subconjunto de medidas, que combina las de ambos grupos. Por ejemplo, en el capítulo 8, que presenta otros enfoques de AA para resolver problemas de anotación funcional, se elige un subconjunto de medidas compuesto por: 1.- dos medidas de rendimiento (AUPRC y AUROC) y 2.- cuatro medidas de interpretación y análisis de las predicciones (número de rutas biológicas extendidas o de clases predichas, número de proteínas añadidas o de nuevos ejemplos predichos, similitud funcional semántica y solapamiento entre las proteínas añadidas a cada clase).

4.2.1. Evaluación del Rendimiento en la Clasificación

Se trata de medidas cuantitativas estándar de evaluación del rendimiento en tareas de Aprendizaje Automático. Pueden ser medidas uni-dimensionales o bi-dimensionales.

Dado que el enfoque adoptado en esta tesis para realizar anotación funcional es fundamentalmente la clasificación, se deben considerar las medidas de evaluación utilizadas tradicionalmente en dicho paradigma. Existen múltiples medidas, tanto numéricas (uni-dimensionales) como gráficas (bi-dimensionales), para validar un proceso de clasificación [Baldi et al., 2000].

Tradicionalmente, las medidas numéricas o uni-dimensionales de rendimiento más frecuentemente usadas son la tasa de aciertos global, la sensibilidad (tasa de aciertos en positivos) (en inglés, también denominada *recall*), la especificidad (tasa de aciertos en negativos), la precisión (tasa de aciertos en positivos, sobre el total de predichos como positivos), o verdaderos y falsos positivos y negativos (*TP, FP, TN y FN*) [Baldi et al., 2000]. También existen medidas unificadas como el coeficiente de correlación de Matthews (del inglés, *Matthews Correlation Coefficient, MCC*) [Matthews, 1975], que combina los *TP, FP, TN y FN* en una sola medida o valor; o la medida F (del inglés, *F-measure* o *F-score*) que unifica la precisión y la sensibilidad, con diferentes aproximaciones [van Rijsbergen, 1979].

En esta tesis, aunque ocasionalmente se pueden utilizar las anteriores, para medir el rendimiento en clasificación se prefiere usar curvas, para no depender de un umbral de confianza en la predicción fijo, frente a las medidas uni-dimensionales que lo suelen necesitar. Entre todas ellas, en esta tesis se seleccionan como más adecuadas las curvas PR (del inglés, *Precision-Recall Curves*), las curvas ROC (del inglés, *Receiver Operating Characteristic*) y las curvas de coste (del inglés, *Cost Curves*). Mayoritariamente se usan curvas PR y ROC (y sus uni-dimensionales correspondientes, AUPRC y AUROC), que es lo más utilizado y comprensible en términos bioinformáticos.

Curvas PR y ROC

Las **curvas PR** [Davis and Goadrich, 2006] representan el ratio de acierto en los positivos predichos o precisión (eje *y*) frente al ratio de acierto en los positivos reales o sensibilidad (eje *x*). Se puede ver un ejemplo en la figura 7.7(a) del capítulo 7.

Las curvas PR son más adecuadas que las curvas ROC [Davis and Goadrich, 2006] cuando se trabaja con conjuntos de datos muy sesgados, es decir, con mucha diferencia en el número de ejemplos de cada clase. Este es el caso de la anotación funcional multi-clase, donde una clase sólo se asigna a unos pocos ejemplos, los cuales se suelen considerar ejemplos negativos para el resto de clases (o la gran mayoría de ellas). De forma que en el conjunto total hay muchos más ejemplos negativos que positivos, para cada una de las clases en particular. Las curvas PR se centran principalmente en analizar los aciertos de la clase positiva, sin prestar apenas atención a los aciertos de la clase negativa, a los que las curvas ROC les da una mayor importancia [Davis and Goadrich, 2006]. Sin embargo, en el dominio de anotación funcional en Biología Molecular interesa principalmente ser certero en las predicciones positivas para un gen o proteína, es decir, las funciones que tiene asociadas, importando menos la precisión en las funciones no relacionadas.

Las **curvas ROC** [Fawcett, 2003] representan el ratio de verdaderos positivos o sensibilidad (eje *y*) frente al ratio de falsos positivos o 1-especificidad (eje *x*). Se puede ver un ejemplo en la figura 7.7(b).

A lo largo de la tesis, también se presentan las curvas ROC equivalentes a las PR elegidas,

porque son más usadas en dominios biomédicos, y preferidas por los expertos biólogos.

Tanto las curvas PR como las curvas ROC se pueden transformar en una medida unidimensional calculando el área bajo su curva (denominados AUPRC o AUROC), factores que se emplean en esta tesis muy frecuentemente.

Curvas de Coste

Las **curvas de coste** [Drummond and Holte, 2006] son una técnica gráfica que permite visualizar el rendimiento (ratio de error o coste esperado) obtenido por distintos clasificadores aplicados a problemas binarios, para el rango completo de posibles distribuciones de clase y de costes de error esperados en la clasificación. Se puede ver un ejemplo en la figura 6.3 en el capítulo 6.

En una interpretación simple, una curva de coste es un gráfico bi-dimensional, que representa la probabilidad de coste (eje x), equivalente [Drummond and Holte, 2006] al porcentaje de instancias de la clase positiva que hay en el conjunto de datos sobre el que se va a aplicar el clasificador, frente al coste esperado normalizado (eje y), que es equivalente al ratio de error cometido, tanto en términos de falsos positivos como de falsos negativos. La interpretación previa asume que el coste de error en la clasificación de ejemplos positivos (FN) es el mismo que el coste de error en la clasificación de ejemplos negativos (FP). Pero cuando el coste de error en la clasificación es diferente entre los positivos y negativos, la interpretación varía. En ese caso, el eje x de la curva de coste no sólo representa la fracción de instancias positivas, sino el producto del coste de error y de la probabilidad de que una instancia sea de la clase positiva. Mientras que el eje y indica la fracción de la diferencia entre los costes máximo y el mínimo posibles en los que se incurre al usar el clasificador [Drummond and Holte, 2006]. Por lo tanto, el eje y muestra el coste esperado normalizado correspondiente al escenario de probabilidad de coste y distribución de clases indicado por el valor del eje x.

En el lado izquierdo del eje y se mide el ratio de falsos positivos (FP) en orden creciente, y en el lado derecho se mide el ratio de verdaderos positivos (TP) en orden decreciente. En consecuencia, la curva de coste de un clasificador se construye a partir de diferentes líneas rectas, las cuales tienen sus extremos a ambos lados del eje y, correspondiendo a varios pares $<$*ratio FP, ratio TP*$>$ obtenidos para diferentes umbrales de clasificación. Los fragmentos de línea no dominados por ninguna otra (es decir, los más bajos) componen la curva de coste completa. Así, en una sola curva se representan los modelos obtenidos tras aplicar distintos umbrales de discriminación entre clases, aplicados sobre la probabilidad a posteriori dada por el clasificador.

Las curvas de coste permiten comparar fácilmente varios clasificadores, representados en una misma figura, procedentes de la aplicación de distintos algoritmos de Aprendizaje Automático o del empleo de diferentes conjuntos de entrenamiento y/o test. La representación gráfica de la curva de coste de un clasificador normalmente contiene también la curva correspondiente al clasificador trivial (línea roja en la figura 6.3, que forma un triángulo), el cual siempre asigna la misma clase a cualquier ejemplo. La curva de coste de un clasificador útil siempre debería estar por debajo de la curva del clasificador trivial, lo cual indica que es un buen clasificador para cualquier distribución entre positivos y negativos en el conjunto de datos. En un gráfico con varias curvas de coste, el mejor clasificador es el que tenga la curva más baja (coste esperado más bajo). Por lo tanto, los puntos de corte (si existen) de una curva con la del clasificador trivial, determinan el rango del eje x para el que no es adecuado usar dicho clasificador, ya que una clasificación por defecto tiene mejor rendimiento. Como regla genérica, a la hora de valorar la bondad de una curva de coste, ésta se puede considerar válida

si toma valores de probabilidad de coste (eje *y*) inferiores a 0,3.

Es interesante destacar que se puede construir una curva ROC equivalente para cada curva de coste. La correspondencia entre estas dos técnicas gráficas es la siguiente: un punto en una curva ROC se corresponde con una línea en la curva de coste equivalente. Las coordenadas de un punto en la curva ROC son los extremos izquierdo y derecho en los ejes *y* de la curva de coste. Cada línea en una curva de coste se compone de muchos clasificadores procedentes de dos variables: diferentes umbrales y diferentes proporciones de positivos en el conjunto de datos. Así, la mayor ventaja de las curvas de coste frente a las ROC es que permiten un comparación directa del rendimiento para cualquier combinación de coste de error y distribución de clases. Aunque ambas curvas permiten mostrar el rendimiento para diferentes umbrales de clasificación, las curvas de coste presentan una información más detallada sobre el rendimiento frente a distintas distribuciones de clases, porque tienen una línea para representar este aspecto, no sólo un punto como las curvas ROC. Otra característica positiva de las curvas de coste es que permiten comparar de inmediato varios clasificadores. Porque en las curvas de coste, la diferencia de error entre un par de clasificadores se puede medir automáticamente a través de la distancia vertical, lo cual no es tan fácil en las curvas ROC, donde se deben combinar las distancias verticales y horizontales [Drummond and Holte, 2006].

En resumen, una curva de coste es equivalente a una curva ROC en la información a partir de la que se construyen, pudiendo convertirse una en otra. Será más adecuado usar curvas de coste cuando se necesite seleccionar el mejor clasificador mediante una simple visualización, y siempre que se conozcan unas ciertas condiciones, como el coste de error en la clasificación y la probabilidad de aparición de la clase positiva en el conjunto de datos sobre el que aplicar el clasificador.

Media-micro y Media-macro

En el dominio de predicción de anotación funcional muchas veces se trabaja con una predicción multi-clase: un gen o proteína tienen más de una anotación del mismo tipo. Se suele obtener un predictor binario diferente por función (clase), siendo los ejemplos positivos las proteínas que tienen asociada la clase a predecir, y los negativos el resto de proteínas sin esa clase asociada (enfoque *1-contra-todos*, aunque con clases solapadas) [García-Pedrajas and de Haro García, 2008; Lee et al., 2009]. Por tanto, los resultados se pueden presentar separados por clases individuales, pero también se suele calcular una media para todas las clases, para tener una evaluación global conjunta. No obstante, hay que analizar la forma más adecuada de calcular dichos valores promedio, tanto en una como en dos dimensiones.

Existen dos métodos convencionales para evaluar el rendimiento medio sobre todas las clases en un problema de aprendizaje multi-clase, denominados media-macro (del inglés, *macro-average*) y media-micro (del inglés, *micro-average*) [Yang, 1999].

- En la *media-macro*, primero se calcula independientemente para cada clase la medida en cuestión (precisión, tasa de aciertos en positivos, área bajo la curva, etc.), con una tabla de contingencia individual, y después se promedian estas medidas por clase, para obtener la media global. Es decir, primero se evalúa localmente y luego globalmente.

- En la *media-micro*, directamente se calcula una tabla de contingencia global, donde cada celda contiene la suma de las celdas correspondientes de las tablas de contingencia individuales de cada clase, y entonces se usa esta tabla global para calcular la medida.

Extendiendo la definición de medida uni-dimensional a bi-dimensional, la curva de media-macro es la media de todas las curvas individuales; mientras que la curva de media-

micro calcula cada uno de sus puntos (par <precisión,sensibilidad> en curva PR o par <sensibilidad,1-especificidad>) contabilizando todos los ejemplos, de todas las clases, a la vez, para cada uno de los umbrales considerados.

Existe una gran distinción entre las medidas promedio *macro* y *micro* [Sebastiani, 2002], pudiendo dar resultados bastante diferentes, sobre todo si las clases tienen distintas frecuencias sobre el conjunto de ejemplos. La media-micro da un peso equivalente a todos los ejemplos, y por lo tanto se considera una media por ejemplo, específicamente, una media por pares ejemplo-clase. Sin embargo, la media-macro da el mismo peso a cada una de las clases, sin importar su frecuencia, siendo por tanto una media por clase. Usar una u otra medida depende de los requisitos del problema.

En los problemas de anotación funcional no suele existir una distribución homogénea de ejemplos en clases. Es decir, a nivel molecular, una función la pueden realizar muchas proteínas diferentes, y otra diferente ser una función en la que sólo están especializadas unas pocas proteínas; y a nivel de proceso, una función puede necesitar la colaboración de cientos de proteínas para llevarse a cabo, aunque otra función diferente pueda concluirse con éxito con sólo diez proteínas. Por lo tanto, en esta tesis se decide usar principalmente la media-macro, para darle la misma importancia a lograr una buena predicción en todas las clases por igual, para no favorecer que las más frecuentes se predigan mejor que las minoritarias, como podría suceder al optimizar una media-micro. No obstante, en algunas secciones de la tesis se muestra la media-micro de ciertas evaluaciones multi-clase, para compararlas con los resultados de la media-macro, demostrando experimentalmente las diferencias entre ambas.

4.2.2. Interpretación y Análisis de las Predicciones

Este segundo grupo de medidas de evaluación, para la interpretación y análisis de la predicción, tienen como finalidad comprender los resultados obtenidos y valorarlos en su contexto de problema real. Para esta tarea, a lo largo de la tesis se usan medidas variadas, según el objetivo biológico específico buscado, como por ejemplo, diversidad en las predicciones en una misma clase, solapamiento entre predicciones en distintas clases, cobertura, relevancia de atributos, significado de las reglas de decisión, similitud semántica o anotaciones funcionales en bases de datos y en la literatura científica.

Cabe destacar que las metodologías de interpretación de los resultados no se utilizan de forma exclusiva sobre las proteínas no anotadas (no etiquetadas), sino que, en algunos casos, también se pueden emplear sobre los conjuntos de elementos anotados usados en el entrenamiento y test.

Este conjunto de medidas de interpretación y análisis es amplio, pero la mayoría son medidas sencillas, de forma que se explican a lo largo de los capítulos 6, 7 y 8, cuando se utilizan por primera vez. Sin embargo, las medidas de similitud semántica sí requieren de una breve introducción, así como una explicación del uso concreto que se hace de ellas en esta tesis.

Similitud Semántica

La similitud semántica es una medida que estima cuantitativamente la similitud funcional entre genes y productos genéticos, a través de sus anotaciones funcionales.

Ante la falta de una evaluación experimental de anotaciones nuevas, la similitud semántica es una alternativa que se ha propuesto como medida de evaluación en reuniones de predicción funcional automática [Friedberg, 2006] (como en *Automated Function Prediction 2005 meeting*), como aproximación para evaluar la predicción de anotaciones. Además, la similitud

semántica se utiliza para otros análisis en Biología Computacional, en los que se necesita conocer el grado de relación de dos o más genes o proteínas en términos de sus anotaciones. Incluso existen diversas herramientas que calculan dichas similitudes automáticamente, como *GOSemSim* [Yu et al., 2010], que se emplea en esta tesis.

Las anotaciones funcionales consideradas deben pertenecer a una ontología, como por ejemplo la Ontología Génica (del inglés, *Gene Ontology, GO*) [Ashburner et al., 2000]. Aunque la ontología de procesos biológicos de *Gene Ontology* (GO-BP) no fue diseñada para evaluar anotaciones funcionales individuales, porque no todas las relaciones entre los términos de GO-BP se apoyan en asociaciones funcionales reales [Chagoyen and Pazos, 2010], es el vocabulario de anotación más extendido y usado, también empleado para evaluar predicción de anotación funcional [Lord et al., 2003].

Las medidas de similitud semántica requieren un vocabulario estructurado, con un conjunto de términos fijos y relacionados, porque las medidas de similitud se basan en la teoría de la información (las que se usan en esta tesis) o están basadas en el grafo de la ontología [Pesquita et al., 2009]. En la teoría de la información, se considera la frecuencia de aparición de un término de anotación en un corpus (por ejemplo la base de datos UniProt), considerándolo más informativo cuanto menos aparezca en el corpus, y menos informativo cuanto más aparezca, porque significa que es un término más general. En GO, dicha frecuencia incluye la suma de probabilidades de aparición de todos sus nodos hijo, siendo el valor de frecuencia 1, el máximo, para la raíz de la ontología (por ejemplo, el término *'molecular function'* en GO-MF). Así, para calcular la similitud funcional entre dos términos GO, se busca el ancestro común más bajo en la jerarquía, y se traduce su frecuencia a medida de similitud. Es decir, cuanto más alta es la frecuencia, menor es la similitud [Lord et al., 2003].

Sin embargo, un gen o proteína suele tener más de un término GO asociado, por lo que si se quieren comparar 2 productos genéticos, se necesita extender esta definición de similitud entre 2 términos, a medidas semánticas entre 2 proteínas o genes (es decir, entre dos listas de términos GO). Para ello se han desarrollado varios enfoques [Pesquita et al., 2009]. Aunque hay algunas medidas basadas en grafos, los métodos más comunes se basan en calcular las similitudes por pares de términos y luego combinarlas, calculando principalmente la media, el máximo, o la media de los mejores (del inglés, *best-match average*). Se debe elegir la medida y combinación más conveniente según el objetivo buscado con la medida de similitud semántica. En *Pesquita et al.* se puede consultar una extensa y detallada revisión de las medidas de similitud semántica [Pesquita et al., 2009].

En esta tesis, como parte de la interpretación y análisis de los resultados, se utilizan en varios puntos las medidas de similitud semántica calculadas sobre las ontologías Proceso Biológico (GO-BP) y Función Molecular (GO-MF) de *Gene Ontology*. Se utiliza GO-BP para calcular la similitud de las proteínas en el contexto de las rutas biológicas, y GO-MF para calcular la cohesión funcional de un grupo de proteínas a nivel molecular.

En concreto, se definen tres objetivos de uso de medidas de similitud semántica, aplicadas en el capítulo 7:

1. **Objetivo 1**: similitud entre 2 conjuntos: comparar proteínas que extienden una ruta biológica con las originales de la ruta.

2. **Objetivo 2**: similitud entre 2 conjuntos: comparar proteínas que extienden una ruta biológica por 2 sistemas de predicción distintos.

3. **Objetivo 3**: similitud de proteínas de un mismo conjunto entre sí.

Como en los objetivos planteados se comparan más de 2 proteínas, es decir, conjuntos de N proteínas con M términos GO cada una, se debe extender a un tercer nivel la definición de similitud semántica, para determinar cómo se combinan las similitudes de una lista de proteínas. Se puede hacer fácilmente extrapolando los enfoques de media, máximo y media de los mejores, definidos para combinar las similitudes de una lista de términos GO, descritos previamente.

Así, para cada uno de los objetivos de uso de similitud semántica planteados, según las necesidades del mismo, se define un cálculo diferente de dicha similitud, para este tercer nivel de combinación de las similitudes de una lista de proteínas:

1. **Objetivo 1**: *media de los mejores en 1 sentido, con GO-BP*: se calcula la similitud por pares de proteínas, tomando la máxima similitud de una proteína con todas las del otro conjunto, y entonces se calcula la media para todas las proteínas del primer conjunto. Aunque el estándar es calcular la media de los mejores en los dos sentidos, sólo se calcula en uno porque interesa el parecido de las proteínas predichas con las originales de la ruta, siendo indiferente la comparación contraria.

2. **Objetivo 2**: *media de los mejores en 2 sentidos, con GO-BP*: el cálculo de similitud es igual que en el objetivo 1, pero en los dos sentidos. Es decir, se calcula finalmente la media de similitudes de proteínas del primer conjunto con las del segundo y las del segundo conjunto con las del primero, porque ahora sí interesa conocer cuánto se parecen entre sí las proteínas de los dos conjuntos de predicciones.

3. **Objetivo 3**: *media entre todas, con GO-MF*: en este caso se calcula la similitud por pares de todas las proteínas respecto a todas las del conjunto, y se calcula la media, porque interesa conocer si todas las proteínas se parecen a todas, no sólo a alguna concreta, como sucedería si se calculara la media de los mejores, como en los objetivos previos.

Para los tres objetivos de uso de similitud semántica en esta tesis, en los dos niveles previos de similitud se utiliza la medida de Jiang y Conrath [Jiang and Conrath, 1997], como medida de similitud entre 2 términos GO (primer nivel), y la media de los mejores en los dos sentidos, como combinación de similitud entre N términos GO (segundo nivel).

Capítulo 5

Modelo de Representación del Conocimiento en un BioRepositorio Multi-Relacional para Anotación Funcional

En este capítulo se satisface el objetivo 1 de la tesis (ver capítulo 3), diseñando un modelo de representación del conocimiento de Biología Molecular para anotación funcional, con un enfoque multi-relacional.

En Biología Molecular se han generado y se siguen generando a diario muchos datos, que se almacenan en miles de bases de datos. El conocimiento biológico es muy extenso, de distinto tipo y generado de forma distribuida en múltiples laboratorios. Todo esto da lugar a problemas porque las bases de datos biológicas son muchas, diversas, heterogéneas, variables y complejas en su contenido (desde formatos antiguos hasta modernos, con datos experimentales o predicciones automáticas, etc.) [Quiles, 2005]. Las bases de datos presentan diferentes estructuras internas, por lo que es inviable diseñar un esquema global que lo integre todo, siendo más adecuado realizar esquemas parciales, más acordes a la tarea concreta a resolver.

Por lo tanto, en esta tesis se propone definir una representación del conocimiento biológico relativo a relaciones y características asociadas a los genes y proteínas, restringida a la tarea de anotación funcional, de forma análoga a cuando en otros estudios se limita la representación a imágenes 3D [Quiles, 2005], por ejemplo. Esta representación sigue un enfoque relacional, y su uso está principalmente centrado en la predicción de anotación funcional con aprendizaje automático. Se pretende que el modelo de representación sea flexible, permitiendo anotar distintos productos genéticos, en distintos organismos y cubriendo las distintas relaciones existentes entre los datos.

El capítulo se estructura de la siguiente forma. La primera sección lista los tipos de relaciones biológicas más frecuentes que se pueden usar en anotación funcional. La segunda agrupa dichas relaciones en seis categorías generales. La tercera utiliza esas categorías de relaciones para diseñar el modelo Entidad-Relación (E/R) genérico o abstracto propuesto en esta tesis, para representar el conocimiento biológico que se pueda necesitar en cualquier problema de predicción de anotación funcional. En la última sección, se especifican las bases que habría que seguir para concretar dicho modelo E/R genérico en un modelo E/R específico para un problema concreto de predicción de anotación funcional. En los capítulos 6 y 7 se exponen los modelos E/R específicos para representar los datos de los dos problemas concretos

de anotación funcional elegidos en esta tesis, generados a partir del modelo E/R genérico propuesto en esta tesis en la sección 5.3.

5.1. Tipos de Relaciones Existentes en Biología Molecular

En los problemas de anotación funcional es necesaria una representación relacional, porque, como ya se ha comentado, todos los datos en Biología Molecular presentan relaciones de diversa índole. A continuación se exponen las más frecuentes:

- interacción física entre 2 proteínas u otros compuestos metabólicos,
- enlace de una molécula con un fragmento de ADN,
- enlace con un factor de transcripción,
- asociaciones binarias por co-expresión (a partir del mismo gen, o una proteína induciendo la expresión de la otra),
- genes que se fusionan a lo largo de la evolución,
- genes con perfiles de expresión semejantes (observados en *microarrays*),
- proteínas en una misma ruta metabólica, de señalización o de regulación,
- genes o proteínas con una misma función molecular,
- genes o proteínas con una cierta similitud entre sus secuencias de nucleótidos o aminoácidos, en la misma especie (parálogos) o en diferentes (ortólogos),
- pertenencia a una misma familia de proteínas,
- compartición de un tipo de dominio en la secuencia,
- grupo de proteínas con una estructura tridimensional semejante,
- proteínas localizadas en la misma zona de la célula,
- tejido común donde se expresan varios genes,
- compuestos relacionados en algún artículo de la literatura científica (extraídos a través de técnicas de minería de textos),
- genes o proteínas con expresión fenotípica similar, y
- genes o proteínas implicados en el desarrollo de la misma enfermedad.

Además, el hecho de que existan muchas relaciones y de diferentes tipos en Biología Molecular, como se acaba de mostrar, justifica que en esta tesis se apueste por representar siempre la información utilizando un enfoque relacional. Así se consigue mantener la semántica asociada, sin las simplificaciones que exige una representación proposicional, además de permitir un mejor almacenamiento y gestión de los datos divididos por módulos. El resto de ventajas de la representación relacional se encuentran detalladas en la sección 2.2.2.

5.2. Generalización de Relaciones

En esta sección se agrupan las relaciones del apartado anterior en unas pocas categorías generales, como paso previo que facilite la definición de un modelo de representación del conocimiento genérico y compacto (ver sección 5.3). Dichas categorías abarcan prácticamente todos los tipos de relaciones que se pueden dar en un dominio de Biología Molecular sobre el que predecir anotación funcional. La mayoría de las grandes diferencias biológicas entre dos dominios, en este contexto, se pueden resolver con un cambio pequeño en la representación (poner, quitar, duplicar relaciones), que generalmente va a estar dentro de este pequeño subconjunto de categorías de relaciones:

- `clase_elemento(ID,valorClase)`: asocia la clase o función de un gen, proteína o grupo. Se debe establecer una relación de esta categoría si se quiere realizar aprendizaje supervisado.

- `gen_proteina(IDgen,IDproteina)`: se trata de una relación binaria *uno a uno* o *uno a muchos*, según la especie, que asocia un gen con todos los transcritos o isoformas que se derivan de su expresión.

- `propiedad_X(IDelemento,valorPropiedad)`: para asociar propiedades individuales a un gen, una proteína o un grupo de ellos. En el modelo de datos final, esta relación se puede simplificar y ser representada como un atributo (con valor variable o constante) de la relación correspondiente; a no ser que sea multi-valuado, como suele pasar con las anotaciones biológicas.

- `par_Y([IDpar],IDgen/IDproteina,IDgen/IDproteina)`: relaciones binarias entre genes o proteínas. El identificador del par es opcional, definiéndose sólo cuando se vayan a asociar atributos al par.

- `elemento_en_grupo(IDgen/IDproteina,IDgrupo)`: representación desagregada de relaciones entre más de dos genes o proteínas. Esta categoría de relación permite definir un grupo, mediante la descripción de todos los elementos que pertenecen a él.

- `propiedad_grupo_Y(IDgrupo,valorPropiedad)`: para asociar propiedades a grupos, es decir, relaciones entre más de dos elementos.

La definición de una propiedad, tanto de un elemento como de un grupo (ver categorías `propiedad_X/2` y `propiedad_grupo_Y/2`), permite representar cualquier tipo de anotación. Adicionalmente se podría añadir un argumento más a alguna relación, en el caso de que ésta tenga atributos propios, o incluso quitar el atributo del valor de la propiedad, si ésta es binaria. Por ejemplo, se deberían añadir atributos al calcular valores agregados, que en esta categoría de relaciones se representarían como propiedad de un grupo. En particular, se podría tener la relación `propiedad_grupo_%_GO(IDgrupo,terminoGO,%)`, que represente el porcentaje de elementos del grupo con una anotación concreta GO; o la relación `propiedad_grupo_longitud_ruta(IDgrupo,long)`, que indique el número de elementos de la ruta. Este tipo de agregados se calculan fácilmente mediante el uso de la lógica inductiva, muy relacionada con los lenguajes de representación relacional del conocimiento (ver sección 2.2.3).

De cada una de estas categorías genéricas de relaciones puede haber varias o ninguna, según el escenario concreto a representar. Por ejemplo:

- de la categoría `par_Y/2` puede haber una para relaciones de homología (`par_homologo(IDprot,IDprot)`) y otra para interacciones proteína-proteína (`par_ipp(IDprot,IDprot)`).
- de la categoría `propiedad_X/2` puede haber una para cada tipo de anotación (`dominio_transmembrana/2`, `GO_funcionMolecular/2`, `familia_proteina_Pfam/2`, etc.).
- de la categoría `elemento_en_grupo/2`, si no hay grupos de más de dos elementos, quizá no haya ninguna relación asociada.

Cada una de las relaciones particulares y cada una de las entidades relacionadas, en principio, se correspondería con una tabla en una base de datos relacional, excepto por las simplificaciones concretas aplicadas al convertir el modelo Entidad/Relación en modelo relacional [de Miguel Castaño et al., 1999], que afectan a propiedades y relaciones binarias y multi-valuadas. Una vez definida la representación, en el conjunto de datos habrá varias filas o tuplas de una relación, tantas como existan en el dominio real.

5.3. Modelo Global del BioRepositorio Multi-Relacional para Anotación Funcional

La figura 5.1 muestra la propuesta de definición para esta tesis de modelo Entidad/Relación (modelo E/R) [de Miguel Castaño et al., 1999] genérico para representar toda la información de Biología Molecular asociada a un proceso de anotación funcional, por supuesto, incluyendo todas las relaciones posibles. Cualquier dato necesario para un escenario de predicción de anotación funcional cabe en este esquema general. Se pueden representar tanto propiedades individuales asociadas a un gen o proteína (entidad) concreta, como cualquier tipo de relación entre ellos.

Como parte de la definición de modelo E/R genérico para datos biológicos usados en predicción de anotación funcional, la tabla 5.1 describe cómo cada categoría posible de relación entre elementos biológicos, definidas en la sección anterior, encaja en alguna de las entidades o relaciones definidas en el modelo E/R de la figura 5.1.

Tabla 5.1: Correspondencia entre categorías de relaciones (sección 5.2) con el modelo E/R (figura 5.1).

Categoría de relaciones	Entidad o Relación en Modelo E/R
`clase_elemento`	relación *anotado/a con*
`gen_proteina`	relación *se expresa en*
`propiedad_X`	relación *anotado/a con* o atributo de entidad *Secuencia* o entidad *Grupo*
`par_Y`	relación *pertenece a* con entidad *Grupo Binario*
`elemento_en_grupo`	relación *pertenece a* con entidad *Grupo N-ario*
`propiedad_grupo_Y`	relación *anotado con*

Cabe destacar algunos aspectos de representación sobre el modelo E/R propuesto de la figura 5.1. En primer lugar, la entidad *Grupo Binario* se podría simplificar a una relación de la entidad *Secuencia* consigo misma, pero de esta forma queda representado que una

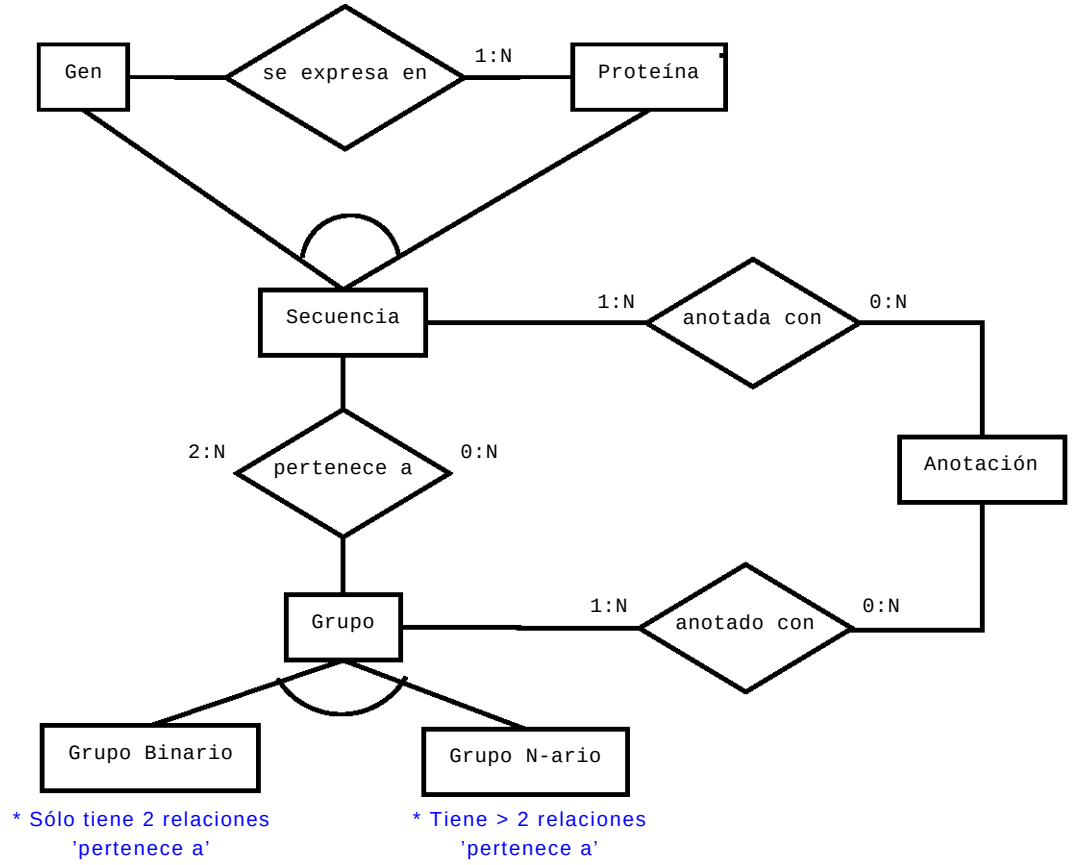


Figura 5.1: Modelo Entidad/Relación BioRepositorio Multi-Relacional.

relación binaria también puede tener anotaciones o propiedades (por ejemplo, una interacción proteína-proteína). Por otro lado, la entidad *Anotación* sirve para representar tanto el objetivo de predicción (es decir, la clase) como una propiedad cualquiera de un gen, proteína o grupo de ellos. Además, la entidad *Anotación* se podría representar directamente como un atributo de las entidades *Secuencia* o *Grupo*, o de la relación *pertenece a*. Dicha simplificación sería válida si el atributo no es multi-valuado, pero con la representación elegida en el modelo E/R propuesto en esta tesis se cubren todos los casos. Respecto a la cadena de nucleótidos o aminoácidos de una *Secuencia*, se puede almacenar la cadena completa en un atributo único. Si se quisiera guardar alguna información más elaborada (como por ejemplo, la proporción de cada aminoácido o de cada par de aminoácidos, o la carga asociada a la secuencia), se tendrían que añadir atributos adicionales o relaciones de tipo *anotada con*, según la complejidad de los datos calculados a partir de la secuencia.

5.4. Aplicación a Bases de Datos Concretas

El modelo E/R de la figura 5.1 se puede denominar modelo genérico o *abstracto*, porque en los modelos E/R particulares para un problema de anotación funcional específico habrá varias entidades y/o relaciones de cada tipo, como se explica en esta sección.

La representación del conocimiento del modelo E/R genérico debe modificarse en función

de lo que se vaya a predecir, adaptándose al problema de anotación concreto. Entonces, para convertir el modelo E/R genérico en uno específico hay que determinar qué relaciones y entidades incluir y cuáles no, así como decidir cuántas de cada tipo; como se hace en las secciones 6.2.2 y 7.2.2 para cada uno de los dos problemas de anotación funcional elegidos en esta tesis.

La mayoría de las grandes diferencias biológicas de un escenario de anotación funcional a otro se pueden resolver con pequeños cambios en la representación, limitados a un subconjunto de casos, descritos a continuación.

En primer lugar, hay que diferenciar si se quiere predecir para un gen, una proteína o un grupo de elementos (de dos o más genes o proteínas). En función de ello, se tiene que definir la relación objetivo `clase_elemento`, de cuya categoría siempre va a existir una única copia. Si el objetivo es anotar un grupo de elementos, va a existir una entidad para el *Grupo* concreto, y una relación *pertenece a*, que no existirían si se predice para genes o proteínas aislados.

En segundo lugar, hay que definir si se predice anotación funcional en un organismo con isoformas o no, es decir, si un gen produce estrictamente una proteína (como suele suceder en especies simples como *E.coli* y *S.cerevisiae*) o si puede expresarse en más de una proteína por procesos de ensamblaje alternativo (típico de especies más complejas, como los humanos). Esto influye en la cardinalidad de la relación *se expresa en* (de la categoría de relación `gen_proteína`), de la cual siempre hay una copia, dependiendo de si hay isoformas (cardinalidad *un gen a muchas proteínas*) o no (cardinalidad *uno a uno*). En ocasiones, a pesar de la existencia de isoformas, conviene obviarlas y trabajar con una forma principal, lo cual se puede representar con una relación adicional *isoforma principal* de la categoría `gen_proteina`, o directamente simplificar la relación *se expresa en* a una cardinalidad *uno a uno*. Si la relación *se expresa en* tiene cardinalidad *uno a uno*, a veces las propiedades individuales de genes y proteínas, e incluso la clase, se pueden representar como atributos de una única entidad, para no complicar innecesariamente el proceso de aprendizaje.

En tercer y último lugar, el número del resto de relaciones y entidades del modelo global depende del conjunto de datos de entrada concreto y del problema a representar. Las propiedades y relaciones son diferentes de una especie a otra, incluso entre todos los genes y proteínas de una misma especie puede haber una cantidad de información distinta, o puede desearse seleccionar un subconjunto diferente entre la información disponible, acorde a las restricciones del problema.

En conclusión, esta representación en un BioRepositorio Multi-Relacional es flexible, permitiendo anotar un gen, una proteína o un grupo; en una especie simple o compleja; y cubriendo distintas propiedades o relaciones en los datos.

Capítulo 6

Predicción de Asociaciones Funcionales entre Pares de Proteínas en *E.coli*

El objetivo de este capítulo es afrontar uno de los dos problemas de anotación funcional seleccionados en esta tesis: predecir *A*sociaciones *F*uncionales entre *P*ares de *P*roteínas (AFPP) en *E.coli* desde una perspectiva unificadora.

En este capítulo, primero se describe el contexto del problema, se detallan los datos a utilizar y su representación (aplicando el modelo E/R genérico descrito en el capítulo 5) y se expone el método de aprendizaje seleccionado. A continuación se presentan los resultados de rendimiento de la clasificación y la comparación con otros métodos de predicción de asociaciones funcionales (frente a algoritmos de aprendizaje automático alternativos, frente a los métodos computacionales individuales que unifica el método propuesto en esta tesis, y frente a una base de datos muy extendida de recopilación de estas asociaciones funcionales). También se expone la utilidad de la propuesta presentada para filtrar interacciones experimentales a gran escala, y el servidor de predicciones EcID donde se almacenan las predicciones que se consiguen en esta tesis.

6.1. Definición del Problema

Como se define en el capítulo 2, de forma general, se podría decir que un par de proteínas asociadas funcionalmente significa que están relacionadas por la función que realizan. Durante todo el capítulo, las interacciones proteína-proteína, con contacto físico entre las proteínas, también se consideran incluidas dentro de las asociaciones funcionales.

Las funciones celulares son casi siempre el resultado de la acción coordinada de varias proteínas [Bader et al., 2008]. Por lo que frecuentemente se infiere la función de una proteína desconocida mediante la identificación de las proteínas con las que interacciona o se asocia funcionalmente [Causier, 2004]. Esta perspectiva de Biología de Sistemas se emplea cada vez más para la anotación funcional, porque los progresos tecnológicos permiten actualmente construir redes de interacción a gran escala [von Mering et al., 2005].

Ante el elevado coste experimental de detección de interacciones (ver sección B.4), los métodos computacionales de predicción permiten priorizar las más probables.

Se han desarrollado varios métodos computacionales para predecir asociaciones funcionales en genomas/proteomas completos (ver sección B.4). Tres de los métodos frecuentemente

probados e implementados (PP, *Phylogenetic Profiles* [Pellegrini et al., 1999]; GC, *Gene Context*) [Dandekar et al., 1998]; GF *Gene Fusion*) [Enright et al., 1999; Marcotte et al., 1999]) tienen en común el uso de información evolutiva. Posteriormente han aparecido variantes para ellos [Bowers et al., 2004; Morett et al., 2003; Wu et al., 2003]. Otros métodos de predicción comparten el uso de alineamientos múltiples de secuencia y principios de co-evolución (I2H, *In silico Two-Hybrid* [Pazos and Valencia, 2002]; MT *MirrorTree*) [Pazos and Valencia, 2001]). Cuando se aplican sobre grandes colecciones de datos producen una cantidad considerable de falsos positivos, probablemente relacionados con tendencias evolutivas adicionales que diluyen la señal de interacción o asociación funcional, aunque se continúa investigando para discernir mejor la información co-evolutiva, dando lugar a métodos más fiables en este aspecto [Pazos et al., 2005; Sato et al., 2005; Juan et al., 2008; Pazos and Valencia, 2008; Herman et al., 2011].

Ahora bien, los resultados derivados del estudio de las proteínas y de sus interacciones y asociaciones funcionales (ya sea a nivel experimental o computacional) no están unificados; sino que, por el contrario, están repartidos en múltiples repositorios de información. Por lo cual, es conveniente integrar las fuentes de conocimiento sobre interacciones y asociaciones funcionales, ya que son heterogéneas en enfoque, cobertura y fiabilidad. Si se usan de forma aislada, se limita el conocimiento que se puede extraer. Cada fuente puede aportar su parte relevante de información, siendo cada una más adecuada para un fragmento del espacio de interacciones donde las características encajen con sus hipótesis. Por lo que para tener una visión global de la red de interacciones entre las proteínas de un organismo se necesita tener una combinación de todo ello.

Algunos estudios previos han utilizado la integración de datos provenientes de distintas fuentes (principalmente experimentales) para mejorar la predicción de asociación funcional entre proteínas en *Saccharomyces cerevisiae* [Qi et al., 2006; Lu et al., 2005]. Dichos métodos confían en la gran cantidad de datos experimentales disponibles para este organismo, y se dedican fundamentalmente a asignar valores de fiabilidad a interacciones experimentales.

En este trabajo se propone también mejorar la predicción de asociaciones funcionales, pero no mezclando datos experimentales, sino basándose en la combinación de métodos computacionales desarrollados independientemente. Además, se pretende descubrir nuevas asociaciones funcionales entre pares de proteínas, no restringidas a interacciones físicas entre proteínas. Por lo tanto, este nuevo enfoque se diferencia de los estudios previos tanto en la información de entrada, como en su consiguiente aplicabilidad. En este caso, se predicen asociaciones funcionales sobre el proteoma de un organismo procariota concreto: *Escherichia Coli (E.coli)*. Se ha elegido por estar mejor caracterizado que otros a nivel molecular y porque las bacterias son un buen conjunto de prueba, por la cantidad de genomas secuenciados y por la arquitectura simple de sus células (sin núcleo definido, sin orgánulos con membrana, ADN concentrado en un único cromosoma, con una sola proteína por gen, etc.). Tiene un total de 4.339 proteínas conocidas, entre las que determinar si existen interacciones o asociaciones funcionales por pares, aportando una probabilidad de confianza de la predicción para cada par.

6.2. Diseño/Materiales y Métodos

En esta sección se describe cuáles son las fuentes de datos originales, cómo se representa y agrupa dicha información para aplicar un algoritmo de aprendizaje automático, y cómo obtener un sistema que integre cinco métodos de predicción de asociaciones funcionales entre pares de proteínas en *E.coli* en uno solo unificado. Cada uno de estos métodos está basado en diferentes enfoques sobre la posible causa que provoca que un par de proteínas interaccione o se asocie

funcionalmente a otra [Valencia and Pazos, 2002]. Cada método predice de forma óptima sobre un subconjunto concreto de todos los posibles pares, donde se cumplen todas sus premisas. Sin embargo, ningún enfoque individual es el más apropiado para todos y cada uno de los pares de proteínas posibles, por lo que resulta adecuado combinarlos.

6.2.1. Fuentes de Datos

Antes de comenzar la descripción, hay que destacar que, excepto los rankings (la última fuente de datos descrita en esta sección), las otras tres fuentes de datos han sido recopiladas o implementadas por investigadores del programa de Biología Estructural y Biocomputación del Centro Nacional de Investigaciones Oncológicas. Para conocer más detalles de los descritos aquí sobre la implementación del cálculo de ortólogos y de cada uno de los cinco métodos computacionales, cuyas salidas se usan en este trabajo, consultar las publicaciones relacionadas [García-Jiménez et al., 2010a; Leon et al., 2009].

Interacciones y Asociaciones Funcionales entre Pares de Proteínas

Cada una de las bases de datos utilizadas (cuantificadas y referenciadas en la Tabla 6.1), contiene información relativa a una evidencia que indica la posibilidad de que exista una interacción o asociación funcional entre pares de proteínas. Estas bases de datos se pueden agrupar en distintas categorías:

- *Datos experimentales*: pares de proteínas que se ha comprobado que interaccionan físicamente, mediante una experimentación a pequeña escala en laboratorio. Fuentes: DIP, BIND e IntAct.

- *Complejos*: se consideran interacciones entre proteínas que pertenecen a un mismo complejo molecular, por estar unidas físicamente entre sí en un grupo. Se establece un enlace funcional por cada par de proteínas que forman parte del mismo complejo. Se usan dos fuentes de datos de complejos. El primer grupo está basado en verificaciones manuales de la literatura científica, que representa un conjunto de alta calidad de complejos muy conocidos, con muy elevada probabilidad de certeza (fuente: EcoCyc complejos). El segundo procede de experimentación por co-immunoprecipitación a gran escala, con menos fiabilidad que el primero, en concreto, con sensibilidad más alta y especificidad más baja (fuente: conjunto de Butland).

- *Regulación*: bases de datos que tienen en cuenta procesos de regulación génica. Se establecen enlaces funcionales entre cada regulador de la transcripción y sus correspondientes genes regulados. Es decir, se incluyen pares de proteínas en los que una de las proteínas cataliza una reacción para que se exprese la otra del par. Fuentes: EcoCyc regulados.

- *Co-regulación*: basándose en el mismo tipo de información que el conjunto previo, se establecen relaciones entre las proteínas que son reguladas por el mismo regulador. Es decir, se incluyen pares de proteínas (asociadas a un determinado gen) que se expresan a la vez. Fuentes: EcoCyc co-regulados.

- *Rutas metabólicas*: pares de proteínas involucradas en la misma ruta metabólica, basándose en que las proteínas que participan en un mismo flujo de reacciones pueden interaccionar entre ellas. Se considera que todas las proteínas asignadas a la misma

ruta están asociadas funcionalmente por pares, aunque no implique necesariamente una interacción física directa entre ellas. Fuentes: KEGG y EcoCyc asociaciones funcionales.

- *Minería de textos*: bases de datos que almacenan asociaciones funcionales extraídas directamente de la literatura científica usando técnicas de minería de textos. Específicamente, los pares se definen cuando hay una mención de ambas proteínas en la misma frase de los resúmenes de un artículo en PubMed, lo cual indica que puede existir alguna asociación entre ellas. Fuente: iHOP.

Tabla 6.1: Bases de datos de interacciones y asociaciones funcionales entre pares de proteínas usadas. Referencia y número de pares de proteínas extraídos de cada una. KEGG en su versión 14 y EcoCyc en su versión 15. Todas ellas conforman las instancias de la clase positiva.

Base de datos	**Referencia**	**Nº Pares**
DIP	[Salwinski et al., 2004]	401
BIND	[Alfarano et al., 2005]	58
IntAct	[Hermjakob et al., 2004b]	2.684
EcoCyc complejos	[Keseler et al., 2005]	950
conjunto de Butland	[Butland et al., 2005]	4.745
EcoCyc regulados	[Keseler et al., 2005]	1.686
EcoCyc co-regulados	[Keseler et al., 2005]	58.275
KEGG	[Kanehisa et al., 2006]	20.860
EcoCyc asociaciones funcionales	[Keseler et al., 2005]	3.446
iHOP	[Hoffmann and Valencia, 2004]	6.686
	Total (sin solapamientos)	**89.401**

Se intenta capturar la naturaleza compleja del dominio utilizando esta gran variedad de bases de datos externas. Unas fuentes proporcionan información más fiable que otras, apareciendo en el listado superior de mayor a menor fiabilidad teórica. Esto se debe, por una parte, a la evidencia en la que se fijan, que biológicamente puede guardar más o menos relación con las interacciones o asociaciones funcionales entre proteínas; y por otra, a la forma de añadir contenido a la base de datos, ya que en algunos casos se permite que cualquiera que disponga de alguna información la introduzca, mientras que en otros está bastante más restringido a expertos. No obstante, hay que aclarar que se trata de una fiabilidad teórica, por el tipo de evidencia usada, pero no por las bases de datos concretas disponibles para cada evidencia, entre las que pueden variar los niveles de fiabilidad. Por ejemplo, el conjunto de Butland [Butland et al., 2005] pertenece a la categoría de complejos, por ser la evidencia biológica en la que se basa; dicha categoría es la segunda con mayor fiabilidad teórica, y sin embargo se sabe que el conjunto de datos de Butland es de mala calidad en comparación a otros conjuntos de complejos.

Métodos de Predicción Computacional

Otra fuente de datos es el grado de asociación (salida o puntuación; del inglés, *score*) proporcionado por varios métodos computacionales de predicción de interacción o asociación funcional entre proteínas [Valencia and Pazos, 2002]. Los fundamentos subyacentes de cada método computacional de predicción son variados, y se suelen dividir en cinco grupos, de los que se ha tomado un representante de cada una de ellos.

Uno está basado en la similitud de perfiles filogenéticos (PP, *Phylogenetic Profiles*), examinando la presencia o ausencia de genes en especies relacionadas [Pellegrini et al., 1999]. Otro método se basa en la conservación de genes adyacentes en diferentes especies (GC, *Gene Context*) [Dandekar et al., 1998]. Un tercer procedimiento se fija en los eventos de fusión de genes, buscando los mismos dominios de proteínas en distintos genomas (GF, *Gene Fusion*) [Enright et al., 1999; Marcotte et al., 1999]. Los dos métodos restantes están basados en la coevolución de proteínas, estudiando la similitud de sus árboles filogenéticos (MT, *MirrorTree*) [Pazos and Valencia, 2001], o cuantificando el grado de co-variación entre los pares de residuos de las proteínas (mutaciones correlacionadas) (I2H, *In silico Two-Hybrid*) [Pazos and Valencia, 2002]. Ver la sección B.4 para una descripción más detallada de los métodos de predicción originales.

Cabe destacar que ninguno de estos métodos computacionales distingue entre una predicción entre pares de proteínas asociadas funcionalmente o más restringida a una interacción proteína-proteína, porque las evidencias que usan no son suficientes para ello.

Características Individuales de Proteínas

Una tercera fuente de datos son un par de características básicas de una proteína. Por un lado, la longitud de la secuencia de aminoácidos y, por otro, el nº de secuencias ortólogas a la proteína dada. Se han seleccionado ambas características por estar alta e intrínsecamente relacionadas con el rendimiento de algunos de los métodos computacionales descritos en la sección anterior.

Ranking de Predicciones Centrado en la Proteína

Esta cuarta fuente de datos se compone de valores derivados a partir del grado de asociación (o puntuación) de los métodos de predicción computacional, descritos previamente.

Varios de estos métodos computacionales tienen algunos sesgos, produciendo muchas predicciones con puntuaciones bajas. Por ejemplo, algunos perfiles filogenéticos son más usuales y, por lo tanto, PP sobre-predice asociaciones entre las proteínas correspondientes. Así, para muchas proteínas, los métodos individuales predicen más interacciones y asociaciones de las razonables. Entonces, aunque se considere que cada par asociado es independiente de otro, no lo es realmente desde una semántica biológica.

Para hacer frente a esta situación, se definen características para identificar el subconjunto de proteínas que es más probable que interaccione o se asocie funcionalmente con una dada. Para cada proteína, y en función de la puntuación proporcionada por cada uno de los métodos, se crea la lista ordenada de potenciales 'compañeras de par' y en ella se calcula la posición que ocupa la otra proteína del par. Entonces, la nueva fuente de datos está compuesta por la posición para cada una de las 2 proteínas del par en la lista de predicciones ordenada, construida para el otro elemento del par. Esto se repite para cada uno de los 5 métodos, generando diez nuevos valores por par.

6.2.2. Representación del Conocimiento

Modelo E/R para Predicción de Asociaciones Funcionales entre Pares de Proteínas en *E.coli*

Analizando este problema de anotación funcional desde el punto de vista del AA, la unidad a anotar está formada por un grupo de 2 elementos (proteínas) al que asignar una clase binaria

(sí/no interaccionan o se asocian funcionalmente). Se tienen atributos individuales (longitud y nº de ortólogos), atributos asociados al grupo (grado de asociación según cada método) y atributos relacionales (posición en el ranking de una proteína frente a la otra). Todos estos atributos tienen valores numéricos.

Por lo tanto, teniendo en cuenta este resumen de información, las fuentes de datos previamente descritas en la sección 6.2.1, y las instrucciones de la sección 5.4 para convertir el modelo genérico E/R del BioRepositorio multi-relacional (ver 5.1) en uno específico, se obtiene el modelo E/R de la figura 6.1 para este problema concreto de anotación funcional.

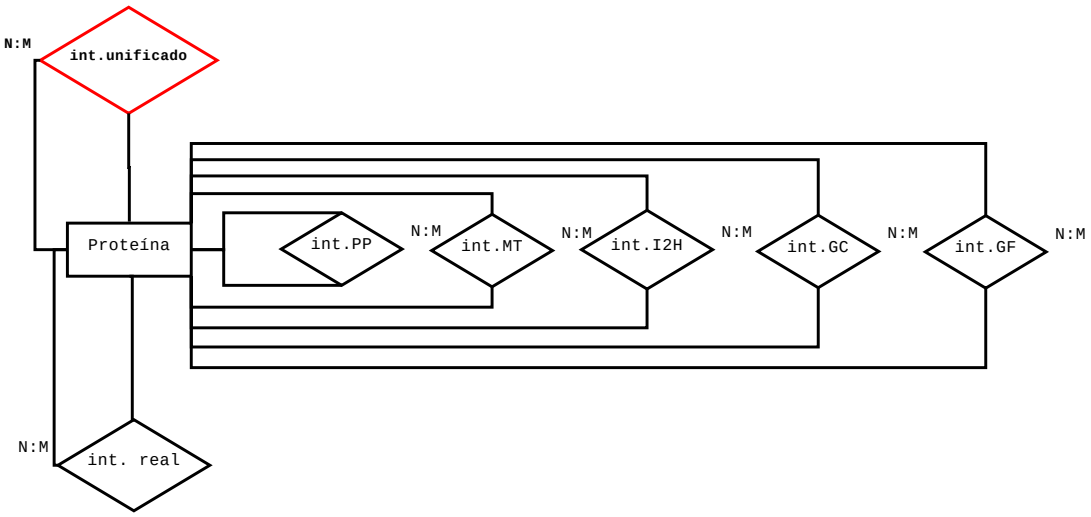


Figura 6.1: Modelo Entidad/Relación para predicción de asociaciones funcionales entre pares de proteínas en *E.coli*. En rojo, el objetivo de predicción.

El modelo global aplicado al problema de predicción de asociaciones funcionales entre pares de proteínas en *E.coli* se reduce a 1 entidad y 7 relaciones binarias, como muestra la figura 6.1.

Sólo se tiene la entidad *Proteína* porque la relación *se expresa en* (o `gen_proteina`) (ver 5.1) es de cardinalidad *uno a uno* en *E.coli*, y no se usan datos del gen, por lo que ni siquiera hace falta representar su entidad asociada *Gen*.

Las 7 relaciones binarias de la figura 6.1 se corresponden con la simplificación de 7 relaciones de tipo *pertenece a* entre la entidad *Proteína* y 7 entidades diferentes de tipo *Grupo Binario*, una para cada método de predicción (5 predictores individuales, 1 predictor unificado, y 1 relación para las interacciones o asociaciones funcionales reales). La relación *int.Unificado* representa el objetivo de predicción, por ello se señala en rojo en la figura 6.1.

El *score* o grado de asociación según cada método es un atributo correspondiente a cada uno de los grupos binarios (simplificados en una relación binaria cada uno). Por otra parte, la posición en el ranking de una proteína frente a otra es un atributo asociado directamente a la relación, y derivado a partir de todas las asociaciones funcionales de una proteína definidas por un método concreto. No es necesario definir entidades de tipo *Anotación*, porque las propiedades de las entidades *Proteína* y *Grupo Binario* toman un valor único y diferente para cada ejemplo, pudiendo representarse de forma sencilla como atributos.

Transformación a Representación Proposicional

Partiendo del modelo de representación Multi-Relacional de la figura 6.1 y de las correspondencias con las categorías de relaciones biológicas descritas en la sección 5.4, en este problema hay pocos tipos diferentes de relaciones entre elementos individuales (proteínas), y la mayoría de propiedades están asociadas al grupo o par de proteínas. Además, todos los atributos son numéricos. Por lo tanto, el conocimiento se puede compactar en una representación proposicional o atributo-valor, con una tupla por cada par de proteínas, con menos de 20 atributos. Mantener la representación relacional daría lugar a un modelo de datos innecesariamente extenso y complejo, que probablemente no aportaría ventajas durante el aprendizaje. No obstante, sí se evitarían muchos valores desconocidos y la repetición de los valores de longitud y nº de ortólogos para cada secuencia de proteína. Así, en este problema se va a usar Aprendizaje Automático Proposicional.

Por lo tanto, cada par de proteínas (instancia), que puede estar asociado funcionalmente o no (clase positiva o negativa), se representa proposicionalmente mediante 19 atributos numéricos que se pueden clasificar en 3 grupos:

- 5 grados de asociación, proporcionados por los 5 métodos individuales de predicción, descritos en la sección 6.2.1.
- 4 (2 por proteína) propiedades de las secuencias de proteínas: longitud y nº de ortólogos.
- 10 posiciones en ranking (una por método y por proteína) descritos en la sección 6.2.1.

En los dos últimos grupos, que existe un valor para cada proteína del par, se ordenan como valor mínimo y máximo de cada par, para mantener una coherencia en todas las instancias. Los rangos de valores así como otros indicadores estadísticos se muestran en la Tabla 6.2.

Los valores desconocidos para los 19 atributos se consideran como valores 'indeterminados' y no se reemplazan con indicadores específicos, porque en el contexto de este dominio, la ausencia de información presenta diferencias con la existencia de ruido en los datos. Así, no tener el valor de salida de un método individual de predicción implica que dicho método no se puede aplicar a un determinado par de proteínas, porque no se cumplen las condiciones necesarias (ej: no hay un número mínimo de ortólogos que permita analizar el perfil evolutivo) o porque no se verifica la evidencia en la que está basado el método (ej: no se produce un evento de fusión de genes). Aunque usar indicadores para reemplazar los valores desconocidos produce resultados similares, se ha preferido la alternativa de no reemplazarlos, ya que se ajusta mejor a la semántica del dominio, tiene un ratio de recuperación positiva ligeramente mejor y gestiona directamente los valores desconocidos.

6.2.3. Construcción de Conjuntos de Datos

De las fuentes de datos expuestas en la sección 6.2.1, según describe el apartado previo, todas se utilizan como atributos de entrada excepto las fuentes de interacción y asociación funcional por pares de proteínas, que son las que se van a utilizar para definir los conjuntos de ejemplos positivos y negativos, necesarios para aplicar AA.

Clase Positiva

Cuando se hacen predicciones de interacción y asociaciones funcionales entre pares de proteínas, una cuestión particularmente controvertida es la definición de qué se considera una

Tabla 6.2: Estadísticas de los atributos para la predicción de asociación funcional entre pares de proteínas. Total de instancias: 2665180. La media y la desviación típica se calculan sin tener en cuenta las instancias con valores desconocidos.

Atributo	**Valores conocidos / desconocidos** (total y porcentajes)	**Rango de valores**	**Media**	**Desviación típica**
I2H	1054149 / 1611031 (39,55 % / 60,45 %)	[0-35,349]	0,771	0,585
MT	1054149 / 1611031 (39,55 % / 60,45 %)	[0-0,991]	0,615	0,194
PP	2591226 / 73954 (97,23 % / 2,77 %)	[0,088-1]	0,646	0,168
GC	11690 / 2653490 (0,44 % / 99,56 %)	[1-145]	3,651	8,093
GF	668 / 2664512 (0,03 % / 99,97 %)	[1-157]	7,266	18,837
longitud sec. mín.	2665180 / 0 (100,00 % / 0,00 %)	[24-1538]	248,032	118,405
longitud sec. máx.	2665180 / 0 (100,00 % / 0,00 %)	[46-2003]	459,751	215,676
nº ortólogos mín.	1594687 / 1070493 (59,83 % / 40,17 %)	[16-113]	332,729	16,646
nº ortólogos máx.	2523554 / 141626 (94,69 % / 5,31 %)	[16-113]	55,957	26,614
pos. rank. I2H mín.	1054149 / 1611031 (39,55 % / 60,45 %)	[1-2168]	615,329	465,049
pos. rank. I2H máx.	1054149 / 1611031 (39,55 % / 60,45 %)	[1-2181]	998,099	545,625
pos. rank. MT mín.	1054149 / 1611031 (39,55 % / 60,45 %)	[1-2137]	543,095	396,529
pos. rank. MT máx.	1054149 / 1611031 (39,55 % / 60,45 %)	[1-2181]	1039,962	535,107
pos. rank. PP mín.	2591226 / 73954 (97,23 % / 2,77 %)	[1-2946]	1125,718	769,455
pos. rank. PP máx.	2591226 / 73954 (97,23 % / 2,77 %)	[1-2946]	1877,786	798,469
pos. rank. GC mín.	11690 / 2653490 (0,44 % / 99,56 %)	[1-20]	4,275	2,895
pos. rank. GC máx.	11690 / 2653490 (0,44 % / 99,56 %)	[1-23]	7,588	4,563
pos. rank. GF mín.	668 / 2664512 (0,03 % / 99,97 %)	[1-11]	1,121	0,549
pos. rank. GF máx.	668 / 2664512 (0,03 % / 99,97 %)	[1-25]	2,867	3,104

asociación funcional. En este trabajo se ha elegido una definición inclusiva de asociación funcional, que es consistente con los distintos métodos de predicción que se emplean como atributos. Mientras que algunos de estos métodos se centran más en detectar interacciones físicas (GF, I2H o MT), otros son más adecuados para predecir rutas bioquímicas (GC) o tienen un alcance menos definido aún.

Así, la clase positiva está formada por los pares de proteínas de *E.coli* que tienen una asociación funcional según la mencionada definición inclusiva; es decir, todos los pares que aparecen en alguna de las bases de datos externas descritas en el primer apartado de la sección 6.2.1. Se toma esta decisión porque no existen ejemplos positivos suficientes para construir clasificadores independientes por evidencia o base de datos (ver cantidades de ejemplos en la tabla 6.1). En total, sin considerar los homodímeros (es decir, los pares formados por dos proteínas iguales), el conjunto contiene 89.401 pares de proteínas diferentes.

Con esta definición de instancias positivas se intenta tener una representación extensa de las asociaciones funcionales entre proteínas, al incluir bases de datos variadas. Como consecuencia de la distinta cantidad de información disponible para cada tipo de asociación funcional, los resultados van a revelar más información sobre las capacidades de predicción de los principales contribuidores de instancias a este conjunto de positivos, es decir, los genes co-regulados y las asociaciones metabólicas.

Clase Negativa

Los pares de proteínas para la clase negativa se obtienen de lo que se podría denominar la aplicación del ‘supuesto del mundo cerrado’. En este dominio significaría que cualquier par

de proteínas que no se sabe que está asociado funcionalmente (es decir, que dicho par no es una instancia positiva) es considerado un par que no se asocia funcionalmente entre sí (una instancia negativa). De esta forma, la cantidad de instancias negativas es muy elevada (más de un 99 % del total), debido a la explosión combinatoria a la que dan lugar las 4.339 proteínas de *E.coli*, con un total de 9.411.291 pares posibles. Por ello, se aplica una serie de filtros que permiten reducir el conjunto de instancias negativas basándose en ciertos criterios más informativos que la simple aleatoriedad. Por ejemplo, sólo se escogen instancias que tienen sus dos proteínas en algún par del conjunto de positivos. Así, se intenta reducir la incertidumbre en la información negativa, por considerar sólo proteínas sobre las que se tiene alguna información acerca de su función. Es decir, que al menos se ha trabajado con ellas, y si no se ha encontrado una interacción o asociación funcional, será más probable que no exista, frente a un par de proteínas sobre las que no se ha investigado nada. Del conjunto de pares resultante, se eliminan los homodímeros y también aquellos pares para los que ningún método de predicción individual genera un valor de salida, porque no son relevantes para el aprendizaje. Este proceso arroja 2.575.779 pares de proteínas negativos. Es importante tener presente que este conjunto todavía podría contener algunas asociaciones funcionales no definidas (es decir, falsos negativos).

Conjuntos de Entrenamiento y Test

La construcción de los conjuntos para el aprendizaje debe ocuparse del problema del desbalanceo entre clases, ya que la clase negativa filtrada aún constituye casi el 97 % de todas las instancias. Por lo tanto, los conjuntos de entrenamiento y test se construyen con un 20 % de instancias positivas y un 80 % de negativas, estableciendo así un compromiso entre representar la distribución subyacente y alcanzar un mayor balance de ambas clases para no afectar al rendimiento de la clasificación. Aunque esta distribución entre clases se considera un buen equilibrio, previamente se han probado otras distribuciones que van del 20 % al 50 % de instancias positivas, no presentando ninguna de ellas resultados destacables.

El conjunto de entrenamiento se compone de dos tercios de ejemplos positivos (completando el 20 % explicado previamente), quedándose el conjunto de test con el tercio restante de pares de proteínas positivos. Los conjuntos de entrenamiento y test se completan con instancias de la clase negativa, cogiendo aleatoriamente exactamente 4 veces el número de instancias positivas, para alcanzar el 80 % de instancias negativas en cada conjunto. Así, se emplean todas las instancias positivas disponibles (asociaciones funcionales conocidas), bien en el conjunto de entrenamiento o bien en el de test. Por el contrario, se descartan muchas instancias negativas. De esta forma, según los criterios mencionados previamente, el conjunto de test tiene la mitad del tamaño del conjunto de entrenamiento.

El total de instancias disponibles es de 2.665.180 asociaciones funcionales (89.401 pares positivos y 2.575.779 negativos), que se reducen a 264.752 (16.566 positivas y 248.186 negativas) al aplicar un filtro, tanto a la clase positiva como a la negativa, asociado a los diez atributos de ranking. Este filtro consiste en eliminar las instancias donde ninguna de las posiciones de ranking está entre las 100 primeras, en los rankings de PP, MT y I2H. Este paso reduce el ruido procedente de los pares menos puntuados que no deberían ser predichos por los métodos de entrada. Si el par tiene un score de los métodos GC o GF, la instancia se mantiene, ya que la cantidad de asociaciones funcionales para una proteína dada es mucho menor según estos dos métodos.

Resumiendo, tras aplicar la distribución del 20 % para positivos y el 80 % para negativos, los conjuntos finales se componen de 11.044 positivos y 44.176 negativos en el conjunto de entrenamiento, y 5.522 positivos y 22.088 negativos en el de test.

6.2.4. Complejidad del Dominio

Conviene tener presente algunos problemas implícitos a la naturaleza biológica de los datos, los cuales establecen un alto nivel de complejidad que dificulta la construcción de un predictor de asociación funcional entre proteínas.

- **Incertidumbre intrínseca:** Como en la mayoría de los dominios de Biología Molecular, en este dominio no se puede asegurar que los datos de interacciones y asociaciones funcionales que se utilizan son completamente correctos (tanto en el conjunto de entrenamiento como en el de test), porque no existen experimentos específicos en laboratorio que verifiquen que todas las interacciones y asociaciones funcionales se producen en realidad. Esto añade un cierto grado de imprecisión a los conjuntos de entrada empleados, donde la distribución de clases que se pretende aprender no es exacta. Por ejemplo, se sabe que en los conjuntos de datos hay falsos negativos (correspondientes a asociaciones funcionales que aún no se han descubierto), que pueden ser muchos en comparación a la cantidad de verdaderos positivos, pero pocos en comparación al conjunto total de asociaciones funcionales que existen en el interactoma del organismo considerado.

- **Desbalanceo extremo entre la cantidad de positivos y negativos:** Como muestra de ello se presentan algunos valores numéricos. Si se tienen en cuenta todos los posibles pares resultantes de la combinación de todas las proteínas de *E.coli* (4.339 proteínas), se alcanza un valor de 9.411.291 posibles ejemplos sobre los que evaluar la existencia de asociación funcional o no, de los que menos del 1 % corresponde a instancias positivas. Aplicando algunos filtros para mejorar la utilidad y fiabilidad de los ejemplos negativos considerados, la cantidad de positivos alcanza tan solo algo más del 6 %.

- **Gran porcentaje de valores desconocidos en los atributos:** Dadas las limitaciones de aplicación de cada uno de los métodos computacionales de predicción (como por ejemplo, la necesidad de un cierto número de secuencias ortólogas en el alineamiento múltiple, la ocurrencia de un evento poco frecuente, o la exigencia de la secuenciación de la proteína completa, entre otros) es lógico que existan muchos pares de proteínas sobre los que no se puede aplicar un método, y por tanto no se tiene un valor asociado al atributo correspondiente a dicho método. Como muestra, se puede mencionar que en los métodos de conservación de genes adyacentes (GC) y de eventos de fusión de genes (GF), los scores son desconocidos en más del 99 % de los casos.

Estos tres aspectos comentados dan una idea de las características del conjunto de datos que se manejan.

6.2.5. Algoritmos de Aprendizaje

Una vez definidas las instancias a utilizar en las fases de entrenamiento y test, el siguiente paso es determinar el algoritmo de aprendizaje con el que construir el clasificador. Entre las diferentes posibilidades se elige AODE (Averaged One-Dependence Estimators) [Webb et al., 2005], descrito en la sección 2.1.2.

AODE es un algoritmo bayesiano (basado en probabilidades condicionadas), que conserva la simplicidad, eficacia y eficiencia de *Naive Bayes* [John and Langley, 1995], evitando los inconvenientes que ocasiona la exigencia de asumir la independencia total de los atributos. Este nuevo enfoque mejora la precisión de *Naive Bayes*, sin incrementar de forma considerable

los costes computacionales; característica deseable en grandes conjuntos de datos, como los que se manejan en este problema. Consultar la sección 2.1.2 para una descripción detallada del algoritmo.

AODE requiere valores nominales en todos los atributos, que originalmente son continuos. Por lo tanto, se realiza una discretización, usando el criterio de igual frecuencia, con un mínimo de 50 instancias por partición. Este criterio se selecciona porque es el que mejores resultados ha proporcionado empíricamente en este dominio, frente a otros esquemas posibles como el de tamaño fijo de la partición.

Además, cabe destacar que AODE maneja los valores desconocidos de atributos teniendo en cuenta sólo los conocidos para esa instancia, calculando el producto de probabilidades sólo de los atributos existentes. Dada la elevada cantidad de valores desconocidos en este dominio, esta idea para gestionarlos es adecuada, porque no rellena los desconocidos con la media o la mediana del atributo en todo el conjunto, ni obvia por completo la instancia, como hacen otros algoritmos. Rellenar los valores desconocidos con la media o el valor mayoritario (como ocurre en *Naive Bayes* y *BayesNet* [Friedman et al., 1997; Bouckaert, 2004]) no refleja la semántica de los datos de este dominio, ya que atributos con un valor desconocido podría implicar que no existen (que es diferente a que se haya extraviado el valor por ruido en los datos). Igualmente, ignorar las instancias con valores desconocidos no sería viable en este dominio, dado que casi todas las instancias tienen algún valor desconocido (sólo hay 82 instancias completas de un total de 2.665.180). Esta circunstancia se debe a que los métodos computacionales individuales que se quieren unificar sólo dan un resultado en condiciones restringidas (ver descripción de los métodos [Valencia and Pazos, 2002]). Otro enfoque para mantener la semántica biológica de los valores desconocidos es la definición de un operador particular, que diferencie esos valores y los gestione de forma específica, como permite la técnica de Programación Genética, desarrollada en el capítulo 8, en la sección 8.1.

Para seleccionar el algoritmo de aprendizaje, se han aplicado diferentes tipos de algoritmos de clasificación, intentando elegir una versión reciente o mejorada de cada uno, o aquella que aproveche alguna característica específica adaptada a los datos de este dominio. Así, se ha hecho uso de regresión lineal [Mitchell, 1997], como técnica básica para combinar ponderadamente atributos numéricos aproximando una función lineal; árboles de decisión, en su nueva versión ADTree [Freund and Mason, 1999]; razonamiento basado en casos, en su versión Kstar [Cleary and Trigg, 1995], que permite seleccionar la manera en la que debe manejar los valores desconocidos de atributos; redes de neuronas (Perceptrón Multi-capa, *MLP* [Bishop, 1995; Rumelhart and McClelland, 1986]); reglas de decisión (PART [Frank and Witten, 1998]); bosques de árboles aleatorios (del inglés, *random forests* [Breiman, 2001]), cuya eficiencia se ha probado en otros dominios similares [Qi et al., 2006]; y otro método bayesiano, (*BayesNet* [Friedman et al., 1997; Bouckaert, 2004]), que también es un algoritmo relevante en este estudio.

BayesNet [Bouckaert, 2004] es un algoritmo de redes bayesianas [Friedman et al., 1997] que aprende tanto la estructura de la red como la tabla de probabilidades. Para inferir la estructura de la red se usa un algoritmo de búsqueda llamado K2 [Cooper and Herskovits, 1992], que añade arcos con un orden fijo de las variables, usando una métrica bayesiana [Bouckaert, 2004] para evaluar la calidad de la red aprendida. Para estimar las distribuciones de probabilidad condicional de la red bayesiana se usa un estimador simple [Bouckaert, 2004]. Hay una modificación en la configuración por defecto de *BayesNet*, que se refiere al número máximo de padres que puede tener un nodo en la estructura de la red, el cual se fija a 2, aprendiendo así una red de bayes de árbol aumentado (TAN) [Friedman et al., 1997].

Para toda la experimentación realizada con algoritmos de Aprendizaje Automático, se usa la implementación de la herramienta Weka [Witten and Frank, 2005].

6.2.6. Esquema Resumen Sistema de Aprendizaje

La figura 6.2 resume el sistema descrito en las secciones previas, utilizado en este trabajo para predecir Asociaciones Funcionales entre Pares de Proteínas.

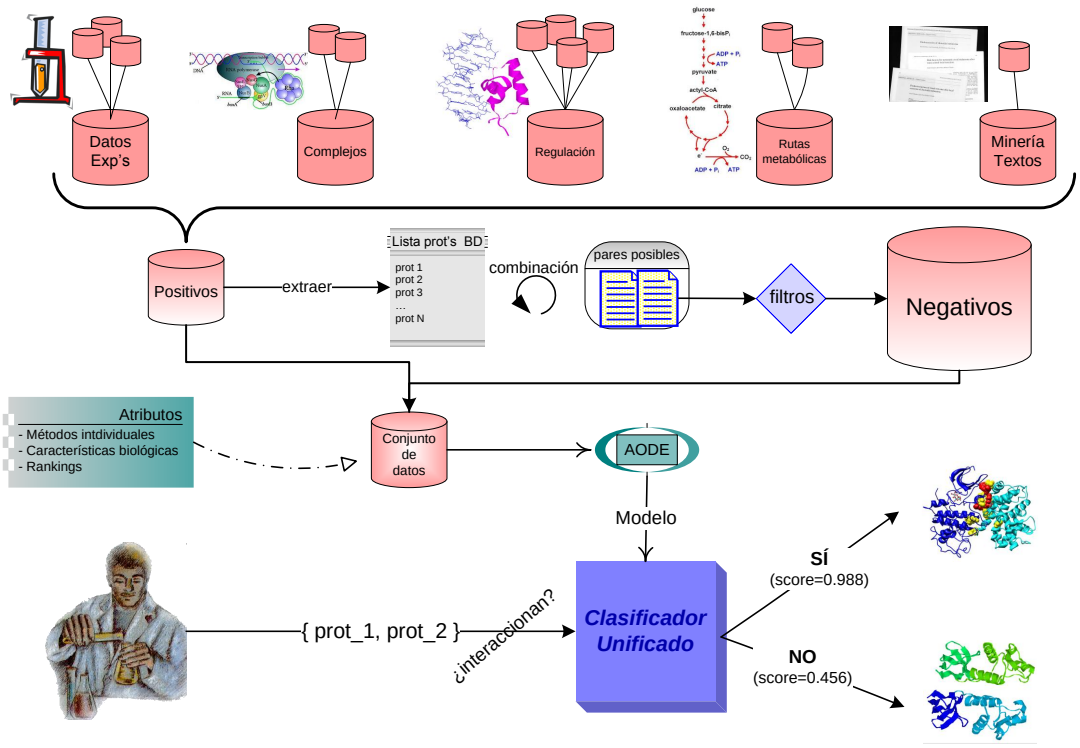


Figura 6.2: Esquema sistema de predicción de AFPP en *E.coli*.

6.3. Resultados e Interpretación

6.3.1. Comparación de Varios Algoritmos

Se exploran varios algoritmos y varios conjuntos de entrenamiento con diferentes proporciones entre instancias positivas y negativas para alcanzar un buen compromiso con el desbalanceo de clases subyacente (porque la mayoría de pares no se espera que estén funcionalmente asociados). Esta experimentación permite determinar si los métodos probados son adecuados para manejar la incapacidad de la mayoría de los algoritmos para gestionar conjuntos altamente desbalanceados (ver detalles en la sección 6.2).

Este proceso exploratorio genera varios clasificadores, cuya fiabilidad se compara a través del conjunto de test de asociaciones funcionales entre proteínas (ver 6.2.3). La comparativa se evalúa con curvas de coste [Drummond and Holte, 2006], que permiten elegir rápidamente el

mejor clasificador por inspección visual directa. Para una descripción del uso de estas curvas, consultar el capítulo 4. Como regla general, los mejores clasificadores están por debajo de los peores ya que tienen un coste más bajo (errores de clasificación). De hecho, la diferencia de error entre un par de clasificadores se puede medir a través de la distancia vertical entre sus curvas.

La figura 6.3 muestra las curvas de coste de diferentes clasificadores, representando los diferentes algoritmos de Aprendizaje Automático usados. Tras una inspección visual de este gráfico se aprecia claramente que el mejor algoritmo en la fase de test es AODE (línea azul, figura 6.3). La curva de coste de AODE aparece por debajo de las curvas de coste de todos los otros algoritmos para la mayoría de los valores de probabilidad, lo que significa que AODE comete menos errores (falsos positivos y falsos negativos) que cualquier otro clasificador para la mayoría de distribuciones positivo/negativo.

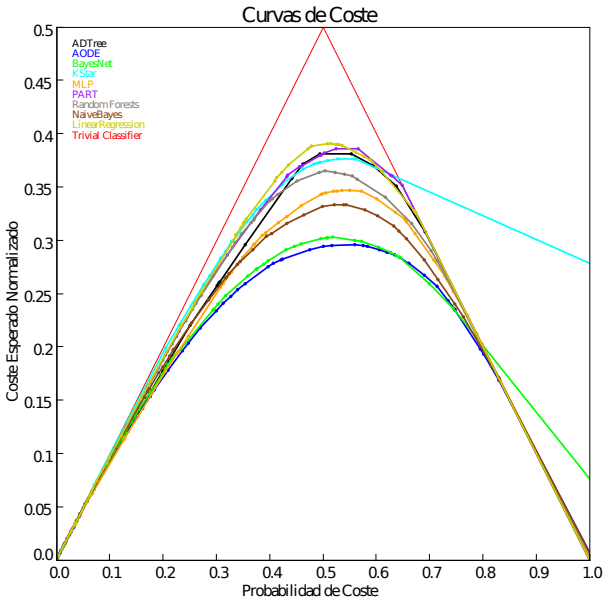


Figura 6.3: Curvas de coste de varios algoritmos de AA que predicen AFPP. El *eje x* representa la probabilidad de coste y el *eje y* el coste esperado normalizado. Cada curva de coste corresponde a un algoritmo de Aprendizaje Automático diferente. Mirando la leyenda de arriba a abajo, los algoritmos son: ADTree, un árbol de decisión; AODE y *BayesNet*, 2 métodos bayesianos; Kstar, un algoritmo de razonamiento basado en casos; MLP, una red de neuronas; PART, un método de reglas de decisión; *Random Forests*, una combinación de árboles de clasificación; *Naive Bayes*; y Regresión Lineal. El último es el clasificador trivial, sin ningún algoritmo asignado. Ver la sección 6.2 para obtener la referencia de cada algoritmo.

BayesNet es el segundo mejor clasificador en términos de rendimiento, lo cual enfatiza que los algoritmos con enfoques bayesianos son los más apropiados para afrontar este problema. Es interesante destacar que *BayesNet* es peor que AODE e incluso que el clasificador trivial (triángulo rojo en la figura 6.3) cuando la probabilidad de coste es mayor de 0,8. Este hecho será irrelevante en la mayoría de los casos, excepto para experimentos que impliquen el filtrado de conjuntos de interacciones altamente fiables obtenidas de fuentes experimentales. Además, *BayesNet* reemplaza todos los valores desconocidos con la mediana de los valores del

conjunto de entrenamiento para el atributo correspondiente, en vez de ignorar dichos valores desconocidos como hace AODE. Respecto a este punto, el enfoque de AODE es más apropiado para la semántica de este dominio, donde un valor desconocido implica la *no existencia*. Este aspecto es importante porque la mayoría de las instancias (o pares de proteínas) tiene al menos un valor desconocido, y la información que representa la ausencia de un valor se espera que sea más inestable y más difícil de extrapolar cuando se predice sobre nuevos pares de proteínas. También es importante destacar que, de acuerdo a las curvas de coste (figura 6.3), el tercer mejor clasificador es *Naive Bayes* [John and Langley, 1995] (que considera que existe una independencia completa entre los atributos de entrada). La comparación de estos tres métodos bayesianos muestra que modelar adecuadamente la dependencia interna entre las características de entrada mejora notablemente los resultados.

Aunque se ha demostrado el valor de otros algoritmos (como los bosques de árboles aleatorios [Breiman, 2001]) sobre trabajos previos relacionados [Qi et al., 2006], AODE se presenta como el más apropiado de las combinaciones exploradas de problema, atributos y sistema experimental. Por lo tanto, el rendimiento superior de AODE sobre la evaluación en el test no se puede tomar como una prueba de superioridad general del algoritmo. Acorde a las características específicas de cada problema de predicción, distintos métodos consiguen diferentes resultados. Además, alguno de estos clasificadores podrían mejorar su comportamiento con la exploración detallada de su espacio de parámetros.

Para propósitos comparativos, la figura 6.4 muestra la evaluación de los algoritmos sobre el test con la ampliamente aplicada curva ROC. El análisis de estas curvas ROC proporciona las mismas conclusiones y apoya la superioridad de los clasificadores bayesianos para este problema, aunque sin grandes diferencias entre los dos enfoques que tienen en cuenta la dependencia entre los atributos. Resumiendo brevemente, estos dos clasificadores bayesianos son claramente mejores que los otros clasificadores probados, aunque AODE proporciona un ligero rendimiento mayor en algunas condiciones. Como consecuencia de los resultados de esta evaluación, se considera AODE como una elección adecuada para resolver este problema, incluso siendo imposible garantizar que será mejor que cualquier otro clasificador en todas las condiciones.

6.3.2. Análisis de Relevancia de Atributos

En esta sección se analiza el rendimiento de AODE para distintas combinaciones de atributos de entrada.

En la tabla 6.3 aparece una comparación de la contribución de las características usadas en el aprendizaje, con distintas medidas de evaluación. Se realiza una inclusión incremental de atributos, de abajo a arriba, incluyendo los rankings al final porque se derivan de los métodos correspondientes.

Se puede observar que todos los atributos aportan valor, dado que los resultados, principalmente evaluados en AUROC y MCC, mejoran siempre que se añade un nuevo grupo de atributos, es decir, al ir de abajo hacia arriba en la tabla 6.3. Para concluir, se verifica que la combinación elegida, con todos los atributos (primera fila de la tabla 6.3), es la que presenta una mejor evaluación, equilibrada entre AUROC y MCC.

6.3.3. Mejora en la Combinación de Distintas Fuentes de Información

En este apartado se compara la capacidad de predicción del clasificador unificado frente a los métodos computacionales individuales existentes previamente, cuyas salidas integra el

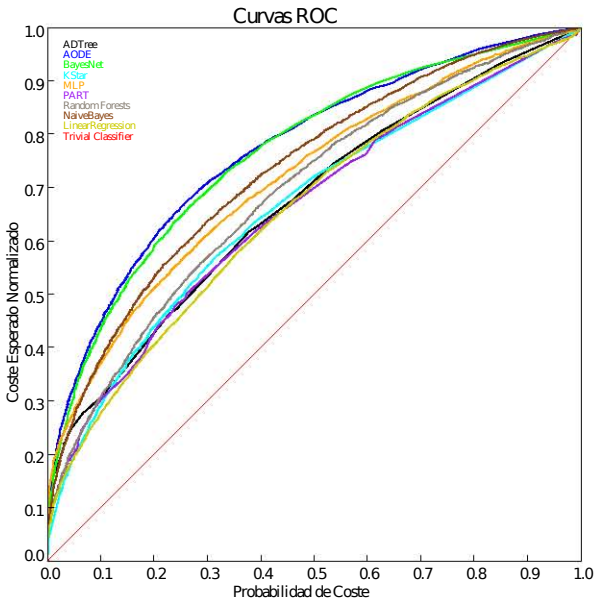


Figura 6.4: Curvas ROC de varios algoritmos de AA que predicen AFPP. El *eje x* representa el ratio de aciertos positivos y el *eje y* el ratio de falsos positivos. La leyenda se debe interpretar como en la figura 6.3, con el mismo orden de algoritmos.

nuevo clasificador.

En primer lugar, es conveniente comparar la exactitud de las predicciones positivas de los métodos originales con las del clasificador basado en AODE, ya que cada método tiene distinta aplicabilidad y es potencialmente capaz de detectar diferentes tipos de asociaciones funcionales. Se compara la precisión de las *'n'* primeras predicciones de un conjunto de test extendido. Éste incluye el conjunto completo de posibles predicciones para *E.coli* después de eliminar aquellos pares usados en el conjunto de entrenamiento. En la figura 6.5 cada línea representa la precisión, medida como el ratio de predicciones positivas verdaderas dividido por el número de predicciones en el conjunto de test extendido, para un número incremental de pares predichos (la figura 6.6 es la equivalente pero restringida sólo al conjunto de test).

Como se puede ver, AODE se comporta mejor que cada uno de los métodos individuales a lo largo de todo el rango de las *'n'* primeras predicciones. Así, AODE tiene una precisión de 0,97 para las 100 primeras predicciones, 0,69 para las 1.000 primeras, 0,56 para las 2.000 primeras, 0,49 para las 3.000 primeras, etc.

Cuando la comparación se hace sobre el conjunto más grande de pares predichos por los 5 métodos (es decir, las 800 primeras predicciones de cada método, que es el máximo de predicciones proporcionadas por el método GF), AODE es 1,41 veces más preciso que el método GC, 3,80 veces más preciso que GF, 9,65 veces más fiable que PP, 32,38 veces más fiable que MT y 47,67 veces más preciso que I2H. Los resultados obtenidos con GC son los más cercanos a AODE y, de hecho, ambos métodos definen casi el mismo perfil, aunque AODE es 20 puntos porcentuales más preciso. La razón de esta diferencia parece ser la información añadida por los otros métodos individuales (GF, I2H, MT y PP), así como los atributos adicionales (longitud de las proteínas y tamaño de la familia de proteínas) que usa

Tabla 6.3: Comparación de relevancia de atributos en predicción de AFPP sobre el conjunto de test. AUROC: área bajo la curva ROC, MCC: Coeficiente de Correlación de Matthews, TP: verdaderos positivos, TN: verdaderos negativos, FP: falsos positivos, FN: falsos negativos.

Atributos	AUROC	MCC	TP	TN	FP	FN
Métodos, Longitud, Nºort., Rankings	0,77	0,35	1.362	21.523	4.160	565
Métodos, Longitud, Nºort.	0,77	0,27	903	21.685	4.619	403
Métodos, Longitud	0,67	0,18	542	21.714	4.980	374
Métodos	0,57	0,14	305	21.903	5.217	185
Longitud	0,67	0,13	306	21.843	5.216	245
Nº ortólogos	0,61	0,06	146	21.891	5.376	197

AODE.

Además, se detecta que algunos métodos individuales son muy imprecisos (menos del 10 %), como es el caso de I2H, MT y PP. Por el contrario, GC produce una proporción de predicciones correctas mucho más elevada, siendo alrededor de 8 veces más preciso que PP (el mejor de los tres métodos más pobres en predicción). GF proporciona muy pocas predicciones debido a que depende de la ocurrencia de un evento particular. La relativa baja frecuencia de eventos de fusión de genes limita la habilidad de GF para predecir la mayoría de asociaciones funcionales.

En segundo lugar, se puede hacer una comparación basada en la cobertura de aplicación de cada método. Hay que resaltar que, en contraposición a los métodos individuales de partida, cuya salida está muy limitada por las restricciones de aplicación de cada uno, el clasificador unificado es capaz de dar una predicción (positiva o negativa, con su probabilidad asociada) para casi cualquier par de proteínas que se le presente. Como ejemplo clarificador, los métodos I2H y GF no dan predicción para el 21 % y 99 % de los casos, respectivamente.

Esta comparación, respecto a la cobertura de aplicación de cada método, queda reflejada en la figura 6.7, que representa mediante curvas de coste los 6 métodos computacionales, aplicados sobre el mismo conjunto de test. Cualquier par de proteínas para el que un método no da un valor de salida, se considera una predicción negativa, con una puntuación 0. Teniendo esto en cuenta, se observa que los métodos individuales tienden a solaparse con el clasificador trivial (triangulo central de la figura 6.7) a lo largo de casi todo el *eje x*. Mientras, el método unificado (línea roja de la figura 6.7) mantiene unos costes aceptablemente bajos en todo el rango.

Por lo tanto, AODE logra una mayor cobertura de aplicación, siendo capaz de aportar predicciones para un conjunto más amplio de pares de proteínas, frente a los métodos individuales, que se encuentran muy limitados a su contexto o restricciones de aplicación.

Por último, es importante tener en cuenta que la definición global de asociaciones funcionales que se usa en este trabajo se centra en los tipos de asociaciones más numerosos en los datos de partida (ver la sección 6.2). En este estudio, las principales fuentes de interacciones y asociaciones funcionales son la co-regulación y las rutas metabólicas. Por lo tanto, es esperable que AODE sea más ventajoso para la predicción de estos dos tipos de asociaciones funcionales, al comparar AODE con otros métodos basados en diferentes fuentes o con diferentes proporciones entre los tipos de asociaciones funcionales.

En conclusión, AODE fusiona varios métodos de predicción mejorando en rendimiento y cobertura a los métodos computacionales individuales que combina y complementa con información adicional.

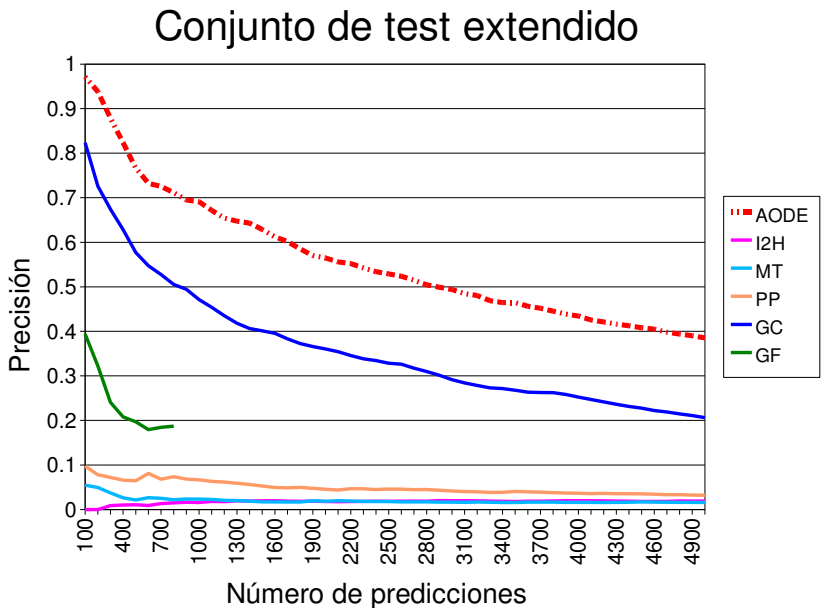


Figura 6.5: Precisión de métodos individuales y método unificado en el conjunto de test extendido. El *eje x* representa el número de ejemplos (pares de proteínas) acumulados, sobre los que se aplica un método, ordenados según la puntuación (o salida) de dicho método. La puntuación es diferente en cada caso, por lo que las *'n'* primeras predicciones de cada método no son las mismas. El *eje y* representa la cantidad de aciertos en la clase positiva entre el total de ejemplos para los que el método en cuestión es capaz de dar una clasificación (nº de predicciones restringidas). AODE, I2H, MT, PP, GC y GF son los diferentes métodos computacionales de predicción.

6.3.4. Evaluación para Diferentes Categorías de Fuentes de Datos

El objetivo de este apartado es presentar unas breves observaciones sobre la posibilidad de predecir asociaciones funcionales por categorías.

En primer lugar, hay que decir que la construcción de un predictor independiente para cada fuente de datos no es recomendable, o incluso inviable, porque se deben dividir los datos de entrenamiento. Con lo que las asociaciones funcionales que quedan pueden ser insuficientes para aprender, mayoritariamente cuando la base de datos disponible no es grande, como por ejemplo la de regulación o de datos experimentales (ver tabla 6.1 con la cantidad de instancias positivas procedentes de cada base de datos).

Así, sin un entrenamiento individual, la única posibilidad es una evaluación con los datos de test por cada categoría de asociaciones funcionales o fuentes de datos. Dado que el criterio para construir el conjunto de ejemplos negativos del sistema AODE aplica a todo el conjunto de asociaciones funcionales, sin división por base de datos, para no añadir fuentes de ruido adicionales, la evaluación por categoría que se presenta se basa sólo en los ejemplos positivos del conjunto de test. Dicho conjunto de test, correspondiente al clasificador global, no sigue una distribución homogénea entre categorías. Según se describen en la sección 6.2.1, dichas categorías son: datos experimentales (248 asociaciones funcionales en test), complejos (676 asociaciones), regulación (143 asociaciones), co-regulación (2.863 asociaciones), rutas

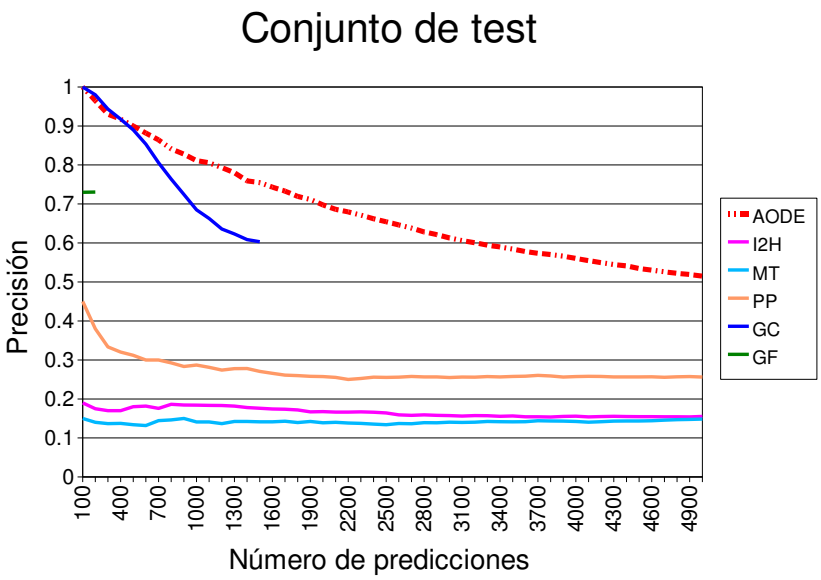


Figura 6.6: Precisión de métodos individuales y unificado en el conjunto de test. Los *ejes x* e *y* se deben interpretar como en la figura 6.5.

metabólicas (1.942 asociaciones) y minería de textos (950 asociaciones), conteniendo ejemplos solapados entre distintas categorías.

Por lo tanto, la comparativa presentada en la figura 6.8 y las conclusiones derivadas en este apartado se ven afectadas por todas estas limitaciones.

Al observar la figura 6.8, como se esperaba, la tasa de aciertos en positivos en los datos experimentales es la más baja, dado que el sistema está entrenado con muchas más asociaciones funcionales que interacciones físicas. Sin embargo, sorprende que para casi cualquier umbral, las asociaciones de minería de textos y las interacciones en complejos sean superiores al resto, porque no son las fuentes de datos más abundantes, correspondientes a co-regulación y rutas metabólicas (ver tabla 6.1).

En conclusión, dadas las restricciones en los conjuntos independientes disponibles para evaluar el sistema global, este análisis no se puede considerar una evaluación definitiva, sino simplemente una observación. Porque el sistema de predicción de asociaciones funcionales está basado en un clasificador global, con todas las fuentes de datos mezcladas. No obstante, el sistema global es prácticamente la única alternativa con suficientes ejemplos para aprender. Por lo tanto, sólo se tiene una aproximación sesgada de la evaluación por fuentes, debido a las diferentes limitaciones del proceso de aprendizaje global.

6.4. Aplicación para Filtrar Interacciones Experimentales

Para mostrar el potencial de AODE, se ha aplicado a un conjunto de datos experimentales no incluido originalmente en los conjuntos de entrenamiento ni test. Para este propósito, se recopila el conjunto de complejos de proteínas detectados por Arifuzzaman et al. [Arifuzzaman

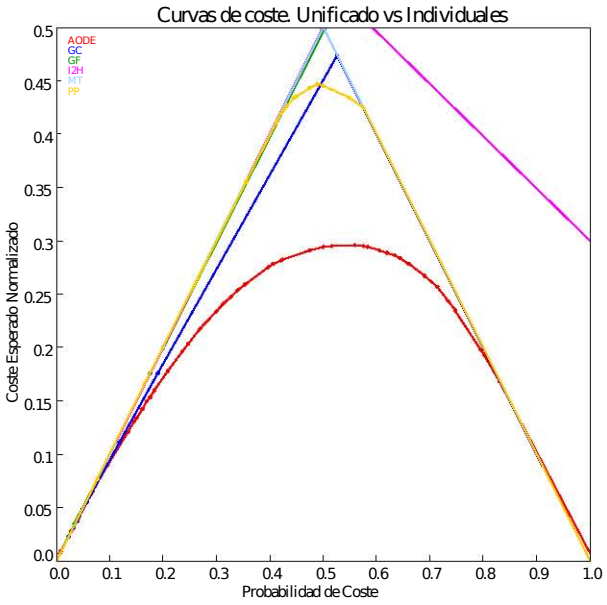


Figura 6.7: Curvas de coste de métodos individuales frente a unificado en conjunto de test. El *eje x* representa la probabilidad de coste y el *eje y* el coste esperado normalizado.

et al., 2006]. Estos datos se obtuvieron mediante un enfoque experimental a gran escala, basado en tecnologías de co-immunoprecipitación, aplicadas a las proteínas de *E.coli*.

Se ha demostrado que las tecnologías a gran escala, aunque valiosas, frecuentemente producen un gran número de falsos positivos debido a diferentes defectos metodológicos. Se ha elegido este ejemplo en particular porque incluye una gran cantidad de asociaciones entre proteínas que no se han podido confirmar por ninguna otra fuente de datos disponible, es decir, que dichas asociaciones no se han detectado por ningún otro método. De hecho, sólo el 7,85 % de los datos se confirman con el extenso conjunto de predicciones de asociaciones funcionales de AODE (cubriendo sólo el 0,64 % de éstas). El número de confirmaciones externas es pequeño (un hecho común cuando se comparan diferentes fuentes de asociación funcional), incluso aunque el conjunto de bases de datos externas recopiladas, con las que se entrena el sistema, incluya otros conjuntos de complejos de proteínas. No obstante, hay que recordar que el conjunto de entrenamiento del sistema combina muchas bases de datos diversas, incluidas algunas con más (como los datos experimentales a pequeña escala) y otras con mucha menos fiabilidad que los complejos (como las fuentes de minería de textos).

AODE se usa para detectar el subconjunto de interacciones que potencialmente tienen significado biológico, tras asignarle un nivel de confianza a cada asociación entre proteínas de este conjunto de datos a gran escala. Para ello, se ordena el conjunto de pares de proteínas recopilados por Arifuzzaman usando la puntuación de AODE como medida de probabilidad de ser una asociación funcional. Entonces se compara el nivel de confirmación de la predicción para los *'n'* pares mejor puntuados (línea verde de la figura 6.9) con el nivel obtenido para el conjunto completo de Arifuzzaman (punto azul de la figura 6.9) y con el nivel de confirmación para aquellos pares predichos para el proteoma completo (línea roja de la figura 6.9).

Los resultados muestran claramente que la combinación de información de AODE puede

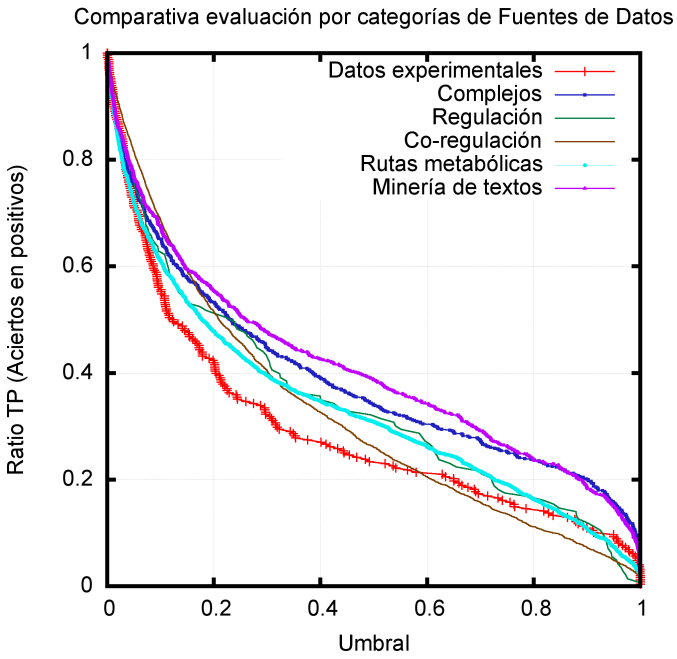


Figura 6.8: Evaluación de la predicción de asociación funcional por diferentes categorías de fuentes de datos. Sistema AODE entrenado con todas las fuentes de datos, y test sobre el subconjunto correspondiente a cada categoría, sólo de positivos. Se representa el ratio de aciertos en positivos (predicciones correspondientes a asociaciones funcionales reales) sobre el total de instancias positivas (asociaciones funcionales presentes en las fuentes origen) para distintos umbrales de corte en la probabilidad de predicción.

extraer un conjunto de asociaciones funcionales significativas a partir de la ruidosa colección de datos original. Por ejemplo, el 68 % de los primeros 100 pares y el 42 % de los primeros 1.000 pares se confirman en la lista puntuada con los valores de AODE. Estas cifras son muy significativas cuando se comparan con el 8 % de confirmación original del conjunto completo de 7.283 asociaciones del conjunto de Arifuzzaman. Cuando se compara el conjunto de Arifuzzaman filtrado (línea verde de la figura 6.9) y las predicciones de AODE (línea roja de la figura 6.9) es importante destacar que las predicciones del conjunto filtrado son menos fiables que las del proteoma completo, porque éstas incluyen pares diferentes. De hecho, la mayoría de los pares fiables del proteoma completo no se recuperan con el experimento de Arifuzzaman y por lo tanto AODE no las puede incluir en el conjunto filtrado.

Estos resultados muestran el poder de combinar diferentes fuentes de datos, de distinta fiabilidad, para asignar fácilmente un nivel de calidad a las interacciones entre proteínas procedentes de experimentos a gran escala, las cuales no poseen a priori un indicador de confianza de la predicción, que permita ordenarlas para extraer las más fiables.

6.5. Comparación con la Base de Datos STRING

STRING [Jensen et al., 2009] es una base de datos dedicada a la predicción de asociaciones funcionales entre proteínas para un conjunto de genomas completamente secuenciados.

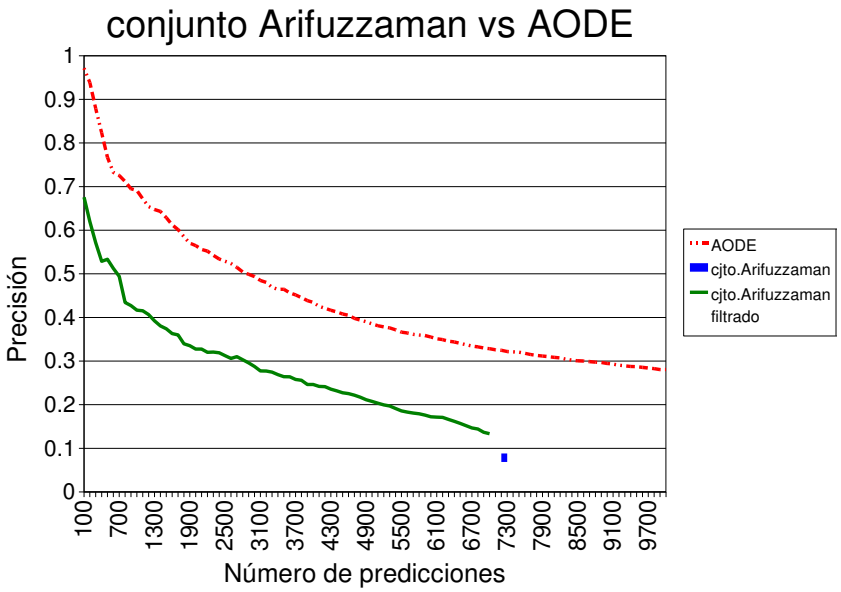


Figura 6.9: Precisión del método unificado sobre el conjunto experimental de Arifuzzaman. Los *ejes x* e *y* se deben interpretar como en la figura 6.5. El conjunto filtrado (la línea verde) se obtiene ordenando los pares de proteínas en el conjunto de Arifuzzaman, según la puntuación de AODE. La precisión del conjunto de Arifuzzaman se representa por su valor medio (el punto azul), ya que este conjunto de datos no se puede ordenar por no tener una puntuación asociada a cada par.

Contiene una recopilación exhaustiva de información, desde importaciones de bases de datos externas hasta predicciones generadas por el equipo de STRING, incluyendo versiones de algunos de los métodos individuales descritos en este trabajo, como la fusión de genes, el contexto genómico o los perfiles filogenéticos.

STRING tiene su propia definición de *Patrón Oro* (del inglés, *Gold Standard*) para las asociaciones funcionales, que está basado en las rutas metabólicas [Jensen et al., 2009]. Este enfoque difiere de la estrategia adoptada en el presente trabajo, porque STRING no incluye relaciones de regulación entre factores de transcripción y genes regulados, ni entre los genes regulados por el mismo factor de transcripción. Incluso lo que es más importante, STRING se aprovecha de la información experimental disponible para predecir nuevas asociaciones funcionales metabólicas, utilizando interesantes datos experimentales adicionales para un par de proteínas dado. En cambio, las predicciones que produce AODE están más enfocadas a asociaciones sin confirmación experimental disponible. Lo que también implica que, en principio, las predicciones de AODE son aplicables a cualquier proteína (dentro del proteoma de *E.coli* aquí analizado). Por lo tanto, la cobertura y la capacidad de descubrir asociaciones desconocidas de AODE debería ser más alto, mientras su capacidad para detectar asociaciones de proteínas bien caracterizadas será necesariamente más bajo.

Para probar estas ideas, se usa AODE para extraer aquellas entradas de STRING con puntuaciones más altas, a partir de un conjunto de 240.885 pares de proteínas de STRING, con el valor de confianza de la predicción mínimo al 0,15. Para obtener una visión de la habilidad de ambos enfoques para detectar asociaciones funcionales desconocidas, todos los pares validados

experimentalmente se eliminan. En la figura 6.10 se compara la predicción de las puntuaciones de STRING y AODE para el conjunto de 121.042 asociaciones comunes a ambos conjuntos, que representan un 50,25 % de los pares de STRING.

Los resultados muestran claramente que ambas definiciones de enlaces funcionales sólo solapan parcialmente. Además, se concluye que AODE puede complementar las predicciones de STRING con una definición más extensa de asociaciones funcionales entre pares de proteínas.

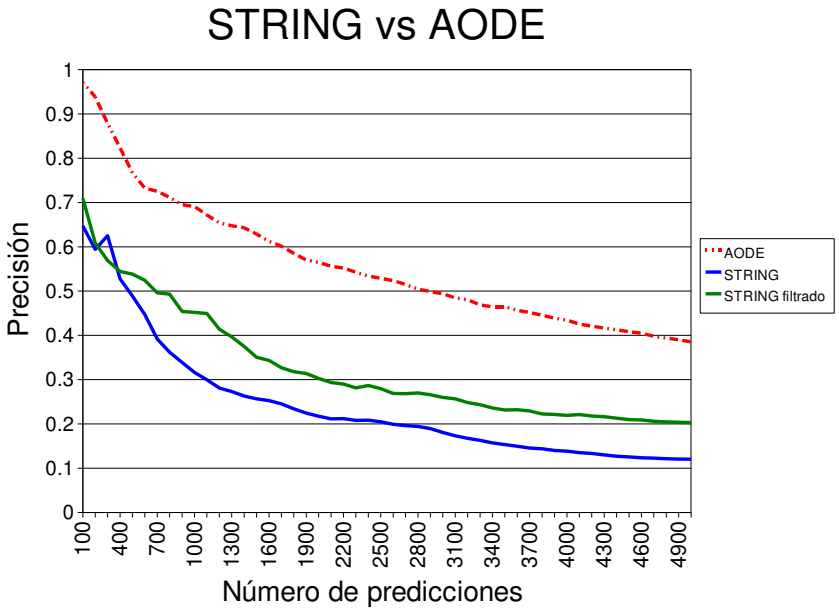


Figura 6.10: Comparación de precisiones del método unificado y STRING, sobre el conjunto de predicciones de STRING. Los *ejes x* e *y* se deben interpretar como en la figura 6.5. El conjunto STRING filtrado (la línea verde) se obtiene ordenando los pares de proteínas en la base de datos externa, es decir, STRING, según la puntuación de AODE. La línea de STRING (azul) se calcula ordenando los datos según la puntuación de STRING, es decir, la puntuación para cada par de la base de datos externa.

6.6. Servidor de Predicciones EcID

Los resultados de cada uno de los 5 métodos de predicción individuales y de su combinación en AODE están integrados en el servidor EcID (del inglés, *E.coli Interaction Database*, [Leon et al., 2009]), permitiendo al usuario extraer y navegar fácilmente por la red de interacciones y asociaciones funcionales entre proteínas.

EcID proporciona dos modos básicos de navegación por la red: el 'Modo Experimental' enfocado en extraer asociaciones funcionales apoyadas experimentalmente (similar al enfoque de STRING), y el 'Modo de Predicción' centrado en suministrar predicciones para las proteínas menos caracterizadas. Las puntuaciones de AODE, calculadas según se describe en este capítulo con el clasificador basado en AODE, se usan en EcID para generar un criterio

de confianza de la predicción, para las asociaciones funcionales mostradas en el 'Modo de Predicción'. Este criterio de confianza, que aporta el sistema unificado de predicción AODE, permite al sistema EcID una medida única para seleccionar un conjunto de relaciones más probable para proteínas pobremente caracterizadas. Además, esto cumple otro de los propósitos originales, como es conseguir resultados para las proteínas menos caracterizadas, además de ordenar las asociaciones bien conocidas entre proteínas que también permite AODE. El servidor es de acceso libre en http://ecid.bioinfo.cnio.es/, presentando un aspecto como el mostrado en la figura 6.11.

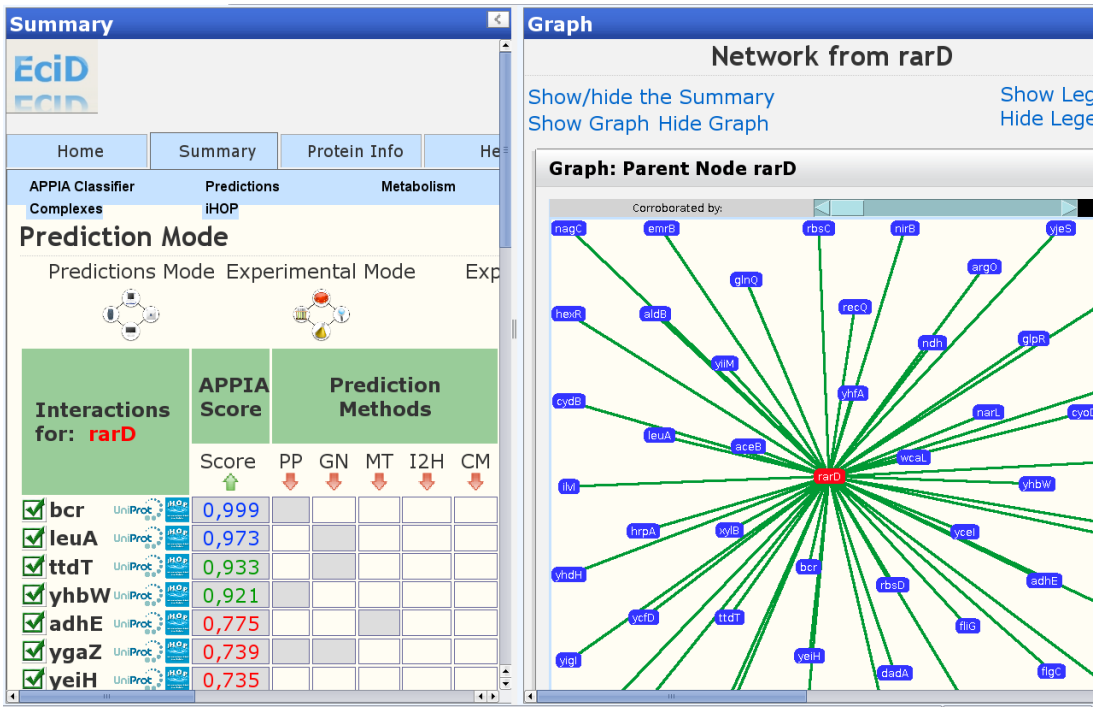


Figura 6.11: Ejemplo de vista de servidor de predicciones EcID.

6.7. Conclusiones

Este capítulo presenta un enfoque basado en Aprendizaje Automático para la predicción de asociaciones funcionales entre proteínas, integrando un par de características básicas de las dos proteínas, los resultados de cinco métodos computacionales heterogéneos, y una medida de los pares potenciales de proteínas asociadas según cada método.

Los resultados sobre un conjunto de test independiente del de entrenamiento confirman la adecuación de los algoritmos bayesianos para afrontar el problema planteado, según un análisis con curvas de coste y curvas ROC sobre algoritmos alternativos de Aprendizaje Automático. El mejor algoritmo es AODE, seguido de *BayesNet*, con resultados ligeramente peores. Adicionalmente, AODE es adecuado para problemas con valores desconocidos, siendo flexible en su gestión, lo que es conveniente en este caso donde no siempre es posible obtener predicciones con los cinco métodos individuales para los mismos pares de proteínas. Además, AODE proporciona una estimación cuantitativa, útil como medida unificada de fiabilidad de la

asociación de cada par de proteínas (utilizada como agregado en el servidor de predicciones EcID).

Todos los métodos e información de entrada del clasificador provienen de experimentos de secuenciación genómica. Por lo tanto, el clasificador unificado aporta conocimiento preferentemente sobre las asociaciones funcionales poco caracterizadas, en vez de sobre las bien definidas. En este sentido, el nuevo enfoque propuesto en este capítulo se diferencia de otros enfoques populares y exitosos, como STRING, porque es independiente de la información experimental disponible para un par de proteínas considerado.

Por otro lado, este clasificador basado en AODE mejora los resultados de los métodos originales individuales que incorpora. Tanto en precisión como en cobertura de aplicación. Especialmente es mucho mejor en este segundo aspecto, siendo capaz de aportar predicciones para cualquier par de proteínas con un solo método aplicado, mientras que cada método individual está limitado a una región del espacio de predicciones que satisfaga sus restricciones. Sin embargo, existe una limitación en el sistema de predicción presentado, debido a que integra otros métodos existentes. Por tanto, las actualizaciones del clasificador basado en AODE requerirían que se actualicen previamente los métodos que se combinan, o que se incorporen otros nuevos.

El clasificador presentado en este trabajo también permite refinar y enriquecer otros listados de interacciones y asociaciones funcionales existentes. Particularmente, los pares de un experimento de co-immunoprecipitación a gran escala, que se ordenan y filtran según su significado funcional. Este resultado destaca el valor de este tipo de enfoques para eliminar la considerable cantidad de falsos positivos que generan los enfoques experimentales a gran escala, por la falta de una medida de confianza de la predicción, lo cual es todavía uno de los principales inconvenientes de dichas técnicas, que no definen cuáles son las predicciones más fiables.

Capítulo 7

Extensión de Rutas Biológicas en Humanos

En este capítulo se describe el segundo problema afrontado en esta tesis de anotación funcional dentro de la Biología de Sistemas, junto con los resultados obtenidos y su interpretación. Se trata de añadir proteínas a una red en la que se comparte función a nivel de proceso biológico, a partir de combinaciones de propiedades simples de la secuencia e información relacional de proteínas, interaccionando por pares o en complejos.

El contenido del capítulo en detalle es el siguiente. En las dos primeras secciones se describe el contexto del problema, los datos recopilados, su representación según el modelo de conocimiento genérico presentado en el capítulo 5, y el método de aprendizaje propuesto (híbrido entre relacional y proposicional) para resolver el problema de anotación de rutas biológicas. En la sección 7.3 se presentan los resultados del aprendizaje desde una perspectiva puramente computacional. En la sección 7.4 se describe una serie variada de interesantes análisis interpretativos de las predicciones, teniendo presente la realidad biológica a la que pertenecen. En las secciones 7.5 y 7.6 se compara la propuesta de esta tesis de extensión de rutas biológicas con otros métodos que podrían resolver la misma tarea de anotación funcional, pero que se basan en otros principios. La sección 7.7 expone la relevancia biológica de las proteínas añadidas a rutas biológicas concretas por el método propuesto en esta tesis, detallando casos particulares de proteínas cuya predicción parece estar justificada por anotaciones en bases de datos biológicas y publicaciones científicas, a falta de una verificación experimental. Finalmente, se expone una discusión y resumen de las conclusiones del capítulo completo.

7.1. Definición del Problema

Las rutas biomoleculares representan una recopilación abstracta del conocimiento sobre procesos metabólicos, de regulación y de señalización, organizados como cascadas de interacciones entre proteínas, con la implicación de otros compuestos moleculares. Son las responsables de lograr resultados fenotípicos biológicos específicos [Cary et al., 2005; Ooi et al., 2010].

El núcleo de la biología de sistemas son las redes bioquímicas y de señalización. Las rutas metabólicas y de señalización son cada vez una parte más importante de la organización del conocimiento en la Biología de Sistemas [Kemper et al., 2010]. Además, en los últimos años se ha mostrado un renovado interés en almacenar y anotar rutas [Demir et al., 2010; Cerami et al., 2011], implicando varios retos que incrementan el interés por las rutas biológicas. Entre

dichos retos está el crecimiento de la cantidad de información experimental disponible, las limitaciones obvias en las bases de datos y los recursos de anotación, y las variadas definiciones sobre qué constituye una ruta. Adicionalmente, en humanos, la des-regulación de los sistemas de señalización ha estado implicada en diversas patologías, como el cáncer [Vogelstein and Kinzler, 2004], la degeneración neuronal, la atrofia muscular, la inmunodeficiencia y la diabetes [Sakharkar et al., 2007].

Como consecuencia de las distintas definiciones de ruta biológica encontradas en la literatura científica [Bader et al., 2006], su implementación no es la misma en bases de datos diferentes, como por ejemplo Reactome [Matthews et al., 2009], KEGG [Kanehisa and Goto, 2000] o MetaCyc [Caspi et al., 2010]. De hecho, existen esfuerzos dirigidos a desarrollar una forma estándar para representar, anotar y mostrar una visión común de las rutas biológicas (como *Pathway Commons* [Cerami et al., 2011]), pero se complica por las complejas diferencias, que en muchos casos se deben a criterios científicos distintos sobre la definición de una ruta.

Debido a la naturaleza de integración de las rutas, se requiere un esfuerzo humano sustancial para construirlas. Los diseñadores expertos construyen una ruta después de leer e interpretar numerosos artículos científicos [Kemper et al., 2010], quedando frecuentemente representada una ruta a través de interpretaciones conjuntas de hechos dispersos por la literatura [Kell and Oliver, 2004; Luciano and Stevens, 2007]. Además, distintos expertos podrían tener una interpretación diferente de los hechos. De forma que, ante el mismo conjunto de información, el diseño de la ruta depende de quién lo haga, variando el resultado final en la inclusión de algunas de las proteínas.

Por todas estas razones, en este capítulo existe un interés por explorar las posibilidades de expandir rutas biomoleculares con proteínas potencialmente relacionadas, pero que no se incluyeron en la definición original de las rutas. Estas proteínas adicionales, relacionadas con el mismo proceso biológico (tales como reguladores), no se consideraron como pertenecientes a la ruta por ruido en los procedimientos experimentales, por ausencia de información en el momento del diseño o por la opinión personal de los expertos que la diseñaron [Lu et al., 2007]. Así, un reto aún pendiente es incluir estas proteínas en el proceso biológico en el que influyen, siendo una tarea de difícil generalización por la variabilidad en la definición de las rutas y por la falta de documentación en la que se basan las decisiones de los expertos que las definen, especialmente las razones por las que no se incluyen proteínas.

Esta aproximación de extensión de rutas difiere de aquellas desarrolladas para descubrir nueva rutas [Karp et al., 2002; Adriaens et al., 2008; Prather and Martin, 2008], a través de redes de interacción y otras características de las secuencias. Es decir, extenderlas es distinto de predecir nuevas rutas.

Cuando se conserva la definición de la ruta original, el uso de homología es uno de los enfoques más útiles para expandirlas [Korcsmaros et al., 2011]. Otra aproximación es usar las anotaciones de dominios de las proteínas (por ejemplo las de InterPro [Hunter et al., 2009]), con las que se han extendido las rutas de KEGG [Frohlich et al., 2008]. Recientemente, se ha propuesto un método alternativo para expandir rutas incorporando exclusivamente proteínas conectadas a la red por interacción [Glaab et al., 2010]. Sin embargo, en esta tesis se propone el sistema de Extensión basado en Representación Relacional (ERR) para extender rutas, usando principalmente características extraídas de las secuencias.

El enfoque del sistema ERR está relacionado con un conjunto de métodos diseñados para predecir función por medio de combinaciones de propiedades de secuencia sencillas, usados en diferentes escenarios [Jensen et al., 2002b, 2003a; Bendtsen et al., 2004]. Entre estas

aproximaciones, la más similar a ERR es el método *ProtFun*, desarrollado por el grupo de Soren Brunak, en particular cuando se emplea para asignar diferentes categorías de la ontología de Procesos Biológicos de *Gene Ontology* (GO-BP) a los genes humanos [Jensen et al., 2003a].

Hay que ser conscientes de que ambos sistemas, ProtFun y ERR, usan características moleculares para predecir procesos biológicos, lo cual es aparentemente contradictorio, porque las entradas y el objetivo de predicción no están en el mismo nivel de función biológica. A pesar de ello, en esta tesis se considera un enfoque con sentido, porque el sistema ERR no busca características comunes en la ruta global, sino propiedades específicas asociadas a diferentes fragmentos de la ruta. Esta idea concuerda con la heterogeneidad implícita de las rutas biológicas a nivel molecular. Metafóricamente, al igual que para conseguir construir un edificio se necesita un conjunto de personas expertas en cada labor (arquitectos, albañiles, electricistas, transportistas, etc.), en una ruta biológica también se necesita un grupo de proteínas encargadas de cada tarea particular (catalizar una reacción, trasladar agua a través de la membrana, fosforilar un compuesto, enlazar con una molécula de ADN, etc.), porque no todos los miembros de la ruta realizan la misma tarea a nivel individual. Así, al inducir conocimiento a partir de características específicas, las proteínas añadidas se parecerán a algunas proteínas concretas de la ruta original, en vez de a todas ellas.

El sistema ERR se concibe para resolver una tarea de predicción de función, considerando que una ruta o proceso biológico (es decir, una función) se asigna a una proteína o gen. También se puede considerar como predicción de pertenencia a un grupo de elementos relacionados (los que forman la ruta). Por lo tanto, el sistema se diseña de forma genérica, para que se pueda aplicar a otros vocabularios de anotación funcional, incluso compartiendo datos.

A continuación, se analizan en detalle las diferencias del sistema ERR con otros enfoques de anotación funcional muy semejantes; tanto frente a trabajos de perfil más biológico (Jensen y Brunak), como frente a otros de perfil más computacional (King y Vens).

Por un lado, las diferencias más importante con otros métodos de predicción basados sólo en secuencia y aplicables en ausencia de homología, como los desarrollados por Jensen y Brunak, son: 1) el uso de propiedades de secuencia más simples, y 2) el uso de información relacional (en este caso, las interacciones proteína-proteína y entre complejos).

1. ERR usa propiedades de secuencia muy simples, mientras que las aproximaciones de Jensen y Brunak usan características complicadas (modificaciones post-traducción, sitios de división de pro-péptidos, sitios de glicosilación, etc.). Estas propiedades se calculan a partir de la secuencia, pero necesitan un predictor dedicado a cada una, desarrollado en el grupo de Brunak, con complejos detalles internos que dificultan en gran medida que otros investigadores las pueden calcular. Sin embargo, las propiedades que usa el sistema ERR son muy sencillas (longitud, carga, etc.) y las puede obtener cualquier investigador fácilmente sin apenas procesamiento alguno. No obstante, el predictor resultante podría ser menos potente.

2. ERR extiende rutas basándose en una representación relacional. Porque, a parte de una colección de características de la secuencia, este sistema también usa información extraída de una red de interacciones, en concreto conexiones entre proteínas. Esto hace que ERR se pueda considerar un método de predicción híbrido, a medias entre los enfoques basados en redes y los basados en propiedades, descritos en la sección B.3. La representación relacional permite incluir información del contexto, como por ejemplo, las características de la secuencia de un compañero de interacción. Este conocimiento del dominio es interesante porque puede influir positivamente en la predicción, aunque los datos no pertenezcan a las características propias de la proteína principal. De este

modo, la representación relacional permite que el sistema ERR use una sofisticada combinación de algoritmos de aprendizaje automático relacional y proposicional, de forma que primero se extraen patrones frecuentes relacionalmente y posteriormente se inducen árboles de decisión proposicionalmente.

Por otro lado, se definen las diferencias más importantes del sistema ERR frente a otros enfoques más computacionales, que sí usan datos y representación relacional, como son los trabajos desarrollados por King y/o Vens. Las divergencias principales son que ERR: 1) usa datos de entrada más sencillos (sin homología), y 2) se aplica a un organismo más complejo.

1. La predicción en ausencia de homología (directa e indirecta), basada estrictamente en secuencia (consultar dificultades e implicaciones de este enfoque de predicción de función en la sección B.3.8), es la diferencia más importante del sistema ERR con los previos mencionados, que aplican la misma combinación de aprendizaje relacional y proposicional para anotación de función. Principalmente se trata de los trabajos del grupo de King, aplicando el método *DMP* (del inglés, *Data Mining Prediction*) [King et al., 2000b] sobre datos de homología y otras múltiples fuentes de información (anotaciones, datos de expresión, estructura terciaria, etc.), entre las que se incluyen atributos calculados a partir de algún tipo de relación de similitud. Por ejemplo, en uno de sus últimos trabajos [Clare et al., 2006], con el método *DMP*, se usan al menos tres fuentes de datos que implican homología, como son: anotaciones que se han podido producir por relaciones de similitud entre proteínas ortólogas, como los dominios InterPro; predicción de estructura secundaria calculada con información de similitud con el método *Prof* [Ouali and King, 2000]; y la inclusión directa de las relaciones de homología en predicados relacionales con su valor esperado (*e-value*) correspondiente. Con fines de análisis, en el capítulo 8 se estudia la inclusión de relaciones de homología (indirecta y directa) para predecir anotación funcional, en las secciones 8.6 y 8.7.

2. Otra diferencia importante con los trabajos de King y derivados es que ERR se aplica a un organismo más complejo, como es el humano, en vez de a especies procariotas o eucariotas simples. En concreto, King y Vens usan la misma combinación de aprendizaje relacional y proposicional para asignar términos de GO y MIPS a los genes de la especie vegetal *Arabidopsis thaliana* [Clare et al., 2006], y de la levadura *Saccharomyces cerevisiae* [Vens et al., 2008] (consultar más detalles de la evolución de esta combinación en la sección 2.3.1). En humanos, algunas asunciones para predicción basada en secuencia no se cumplen (como la proximidad en el cromosoma), por lo que no se pueden aplicar, restringiendo más la información de entrada disponible para predecir. No obstante, hay que decir que el vocabulario de anotación al que se aplica el sistema ERR en este trabajo tiene menos términos que los vocabularios utilizados en los trabajos de King y derivados.

En resumen, en comparación a otros enfoques semejantes, la tarea de predicción por parte del sistema ERR se dificulta notablemente. Por un lado, al limitar la información de entrada a características extraídas sólo de la secuencia y que no incluyan homología directa ni indirecta (diferencia principal con trabajos de King) y, por otro lado, al intentar a su vez aprovechar el restante conocimiento relacional y sólo usar propiedades simples (diferencias principales con trabajos de Jensen y Brunak). Pero dichas limitaciones permiten que el sistema ERR sea aplicable a un conjunto restringido de proteínas, poco caracterizadas, sin información experimental, ni homólogos conocidos.

En este capítulo se extienden las rutas del proteoma humano de Reactome [Matthews et al., 2009]. Se elige Reactome por ser una base de datos de rutas biológicas de autores expertos, revisada por pares y verificada manualmente, además de haberse usado ampliamente en otras investigaciones [Glaab et al., 2010; Jassal, 2011].

En las siguientes secciones se describe el sistema ERR y se presentan los resultados de la extensión de Reactome, así como el análisis de las proteínas predichas, tanto desde un punto de vista estadístico, como desde su interpretación a partir de sus anotaciones funcionales. Además, ERR se compara con otros métodos del estado del arte que pueden resolver la misma tarea, pero basados en una información de entrada diferente (sólo secuencia o sólo redes de interacción).

7.2. Diseño/Materiales y Métodos

En esta sección se describe cuáles son las fuentes de datos originales, cómo se representa y agrupa dicha información para aplicar algoritmos de aprendizaje relacional, y cómo obtener y aplicar un sistema que anote proteínas humanas con rutas metabólicas, de señalización y de regulación.

7.2.1. Recopilación de Fuentes Originales de Datos

Para el desarrollo del sistema de predicción se ha recopilado información de diferentes fuentes para construir un conjunto de datos propio, dado que no existe ninguno sobre el que aplicar aprendizaje para afrontar este problema. Éste incluye tanto características individuales asociadas a las secuencias proteínicas y génicas, como relaciones entre proteínas.

Las secuencias de proteínas provienen de Ensembl [Hubbard et al., 2009], en concreto de la Enciclopedia de genes y variantes de genes (consorcio GENCODE) [Harrow et al., 2006] (tomadas de la versión 3c de Marzo del 2010). Se trata de transcritos de Ensembl verificados manualmente y producidos por el grupo HAVANA (del inglés, *the Human and Vertebrate Analysis and Annotation*) del instituto *Welcome Trust Sanger*, que forma parte del consorcio GENCODE. Tomando como entrada estas secuencias de aminoácidos en formato FASTA, se calculan 3 propiedades numéricas asociadas a la secuencia de proteína (longitud, carga positiva y carga negativa) usando la herramienta BioWeka [Gewehr et al., 2007]. Se incluyen predicciones simples sobre las secuencias de proteínas (en concreto, si la proteína contiene algún dominio transmembrana, de señal o de hélice super-enrollada (del inglés, *coil-coiled*) provenientes de Ensembl versión 56 [Hubbard et al., 2009] (a través de BioMart [Smedley et al., 2009]). Las características de la secuencia génica (nombre del cromosoma, longitud, orientación del gen en el cromosoma y contador de transcritos o isoformas) se extraen también de la misma versión de Ensembl.

Sobre los datos relacionales, se incluyen dos tipos de relaciones entre proteínas: interacciones proteína-proteína y complejos de proteínas, ambas representadas como parejas de proteínas. El primer grupo de datos relacionales consiste en pares de interacción proteína-proteína, extraídos del repositorio BioGRID (versión 2.0.59) [Stark et al., 2006], que integra las bases de datos de interacción más importantes como MINT [Chatr-aryamontri et al., 2007], IntAct [Hermjakob et al., 2004b] y HPRD [Peri et al., 2003]. Se seleccionan los pares de BioGRID sólo de relaciones binarias físicas, identificadas por los códigos de evidencia *Co-crystal structure, Far Western, FRET, PCA* y *Two-Hybrid*. Los complejos de proteínas conforman el segundo grupo de relaciones. Cada complejo o grupo de proteínas se considera como un conjunto de pares de proteínas, ya que las bases de datos de complejos los representan

como pares independientes, y la información disponible para reconstruir el complejo no es completa ni está verificada para muchos casos. Los datos sobre complejos se extraen de la misma versión de BioGRID, seleccionando en este caso las relaciones identificadas por los códigos de evidencia *Affinity Capture, Co-purification* y *Reconstituted complex*.

Finalmente, se toman los datos referentes al objetivo de anotación, es decir, las rutas biológicas de Reactome [Matthews et al., 2009]. En concreto, 37 de las 52 rutas de alto nivel para humanos, en su versión 30. Estas 37 rutas se corresponden con aquellas que alcanzan un tamaño mínimo de al menos 32 proteínas en la ruta original, mínimo necesario para aprender con el sistema ERR.

En resumen, se recopilan 22.304 genes, 72.731 proteínas isoformas, 229.407 pares de interacción proteína-proteína, 478.420 pares de interacción en complejos y 37 rutas, con una media de 142 proteínas no redundantes por ruta. La sección 7.2.4 describe en detalle el lenguaje de representación del conocimiento específico para todos estos datos.

Las fuentes de datos empleadas usan diferentes identificadores de genes y proteínas. Para unificarlos, todos los identificadores originales se mapean a los identificadores de Ensembl (de proteína, *ENSP*, o de gen, *ENSG*), usando el sistema de referencias cruzadas de BioMart [Smedley et al., 2009].

7.2.2. Representación del Conocimiento

En esta sección se presenta la conversión del modelo E/R genérico del BioRepositorio multi-relacional (propuesto en esta tesis en la sección 5.3) para su aplicación a este problema específico de extensión de rutas biológicas, siguiendo las indicaciones de la sección 5.4.

Partiendo del modelo E/R global de la figura 5.1 para obtener el modelo E/R específico de la figura 7.1 se utilizan bastantes entidades y relaciones de cada uno de los tipos generales (ver sección 5.3), para cubrir la diversidad y restricciones del problema de extensión de rutas en humanos.

Se define una entidad *Proteína* y otra *Gen*, con sus propiedades individuales cada una, y una relación *se expresa en* (ver tabla 5.1) con cardinalidad *uno a muchos* en humanos, representando la presencia de isoformas. La relación *IPP* es una simplificación de una relación *pertenece a* de la entidad *Proteína* con un *Grupo Binario* de dos proteínas que interaccionan, como en el problema descrito en el capítulo 6. Lo mismo sucede para la relación binaria *homólogo*, que representa dos proteínas humanas con alta similitud de secuencia.

Del tipo de entidad *Grupo N-ario* del modelo global, en este modelo particular se definen dos entidades diferentes: *Complejo* y *Ruta Biológica*, cada una con su relación de tipo *pertenece a* (*forma parte de* y *participa en*, respectivamente) con la entidad *Proteína*. La relación *participa en* representa el objetivo de predicción, y por ello aparece marcado en rojo en la figura 7.1.

Del tipo de entidad *Anotación* se definen dos entidades principales (*Dominio* y *Anotación Multi-valuada*), con varias sub-entidades asociadas a cada una (ver figura 7.1). A su vez, cada una está relacionada con la entidad *Proteína* mediante una relación de tipo *anotada con* (*tiene* y *anotada con*, respectivamente), heredadas por sus sub-entidades. Las 3 sub-entidades de *Dominio* pueden existir a la vez, asociando un valor booleano a la proteína. De la misma forma, las 3 sub-entidades de *Anotación Multi-valuada* (*GO*, *Pfam* e *InterPro*), pueden tomar varios valores diferentes de cada una de las cinco categorías inferiores para una misma proteína, por lo que no basta con un simple atributo asociado a la entidad *Proteína*.

Mientras que en este capítulo las relaciones usadas se limitan a interacciones proteína-proteína y en complejos, las entidades y relaciones asociadas a *Anotación Multi-valuada* y

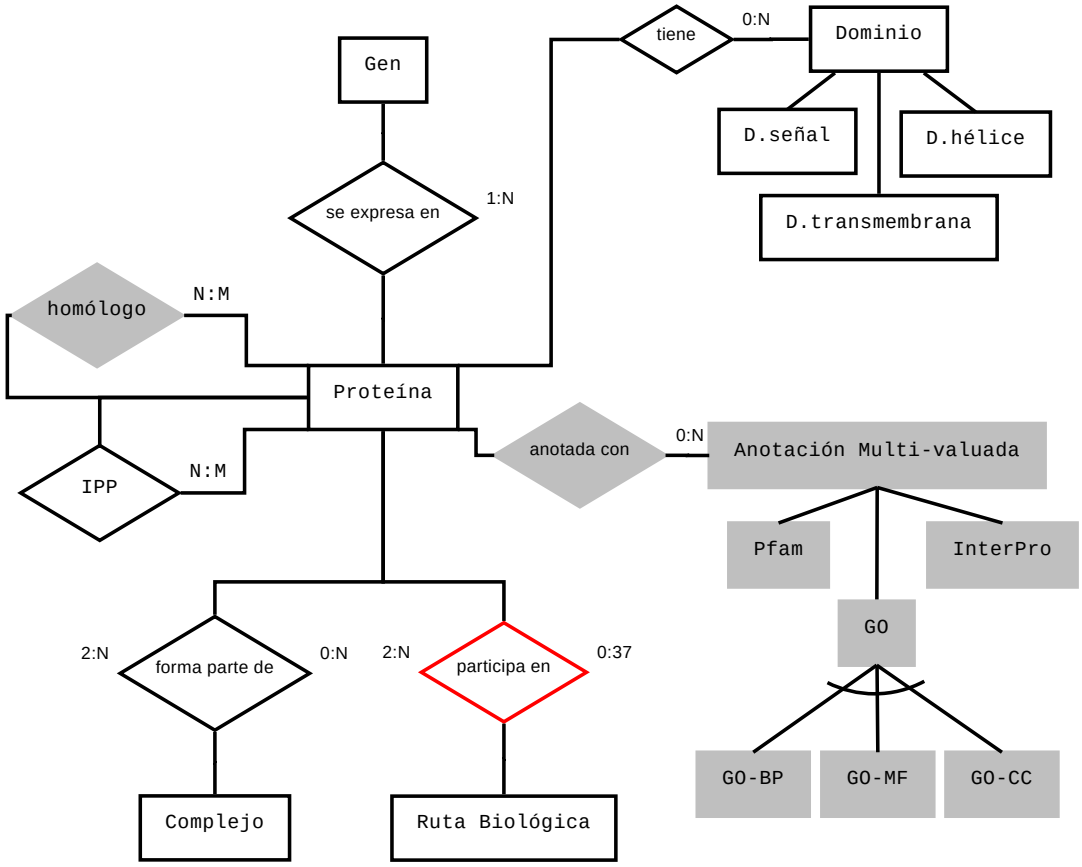


Figura 7.1: Modelo Entidad/Relación para extensión de rutas biológicas en humanos. En rojo, el objetivo de predicción. Sombreado en gris, las entidades y relaciones que sólo se utilizan en el capítulo 8.

homólogo, sombreadas en gris en la figura 7.1, se utilizan en el capítulo 8. En particular, en las secciones 8.5 y 8.6 se realiza un análisis de la influencia de la información relacional en el aprendizaje para anotación funcional, empleando estas otras fuentes de relaciones adicionales.

7.2.3. Construcción de Conjuntos de Datos

En las siguientes sub-secciones se detalla la construcción de los conjuntos de datos que se necesitan para resolver el problema planteado con aprendizaje automático. Por un lado, el conjunto de entrenamiento y el de test, etiquetados con las anotaciones de Reactome, con los que se construye y valida el modelo. Por otro lado, el conjunto de aplicación, con las proteínas no anotadas, con las que se extienden las rutas de Reactome.

Conjuntos de Entrenamiento y Test

Dado que la hipótesis de partida es predecir rutas basándose sólo en características de la secuencia, no en homología, el objetivo en este trabajo es capturar propiedades 'funcionales' de las secuencias, no similitudes entre cadenas de aminoácidos. Por lo tanto, se construye un conjunto de datos no redundante. De esta forma, en el proceso de aprendizaje se evitan

sesgos en la evaluación del rendimiento de la predicción, debidos a relaciones indirectas entre proteínas similares en los conjuntos de entrenamiento y test, comentadas en [Hobohm et al., 1992] y en el capítulo 8, en la sección 8.8. La reducción de redundancia es un proceso conservador típico, siendo la mejor opción cuando la relación entre el origen evolutivo y las características de la secuencia no es fácil de determinar. Como desventaja, el tamaño del conjunto de datos decrece en un porcentaje elevado, al tener que ignorar muchas proteínas anotadas.

Se elimina la redundancia en dos aspectos distintos: isoformas y similitud de secuencia. Un mismo gen se puede expresar dando lugar a varias proteínas o transcritos, llamados isoformas, por procesos de ensamblaje alternativo. Aunque habitualmente no se utilice en la predicción de anotación, hay que tener presente que en este trabajo el número de isoformas se preserva en el aprendizaje como una característica de la secuencia génica. Sin embargo, para reducir la redundancia del conjunto de datos, se selecciona sólo una forma principal entre todas las proteínas expresadas desde el mismo gen. En este caso, se define la isoforma principal como la proteína con el número más alto de anotaciones en Reactome, ya que es el objetivo de predicción. Si varias isoformas tienen el mismo número de anotaciones, la secuencia más larga se considera la forma principal. Tras la eliminación de redundancia por isoformismo, el número de proteínas decrece de 72.731 a 3.510 isoformas principales con anotación en Reactome. Cabe destacar que más del 97 % de las isoformas con más anotaciones en Reactome coinciden con las isoformas de mayor longitud. Según los estudios publicados [Tress et al., 2008], lo razonable sería que coincidieran hasta aproximadamente un 75 % de las isoformas. Así, este 97 % podría indicar que en las bases de datos asignan explícitamente las anotaciones conocidas a la proteína isoforma de mayor longitud, aunque no se haya experimentado explícitamente con ella, incluyendo imprecisión en los datos, como ocurre con frecuencia en los problemas de Biología Molecular. Otra alternativa para seleccionar la isoforma principal sería utilizar el método APPRIS [Rodríguez et al., 2012], actualmente en desarrollo, que combina diferentes fuentes de información para una definición más fiable de la isoforma principal.

A continuación, sobre el conjunto previo de proteínas sin isoformas, se realiza una reducción por similitud de secuencia, basada en alineamientos BLAST [Altschul et al., 1990]. Se utiliza uno de los dos algoritmos de Hobohm [Hobohm et al., 1992], originarios de la herramienta *PDBselect* [Hobohm et al., 1992], que posteriormente se han ampliado [Griep and Hobohm, 2010] y usado en estudios previos [Emanuelsson et al., 1999; Jensen et al., 2003a; Wang et al., 2009]. El algoritmo Hobohm 2 maximiza el tamaño del conjunto no redundante de salida, en un número mínimo de pasos. El algoritmo elimina toda proteína del conjunto que tiene alguna otra similar dentro de él, yendo de las de mayor similitud a las de menos. Se implementa una ligera modificación de este algoritmo, con una medida de similitud basada en secuencia en vez de en estructuras. En el algoritmo Hobohm 2 original se utiliza como medida de similitud de proteínas la función HSSP [Sander and Schneider, 1991]. Dicha función está basada en el mismo algoritmo de alineamiento de secuencias con programación dinámica [Smith and Waterman, 1981] que usa BLAST [Altschul et al., 1990], y aquí se usan los resultados de BLAST como medida de similitud. Adicionalmente, el algoritmo Hobohm 2 original aplica un umbral por homología estructural que aquí no se usa.

Otras opciones para obtener un conjunto no redundante serían usar la base de datos sin homólogos *PDB select* [Hobohm et al., 1992; Griep and Hobohm, 2010] (que busca proteínas con estructuras 3D diferentes almacenadas en PDB [Berman et al., 2000]), o las herramientas *CD-HIT* [Li and Godzik, 2006] o *RedHom* [Lund et al., 1997], que reducen un conjunto de secuencias a un subconjunto representativo con baja similitud de secuencia. Una tercera

opción, que reduce menos el tamaño del conjunto resultante, sería restringirse a eliminar las proteínas redundantes entre los conjuntos de entrenamiento y test (como se hace en [Jensen et al., 2003a]), sin necesidad de que absolutamente todas las proteínas de un mismo conjunto sean no redundantes entre sí, independientemente de dónde estén, como sí se asegura al decidir aplicar el algoritmo de Hobohm.

Se calcula la similitud de secuencia en el proteoma humano completo con BLASTP [Altschul et al., 1997] sobre las secuencias FASTA de HAVANA (ver sección 7.2.1). Los parámetros de BLASTP, diferentes de la configuración por defecto, son: 0,01 como umbral sobre el valor esperado (*e-value*), 500 secuencias de salida como máximo y BLOSUM62 como matriz de puntuación entre aminoácidos. Tras la ejecución de BLASTP, se aplica un filtro para obtener todos los pares de proteínas similares con una identidad de secuencia superior al 30 %. La reducción por similitud de secuencia sobre el conjunto de secuencias isoformas principales deja 1.654 proteínas anotadas en las 37 rutas de Reactome consideradas. Estas proteínas se dividen aleatoriamente en dos tercios para el conjunto de entrenamiento (1.108 proteínas) y un tercio para el conjunto de test (546 proteínas). Finalmente, no hay redundancia ni entre el conjunto de entrenamiento y el de test, ni dentro de cada conjunto entre sí.

Conjunto de Aplicación

Por otro lado, de las proteínas presentes en las 37 rutas de Reactome de interés, 18.794 no están anotadas (22.304 isoformas principales menos las 3.510 con alguna anotación en Reactome).

Igual que el conjunto de proteínas anotadas, el conjunto de proteínas no anotadas debe ser no redundante. Primero, se eliminan todas las proteínas similares en secuencia con las proteínas de los conjuntos de entrenamiento y test, quedando 14.016 proteínas. Segundo, sobre las proteínas resultantes se ejecuta la implementación modificada del algoritmo Hobohm 2 descrita previamente, reduciendo el conjunto de aplicación a 8.187 proteínas, que no tienen similitud de secuencia, ni entre ellas, ni con el conjunto de entrenamiento ni con el de test.

Estas 8.187 proteínas no anotadas en Reactome se usan como entrada para expandir las rutas, a través del sistema de predicción ERR presentado en este capítulo. Estas proteínas son no redundantes, pero no se descarta que sus homólogos pudieran expandir la misma ruta.

7.2.4. Lenguaje de Representación del Conocimiento

En las rutas biológicas, las interacciones proteína-proteína y los complejos de proteínas son relaciones importantes. Por lo tanto, se ha decidido incluir estos tipos de interacciones en el proceso de aprendizaje, como información relacional que pueda influir en la predicción final. En el Aprendizaje Automático clásico, los datos se representan de forma proposicional. Es decir, se tiene una tabla, con una fila por proteína, y una lista de columnas (o características) para cada proteína específica. La representación proposicional de los datos recopilados en este estudio requeriría miles de atributos booleanos por proteína: uno para cada uno de los potenciales compañeros de interacción en el proteoma completo. Además, la mayoría de las columnas tomarían valor falso. Por el contrario, en una representación relacional [Dzeroski and Lavrac, 2001], es suficiente con definir un predicado binario y, en caso de que la pareja de interacción exista realmente, incluir una instancia de dicho predicado, de forma flexible. La representación relacional también permite tener en cuenta las propiedades de secuencia del compañero de interacción en el aprendizaje, a través de enlaces por su identificador. Por ejemplo, se podría anotar una *proteína A* con la ruta *tráfico de membrana*, justificando que la

proteína A tiene una interacción en un complejo con una *proteína B* que contiene un dominio transmembrana. Por lo tanto, el enfoque relacional permite una representación más intuitiva para conceptos relacionados, y facilita la inclusión de información adicional asociada a una relación entre objetos.

El principal lenguaje de representación relacional es la programación lógica, un subconjunto de la lógica de primer orden, también llamada lógica de predicados, donde cada elemento es un predicado. Todos los datos recopilados, descritos en la sección 7.2.1, se representan como predicados en sintaxis *Prolog* (ver la figura 7.2 con el lenguaje de representación completo, y la figura 7.3 con un ejemplo con datos concretos de un par de proteínas). Esta representación permite aplicar Aprendizaje Relacional.

```
protein(proteinID,length,positiveCharge,negativeCharge).
protein_class(proteinID,reactomeID).
protein_gene(proteinID,geneID).
gene(geneID,chrName,length,strand,numTranscriptsOrIsoforms).
transmembrane_domain(proteinID).
ncoils_domain(proteinID).
signal_domain(proteinID).
ppinteraction_pair(proteinID,proteinID).
complex_interaction(proteinID,proteinID).

// discretized(gene(A,B,C,D,E),[C],[5715,20226])..
// discretized(gene(A,B,C,D,E),[E],[1,3])
// discretized(protein(A,B,C,D),[B],[300,396,629]).
// discretized(protein(A,B,C,D),[C],[0.086957,0.109316,0.129964]).
// discretized(protein(A,B,C,D),[D],[0.072897,0.110656,0.133171]).
gene(+ID,W,X,Y,Z), X < 3860.
gene(+ID,W,X,Y,Z), X > 30447.
gene(+ID,W,X,Y,Z), Z = 1.
gene(+ID,W,X,Y,Z), Z > 4.
protein(+ID,X,Y,Z), X < 300.
protein(+ID,X,Y,Z), X > 629.
protein(+ID,X,Y,Z), Y < 0.086957.
protein(+ID,X,Y,Z), Y > 0.129964.
protein(+ID,X,Y,Z), Z < 0.072897.
protein(+ID,X,Y,Z), Z > 0.133171.
```

Figura 7.2: Lenguaje de representación del conocimiento en el dominio de predicción o extensión de rutas metabólicas.

Para incrementar la expresividad de los argumentos numéricos de los predicados `protein/4` y `gene/5`, se discretizan en 4 particiones (ver los resultados en las líneas '*discretized*'de la figura 7.2), y se añaden comparaciones numéricas al lenguaje de representación (ver diez últimas líneas de la figura 7.2). Así, se consigue discriminar los valores del primer cuartil (los valores más bajos) y del cuarto cuartil (los valores más altos). Estos cuartiles representan, por ejemplo, secuencias cortas (`protein(+ID,length,A,B), length <300`), secuencias cargadas positivamente (`protein(+ID,A,positiveCharge,B), positiveCharge >0.129964`) o genes con muchos transcritos (`gene(+ID,A,B,C,transcripts), transcripts >4`).

7.2.5. Método de Predicción

El sistema de Extensión basado en Representación Relacional (ERR) propuesto se divide en dos pasos: primero, extraer patrones frecuentes relacionales y, segundo, aplicar un algoritmo

```
protein('ENSP00000299992',1775,0.065916,0.092394).
length > 629          /*secuencia de proteína larga*/
posCharge < 0.086957  /*baja carga positiva*/
transmembrane_domain('ENSP00000299992').
signal_domain('ENSP00000299992').
protein_gene('ENSP00000299992','ENSG00000166763').
gene('ENSG00000166763','15',118700,-1,3).

protein('ENSP00000373536',230,0.134782,0.104348).
length < 300          /*secuencia de proteína corta*/
posCharge > 0.129964  /*alta carga positiva*/
protein_gene('ENSP00000373536','ENSG00000165863').
gene('ENSG00000165863','10',6275,-1,3).
ppinteraction_pair('ENSP00000373536','ENSP00000340995').
ppinteraction_pair('ENSP00000373536','ENSP00000363440').
ppinteraction_pair('ENSP00000373536','ENSP00000363453').
ppinteraction_pair('ENSP00000373536','ENSP00000379226').
ppinteraction_pair('ENSP00000373536','ENSP00000395815').
ppinteraction_pair('ENSP00000373536','ENSP00000410143').
```

Figura 7.3: Ejemplos de representación del conocimiento en el dominio de predicción o extensión de rutas metabólicas.

de construcción de un árbol de decisión proposicional. Esta descomposición de la predicción en dos, aunque con otras configuraciones y datos, ya se ha aplicado en otros trabajos previos relacionados con anotación funcional, con herramientas distintas [Clare et al., 2006] o iguales [Vens et al., 2008], con las diferencias que se comentan en la sección 7.1. Otra opción posible sería utilizar un método de predicción de un solo paso, bien relacional o bien proposicional, enfoques desarrollados y analizados en el capítulo 8, sección 8.4.

En el primer paso, se extraen los patrones frecuentes (es decir, una secuencia relevante de predicados) con WARMR [Dehaspe and Raedt, 1997], un algoritmo de extracción de reglas de asociación en lógica de primer orden, el cual toma como entrada un conjunto de datos relacional. Está implementado en la herramienta ACE [Blockeel et al., 2000, 2006a]. WARMR encuentra todos los patrones que satisfacen un sesgo del lenguaje y cubren una frecuencia mínima en el conjunto de datos de entrada. Hace una búsqueda por niveles, técnica similar a la del algoritmo APRIORI [Agrawal et al., 1996], rápida y eficiente para bases de datos grandes. APRIORI limita la generación de patrones con los valores de soporte (proporción de ejemplos que cumplen el patrón) y confianza. Pero siempre añade los atributos de manera indiscriminada, sin guiar la búsqueda de patrones frecuentes entre elementos relacionados, como sí permite WARMR. Sin embargo, WARMR, al seguir un enfoque relacional, permite delimitar la generación de patrones frecuentes según el tipo de los argumentos y las relaciones entre los predicados, no sólo por profundidad como en la versión proposicional (APRIORI). En este problema se aplica WARMR a las proteínas de cada ruta independientemente. De esta forma, se extraen los patrones frecuentes que caracterizan cada ruta particular. Después, se unen los patrones frecuentes para todas las rutas o se mantienen individualmente para construir el sistema predictor. Finalmente, cada patrón frecuente es una entrada del siguiente paso, la inducción del árbol de decisión, con un algoritmo proposicional, transformando cada patrón en un atributo booleano, dependiendo de si el patrón es satisfecho o no por la proteína particular.

En el segundo paso, se construyen árboles de decisión relacionales usando el sistema CLUS [Blockeel et al., 1998]. Este sistema implementa el marco de árbol de agrupación de predicción (del inglés, *predictive clustering tree framework*), que induce árboles de decisión

con un algoritmo similar a C4.5 [Quinlan, 1993], aunque viendo un árbol de decisión como una jerarquía de agrupaciones. El nodo raíz es una agrupación con todas las instancias, el cual se divide recursivamente en grupos más pequeños, de tal forma que la variación intra-agrupaciones se minimice. Con esta idea, este marco permite afrontar problemas de predicción más complejos. Se elige CLUS en vez de otros algoritmos de inducción de árboles de decisión porque CLUS nos permite realizar un aprendizaje multi-clase y multi-etiqueta fácilmente. Esto se corresponde con el presente problema de anotación de rutas, debido a que el número de posibles rutas con las que anotar una proteína es mayor que dos (multi-clase) y cada proteína podría pertenecer a más de una ruta (multi-etiqueta).

El o los árboles de decisión relacionales obtenidos después de aplicar WARMR y CLUS a los datos recopilados, permite asociar nuevas proteínas a las rutas de Reactome.

La figura 7.4 muestra un esquema del método híbrido de predicción, descrito en esta sección, aplicado sobre datos concretos, siguiendo la representación del conocimiento en predicados lógicos utilizada en este problema.

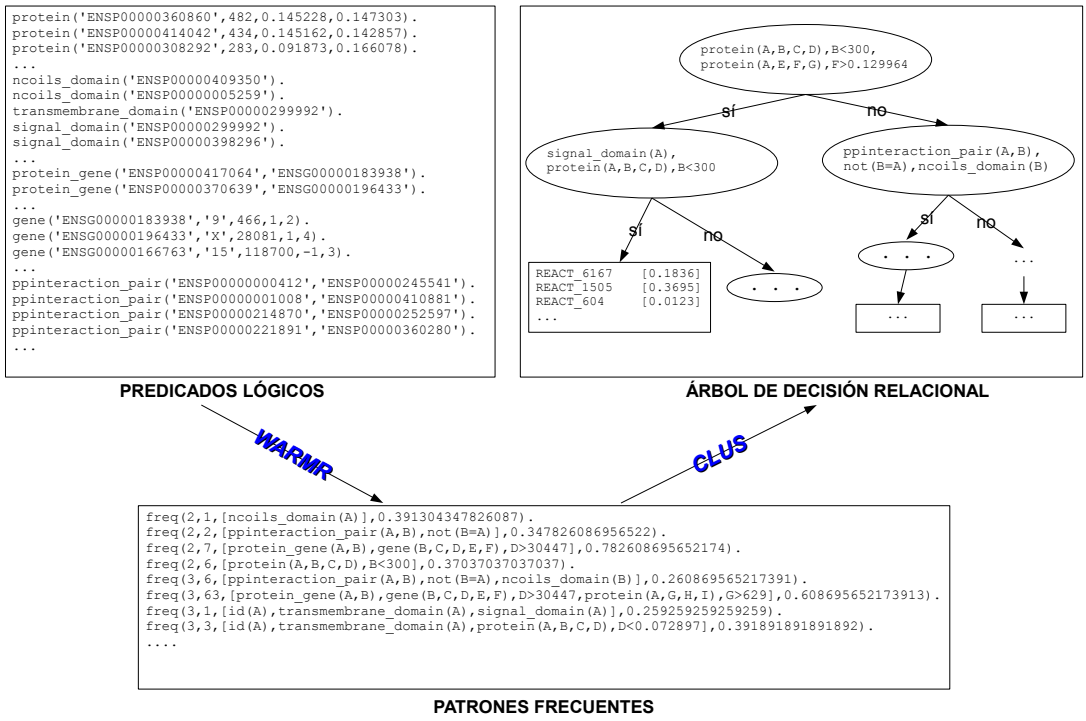


Figura 7.4: Esquema método de predicción del sistema de extensión de rutas de Reactome en humanos.

7.2.6. Aplicación a Proteínas Desconocidas

En este trabajo se expanden las rutas de Reactome aplicando el sistema diseñado a proteínas no anotadas, es decir, al conjunto de aplicación (ver sección 7.2.3). Dicho conjunto contiene 8.187 proteínas sin anotación en Reactome, no redundantes (ni isoformas, ni similares en secuencia) con ninguna proteína anotada (usadas en entrenamiento y test), ni entre sí.

El sistema de aprendizaje asocia una lista de probabilidades a posteriori a cada proteína que clasifica. Así, cada proteína tiene una probabilidad para cada ruta. Por lo tanto, se debe

seleccionar una lista de umbrales para discriminar entre qué proteínas se predicen como pertenecientes a cada ruta y cuáles no. El umbral podría ser el mismo para todas las clases, incluso tomar uno por defecto al 0,5. Pero en este dominio, como las rutas son muy diferentes entre sí, y por lo tanto los valores de probabilidad también divergen mucho, no sería razonable establecer un umbral único, que funcionaría bien en algunas clases y mal en otras. Entre múltiples opciones, en este trabajo se selecciona la siguiente combinación como criterio razonable:

Se establece la expansión de cada ruta en un máximo situado en el 20 % del tamaño de esa ruta original, sin contar las proteínas redundantes. Para cada ruta, se ordenan las proteínas del conjunto de aplicación por valor de probabilidad de predicción decreciente. A continuación, se seleccionan todas las proteínas hasta el último cambio de valor de probabilidad antes de alcanzar el 20 % del tamaño de la ruta. Si no hay un cambio de probabilidad antes del 20 %, el sistema no expande esa ruta con ninguna proteína. El umbral de predicción de cada ruta es el valor de probabilidad más bajo del conjunto de proteínas seleccionadas en el conjunto de aplicación.

7.2.7. Sistemas de Anotación

El método de predicción descrito en la sección 7.2.5 puede generar muchos sistemas de anotación diferentes, dependiendo de los parámetros de configuración. Algunos de los parámetros más relevantes son la frecuencia mínima y la profundidad máxima en WARMR, los niveles de poda en CLUS, la resolución del problema multi-clase con un único árbol de decisión multi-clasificador (como se hace en el capítulo 8, en la sección 8.2) o con N árboles binarios, la extracción de los patrones frecuentes de una o de todas las rutas a la vez (cuya influencia se analiza en el capítulo 8, en la sección 8.3), la medida de probabilidad asociada a cada predicción, etc.

En un árbol de decisión cada rama, desde la raíz a una hoja, es equivalente a una regla de decisión. En el presente problema, cada regla explica por qué una proteína se anota en esa ruta mediante una conjunción de patrones frecuentes. Así, dos reglas procedentes del mismo árbol describen formas alternativas de extender una ruta.

En esta aplicación biológica, se busca extender una misma ruta con proteínas diversas, acorde a la variabilidad molecular de las proteínas que conforman la ruta (ver sección 7.1). Ya que pueden existir criterios alternativos que relacionen la ruta original con las proteínas que la extienden (por ejemplo, localización celular común, interacción física con la ruta, implicación conjunta en causar una enfermedad, etc.), se necesita una gran diversidad de reglas entre aquellas que extienden cada ruta.

Teniendo en cuenta la versatilidad permitida por los parámetros de configuración y la diversidad de reglas buscada, después de una extensa experimentación con configuraciones diferentes, se encuentran dos soluciones distintas:

- el sistema ERR-PRyC (ERR que Prioriza el Rendimiento y la Cobertura), con el mejor rendimiento evaluado sobre el conjunto de test.
- el sistema ERR-PDR (ERR que Prioriza la Diversidad de Reglas), con un equilibrio/compromiso entre una pérdida en rendimiento y una elevada variabilidad de reglas en el conjunto de aplicación.

Se decide considerar dos sistemas alternativos porque, tras extender las rutas de Reactome con el sistema de anotación que sólo prioriza el rendimiento, la variedad de las reglas aplicadas

sobre las proteínas no anotadas no alcanza el objetivo esperado. Es decir, se necesita seguir una selección del modelo guiada también por la diversidad de reglas. A pesar de las múltiples modificaciones en la configuración, no se puede conseguir una mejora en la diversidad de reglas si ésta sólo se mide y evalúa sobre los conjuntos de entrenamiento y test, durante el proceso de aprendizaje. Esto se debe a las restricciones biológicas, tales como la casi ausencia de homogeneidad entre los conjuntos de entrenamiento y test (rutas originales) y el conjunto de aplicación (proteínas no anotadas), en términos del número de reglas aplicadas y de la cantidad de proteínas que cumple cada regla. Por lo tanto, de cara a la aplicación biológica, se decide diseñar una estrategia *ad-hoc* que incremente la diversidad de reglas a la vista del conjunto de aplicación (o de proteínas no anotadas) disponible. Para ello, entre la selección de los parámetros posibles, se elige construir un clasificador individual para cada ruta y no podar los árboles de decisión. La salida de esta estrategia es el sistema de Extensión basado en Representación Relacional que Prioriza la Diversidad de Reglas (ERR-PDR).

La diferencia de configuración más importante entre ambos sistemas de extensión es la medida de la probabilidad asociada a cada predicción. El sistema que prioriza el rendimiento sólo usa la probabilidad de salida del árbol de decisión correspondiente. Mientras que el sistema que prioriza la diversidad de reglas tiene en cuenta la similitud semántica entre las proteínas que cumplen la misma regla en el conjunto de entrenamiento, así como el número de patrones evaluados en la regla. La similitud semántica seleccionada mide la compacidad funcional a nivel molecular del fragmento de la ruta descrito por la regla específica. De esta forma, el sistema ERR-PDR prefiere reglas que cubren un grupo de proteínas coherentes en términos de propiedades moleculares, en vez de valorar exclusivamente una precisión alta. La figura 7.5 detalla la configuración de ambos sistemas.

El estimador-M al que se hace referencia en la figura 7.5 corrige la probabilidad *directa* de salida del árbol para hacer el sistema más robusto. Se aplica en todos los sistemas ERR que se presentan en esta tesis, como mejora de las predicciones frente al uso de la frecuencia de clase relativa simple. La estimación-M que se aplica es la propuesta por CLUS, donde se asume un conjunto virtual de ejemplos adicionales a la hora de calcular la probabilidad de predicción positiva de una hoja del árbol. Se considera el conjunto virtual de tamaño 1, estimando entonces cuántos son positivos con la frecuencia de esa clase en el conjunto completo de datos. Así, para cada regla, la probabilidad corregida es la suma de los ejemplos positivos que se clasifican en la hoja correspondiente más la frecuencia de clase, dividido entre la cantidad total de ejemplos que se clasifican en la hoja (acertados o no) más 1 ejemplo adicional.

Sistema ERR-PRyC

- *Extraer patrones frecuentes:*
  - Para cada ruta por separado (en vez de todas a la vez).
  - Parámetros WARMR (distintos a valores por defecto): frecuencia mínima de **0,005** y profundidad máxima de **5**.
- *Aplicar algoritmo árbol de decisión:*
  - 37 árboles binarios individuales, uno por ruta (en vez de un árbol multi-clasificador único). Estrategia *uno-contra-todos* para definir los ejemplos negativos.
  - Atributos: **iguales para todos los árboles de decisión: unión de los patrones frecuentes de todas las rutas (eliminando patrones repetidos y/o redundantes). 12.144 patrones frecuentes**.
  - Poda: no (valor por defecto en CLUS).
- *Probabilidad de pertenecer a una ruta, asociada a cada regla clasificadora:*

$$\textbf{probabilidad} = frecuencia\ relativa\ de\ clase\ corregida\ con\ estimador\ M \tag{7.1}$$

Sistema ERR-PDR

- *Extraer patrones frecuentes:*
  - Para cada ruta por separado (en vez de todas a la vez).
  - Parámetros WARMR (distintos a valores por defecto): frecuencia mínima de **0,2** y profundidad máxima de **4**.
- *Aplicar algoritmo árbol de decisión:*
  - 37 árboles binarios individuales, uno por ruta (en vez de un árbol multi-clasificador único). Estrategia *uno-contra-todos* para definir los ejemplos negativos.
  - Atributos: **diferentes para cada árbol de decisión: sólo patrones frecuentes extraídos de la ruta concreta. 247,2 patrones frecuentes por ruta, en media; con desviación típica de 117,3**.
  - Poda: no (valor por defecto en CLUS).
- *Probabilidad de pertenecer a una ruta, asociada a cada regla clasificadora:*

$$\textbf{probabilidad} = \frac{Similitud\ corregida + Frecuencia\ relativa\ de\ clase\ corregida}{2} \tag{7.2}$$

donde:

$$\begin{aligned}
\text{Similitud corregida} &= \frac{SimGO_{MF} + (SimGO_{MF} * Factor\ de\ diversidad)}{2} \\
\text{SimGO}_{\text{MF}} &= similitud\ semántica\ (en\ términos\ de\ Función\ Molecular\ de\ GO)\ calculada\ entre\ todas\ las\ proteínas\ de\ entrenamiento,\ por\ ruta\ y\ por\ regla. \\
&\quad Si\ la\ regla\ cubre\ sólo\ una\ proteína,\ SimGO_{MF} = 0. \\
\text{Factor de diversidad} &= 1 - \frac{1}{núm.\ patrones\ frecuentes\ de\ la\ regla}
\end{aligned}$$

La **Similitud semántica** se calcula como la media de similitud entre todas las combinaciones de proteínas del conjunto, por pares, con la medida de similitud de Jiang y Conrath [Jiang and Conrath, 1997], usando el paquete *GOSemSim* [Yu et al., 2010] de R.

Figura 7.5: Configuración detallada de sistemas de extensión de rutas priorizando rendimiento y cobertura (ERR-PRyC) o diversidad de reglas (ERR-PDR).

7.2.8. Esquema Resumen Sistema de Aprendizaje

La figura 7.6 resume el sistema descrito en las secciones previas, utilizado en este trabajo para extender rutas de Reactome en humanos con proteínas adicionales.

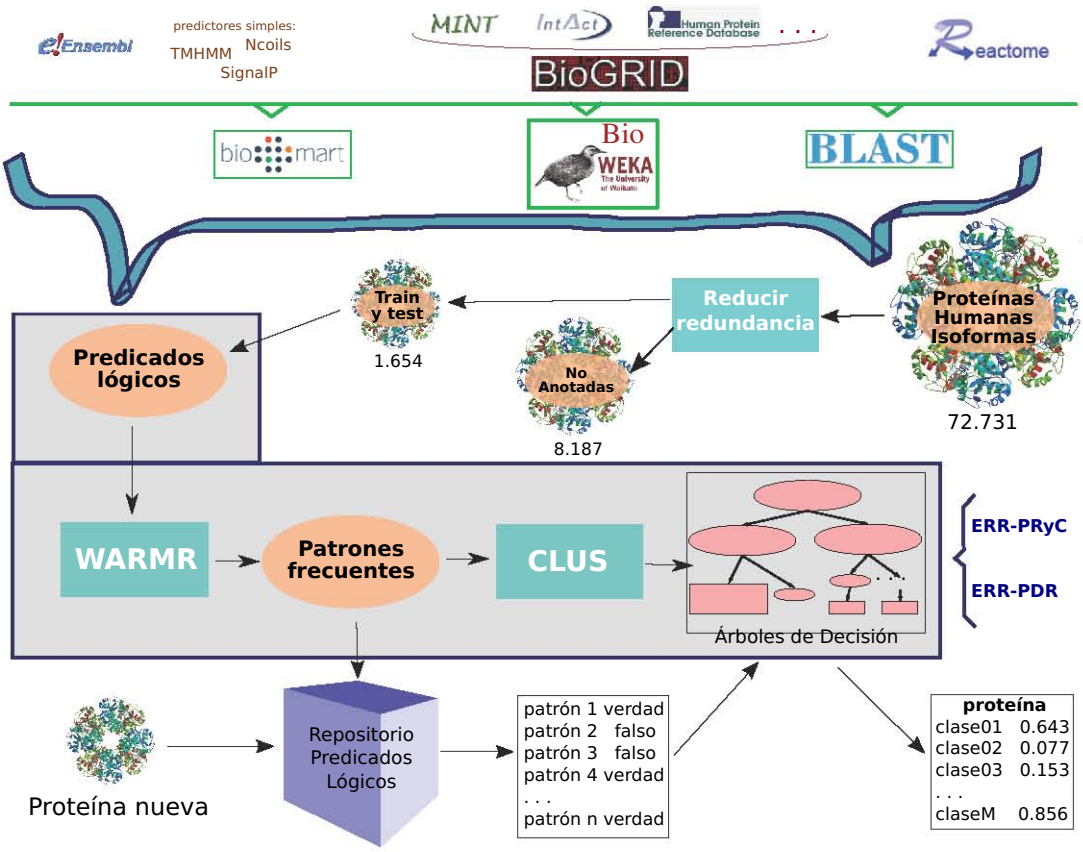


Figura 7.6: Esquema sistema de extensión de rutas de Reactome en humanos.

7.3. Resultados

7.3.1. Evaluación del Rendimiento de la Predicción

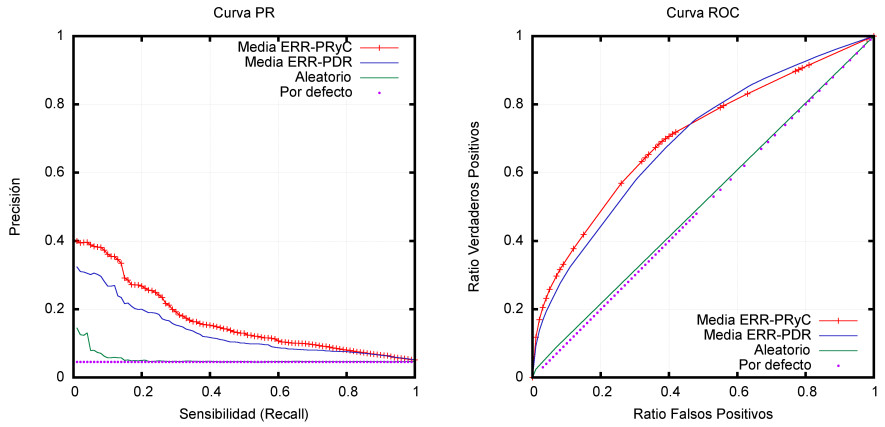
El objetivo de esta sección es analizar cómo rinden los sistemas de expansión de rutas, tanto para una visión global como para cada ruta independiente.

Este análisis computacional de los resultados de aprendizaje se evalúa con las curvas PR (ver capítulo 4). Esta medida de evaluación concuerda con la distribución de clases altamente sesgada de este problema y con un mayor interés en este dominio por las predicciones positivas frente a las negativas (es decir, proteínas que expanden alguna ruta frente a las que no expanden ninguna) [Davis and Goadrich, 2006]. Sin embargo, también se presenta la evaluación con la curva ROC equivalente para los lectores habituados a ella.

Evaluación Global

Se prefiere la *media-macro* frente a la *media-micro* [Yang, 1999; Sebastiani, 2002] a la hora de combinar, en una medida global, las medidas individuales de rendimiento para las 37 rutas de Reactome. Esto significa dar preferencia a la media de todas las áreas bajo la curva (*media-macro*) frente al área bajo la curva de la media de todas las rutas (*media-micro*). Se selecciona esta opción porque la *media-macro* no sesga el resultado hacia las clases más frecuentes, proporcionando una visión homogénea de los resultados para todas las clases [Vens et al., 2008]. Para un análisis más detallado sobre la influencia de elegir la *media-micro* o *media-macro* en problemas multi-clase consultar las secciones 4.2.1 y 8.2.3.

La figura 7.7(a) muestra la *media-macro* de las curvas PR para los sistemas ERR-PRyC y ERR-PDR, el clasificador aleatorio y el clasificador por defecto. El clasificador por defecto se corresponde con un árbol de decisión de una única hoja, el cual proporciona las frecuencias de clase como probabilidades de predicción, para cualquier proteína dada. La figura 7.7(b) representa las curvas ROC equivalentes, donde la diferencia entre ambos sistemas es menor que en las PR.



(a) Curvas PR. (b) Curvas ROC.

Figura 7.7: Curvas media global para todas las rutas: (a) curvas PR y (b) curvas ROC.

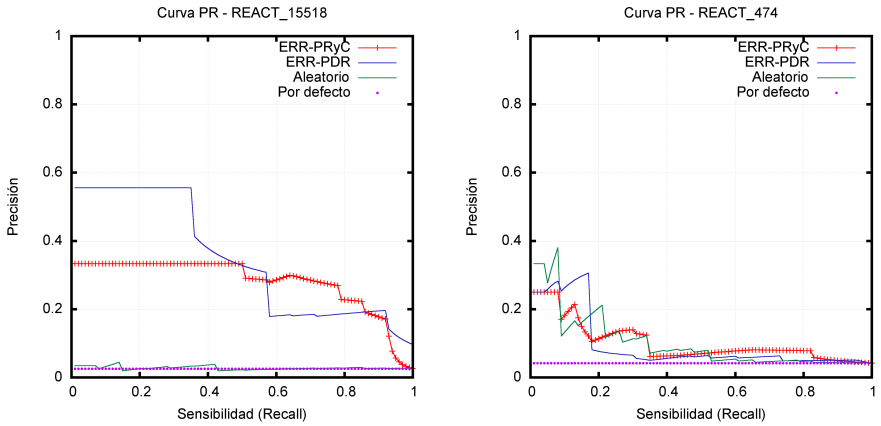
Como medida cuantitativa única de rendimiento se elige el área bajo la curva (AUC, del inglés, *Area Under Curve*) de la PR, y de forma alternativa también de la ROC. Así, en términos cuantitativos medios, los sistemas ERR-PRyC y ERR-PDR logran un AUPRC de 0,1695 y 0,1337, respectivamente, y un AUROC de 0,7028 y 0,6914.

Aunque los valores de AUPRC y AUROC no son muy altos (aún siendo mayores que los de la expansión de rutas aleatoria y por defecto), para la evaluación de estos resultados se debe tener en cuenta la definición de las rutas originales (ver discusión en sección 7.8).

Evaluación por Clase/Ruta

Cuando se analiza el comportamiento por ruta o clase independiente, se puede observar que existe una gran variedad en los niveles de rendimiento, siendo la expansión de unas rutas mejor predicha que la de otras; como se muestra en el anexo D, en las tablas D.1 y D.2 detalladas por ruta, para el sistema ERR-PRyC y ERR-PDR, respectivamente. Así, 16 rutas tienen un AUPRC mayor que la media (extensiones de rutas de alta fiabilidad), y 2 ó 3 rutas presentan un AUPRC

peor que la extensión aleatoria o por defecto correspondiente a esa ruta (extensiones con baja fiabilidad, sombreadas en rosa en las tablas D.1 y D.2). Las figuras 7.8(a) y 7.8(b) presentan dos ejemplos de curvas PR para rutas individuales, con una extensión de alta y baja fiabilidad, respectivamente.



(a) Transporte transmembrana de moléculas pequeñas *(REACT 15518)*.

(b) Metabolismo de carbohidratos *(REACT 474)*.

Figura 7.8: Curvas PR: (a) ruta individual extendida con alta fiabilidad y (b) ruta individual extendida con baja fiabilidad.

El sistema ERR-PRyC es mejor que el ERR-PDR en rendimiento global. Sin embargo, por rutas individuales, el orden del AUPRC de las rutas varía entre ambos sistemas. Además, algunas rutas las extiende mejor el sistema ERR-PDR que el ERR-PRyC, principalmente para baja cobertura, como muestra la figura 7.8(a). Incluso en el sistema ERR-PDR una misma ruta puede estar por encima de la media y en el ERR-PRyC por debajo, o viceversa.

7.4. Interpretación

El objetivo de esta sección es interpretar los resultados de extensión de rutas desde una colección de enfoques variada, tales como conocer: la relación entre la precisión de la predicción y el tamaño de la ruta, los predicados más relevantes para aprender, la cobertura y diversidad de las proteínas añadidas, la cantidad de proteínas que extienden y conectan varias rutas, la similitud semántica de las nuevas proteínas con respecto a las originales y la aportación del aprendizaje relacional para interpretar las características de las proteínas predichas.

Antes de pasar a la descripción detallada por apartados, cabe destacar cuál es la forma en que se exponen los resultados en toda esta sección. Hay que tener en cuenta que se han desarrollado dos sistemas alternativos para expandir rutas, cuyas diferencias fundamentales son:

- El sistema ERR-PRyC es mejor que el ERR-PDR en rendimiento y cobertura (evaluación con curvas PR).
- El sistema ERR-PDR es mejor que el ERR-PRyC en cantidad de reglas diferentes por ruta, y en nº de rutas extendidas.
- Existe mucha variación entre las rutas expandidas por uno y otro sistema.

Aparte de estas tres, de forma global (es decir, en media para todas las rutas) no se pueden extraer más conclusiones interesantes al comparar los dos sistemas, en ninguno de los ámbitos de interpretación (relevancia de atributos, predicados más frecuentes, similitud semántica funcional, etc.). Esto hace que no tenga sentido describir comparaciones genéricas para los dos sistemas. El resto tendrían que ser comparaciones detalladas ruta a ruta para cada uno de los ámbitos, que harían tediosa la lectura de este documento. Por lo tanto, se omite cualquier comparación exhaustiva para cada una de las rutas, y se describen sólo los resultados para uno de los dos sistemas, asumiendo que se tienen presentes las tres diferencias fundamentales con el otro. Se elige el sistema ERR-PDR, porque la mayoría de las interpretaciones tienen un enfoque de utilidad biológica, para lo cual es más conveniente una alta cantidad y diversidad de reglas por ruta, que una mayor cobertura. No obstante, esporádicamente, se comenta algún caso particular cuando se considera relevante.

Ante un interés concreto del lector por alguna ruta, se pueden tomar las dos figuras correspondientes, que se muestran en todos los análisis, y extraer fácilmente la conclusión buscada. Por ejemplo, para responder la cuestión *"a la hora de predecir la pertenencia a la* `ruta X`*, ¿varían las características más relevantes entre el sistema ERR-PRyC y el ERR-PDR?"*, bastaría con observar la figura G.2 del anexo G y la figura 7.10 de este capítulo, y comparar la fila de la `ruta X` en el sistema ERR-PRyC y la fila correspondiente en el ERR-PDR, observando si los puntos de mayor tamaño se encuentran en las mismas columnas y si son del mismo color; o para la pregunta *"¿baja o sube la similitud semántica de las proteínas que extienden la* `ruta Y` *al usar el sistema ERR-PDR en vez del ERR-PRyC?"*, lo más sencillo sería consultar las tablas detalladas por ruta del anexo D.

7.4.1. Relación entre Precisión y Tamaño de la Ruta

Como se puede observar en la figura 7.9, cuanto mayor es el tamaño de la ruta (cantidad de proteínas que la forman, es decir, barra horizontal más larga en la figura 7.9), más fiable es la predicción en términos de AUPRC (es decir, más abajo está la ruta en la figura 7.9). Aunque con excepciones, tales como rutas pequeñas con AUPRCs mayores que la media. Por ejemplo, en el sistema ERR-PDR, en las rutas *Interacciones de las integrinas en la superficie celular* (8) y *Transporte transmembrana de moléculas pequeñas* (16), con un tamaño no redundante menor a 50 proteínas, y con AUPRC tan alto como el de rutas con más de 100 proteínas.

En el anexo G, la figura G.1 es la equivalente para el sistema ERR-PRyC.

7.4.2. Análisis de Predicados Relevantes en el Aprendizaje

En esta sección, para cada ruta, se buscan las propiedades más importantes en el proceso de aprendizaje, con el objetivo de definir los elementos que más influyen para extender cada una de las rutas.

Para lograr este propósito, se emplean las mismas dos medidas de relevancia que en el método de predicción *ProtFun* [Jensen et al., 2003a]. La primera consiste en evaluar el rendimiento después de entrenar el sistema usando cada uno de los predicados individualmente. Los círculos rojos representan estos AUPRCs en la figura 7.10. La segunda medida es la pérdida en AUPRC al eliminar un predicado particular. Entonces, el rendimiento de entrenar el sistema sin un predicado se resta del AUPRC original, obteniendo el rendimiento de la combinación de todos los predicados juntos. La figura 7.10 visualiza estas diferencias en AUPRC en círculos morados. La primera medida sólo pone de manifiesto si un predicado es relevante por sí mismo; pero no tiene en cuenta si un predicado es importante cuando se usa en combinación con otros,

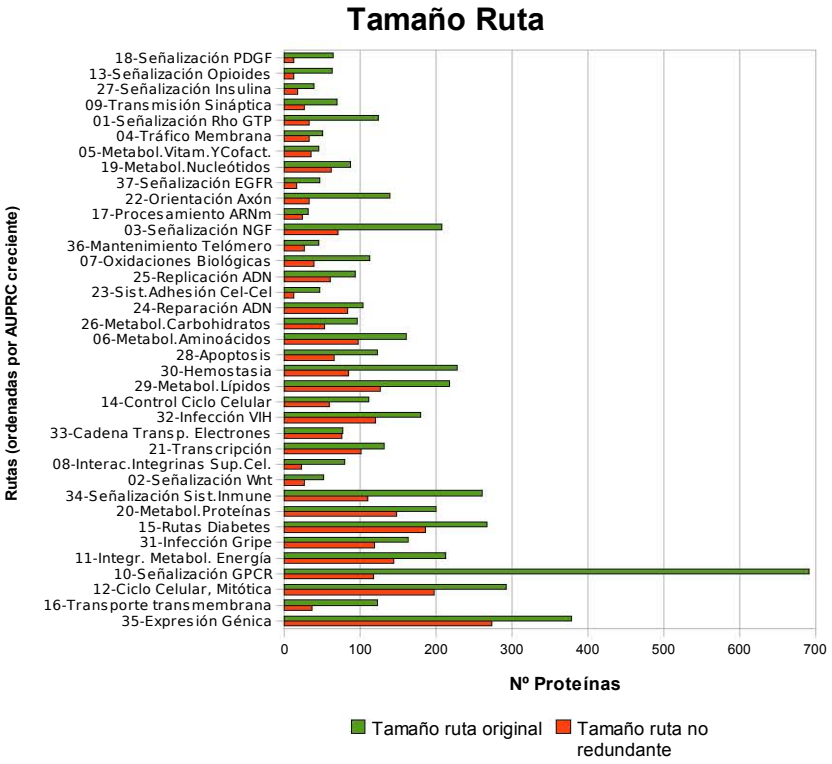


Figura 7.9: Análisis de rendimiento frente a tamaño de ruta. Sistema ERR-PDR. Rutas ordenadas de arriba a abajo, de menor a mayor AUPRC. Las barras verdes representan la cantidad de proteínas de la ruta original, y las barras naranjas la cantidad de proteínas tras eliminar las redundantes entre sí (las que se usan en el aprendizaje).

lo cual se revela con la segunda medida. Como ambas medidas son complementarias entre sí, un predicado se considera relevante cuando alguna de las dos medidas alcanza un valor alto [Jensen et al., 2003a].

Debido a que se usa una representación relacional, como entradas al sistema de aprendizaje se tienen predicados en vez de atributos. Por lo tanto, las columnas de la figura 7.10 se corresponden con predicados lógicos. Cada columna representa un predicado individual, excepto las columnas 5ª a 7ª que son los argumentos numéricos del predicado `protein` (ver los argumentos de los predicados en la figura 7.2), y las columnas 12ª y 13ª del predicado `gene`. Las columnas 'Protein/4' y 'Gene/5' incluyen las comparaciones numéricas de todos sus argumentos. 'Ambas interacciones' es un agregado de los dos tipos de interacciones consideradas (las dos columnas previas).

Al analizar la figura 7.10, se observa que las propiedades relevantes para cada ruta son diferentes. En la predicción *'media'* de rutas (situada hacia el medio de la figura 7.10) no aparece ningún predicado más importante que los otros, por lo que se puede concluir que ninguna característica contribuye más que el resto. Lo que es más, la relevancia es más clara conforme mejora la fiabilidad de la predicción, es decir, conforme se baja en el gráfico. Esto significa que, por un lado, en los casos malos (situados en la parte superior de la figura 7.10, generalmente peores que la aleatoriedad o por defecto), ningún predicado es

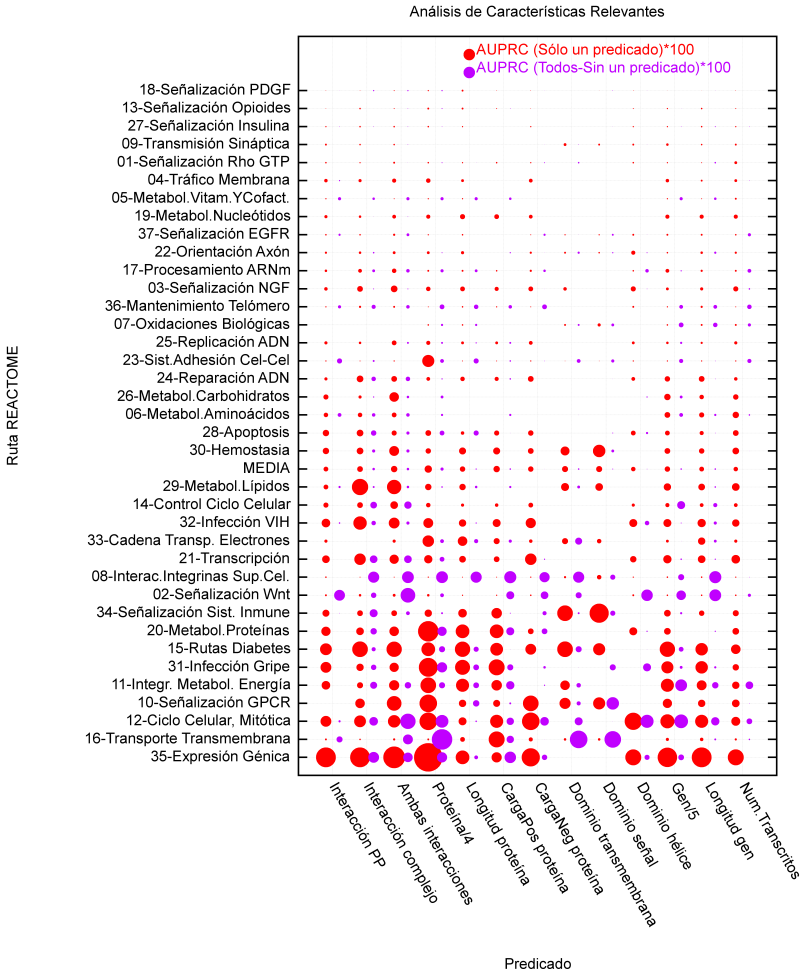


Figura 7.10: Análisis de predicados relevantes en el aprendizaje. Los círculos rojos (izquierda) representan una propiedad relevante por sí misma. Los círculos morados (derecha) representan una propiedad relevante en combinación con otras. Las rutas (filas) están ordenadas, de abajo a arriba, de mejor a peor AUPRC, según el sistema. Sistema ERR-PDR.

relevante. Por ejemplo, en la ruta *Tráfico de membrana* (04) el predicado que representa un dominio transmembrana (`transmembrane_domain`) debería ser relevante para el sistema de aprendizaje, pero no lo es. Por otro lado, en los casos buenos (rutas de la parte inferior de la figura 7.10, por debajo de la media) hay diferencias obvias de relevancia entre predicados distintos de la misma ruta. El predicado más importante (la columna con círculos más grandes) es `protein/4`, que es un agregado de las características que más discriminan: `protein length` y `protein positive charge`. Por el contrario, las interacciones (las tres primeras columnas) no son propiedades tan fundamentales en el proceso de aprendizaje.

Cabe destacar que el tamaño de los puntos es mayor en el sistema ERR-PRyC (ver en el anexo G, la figura G.2). Es decir, la relevancia de propiedades es más acusada en el sistema ERR-PRyC que en el ERR-PDR, pero la diferenciación entre atributos relevantes y no relevantes está más definida en el segundo.

Si se analizan algunos casos específicos en el sistema ERR-PDR, en la ruta *Expresión génica* (35) cada uno de los predicados aisladamente alcanza una buena predicción en casi

todos los casos (se observa en la figura 7.10 que casi todos los círculos rojos de la última fila son de gran tamaño). Mientras que en *Transporte transmembrana* (16), *Interacciones de las integrinas en la superficie celular* (08) y *Señalización de Wnt* (02) casi todos los predicados dependen entre sí, sin contribuir de forma independiente (se ven varios círculos morados de tamaño similar en las filas correspondientes a estas rutas). En la ruta *Transporte transmembrana* (16), los más relevantes son la carga positiva (por si sola), y el dominio transmembrana y el de señal (ambos en combinación con el predicado `protein`). Las interacciones, principalmente en complejos (segunda columna), por sí solas comprenden la información más importante para predecir la ruta *Metabolismo de lípidos* (29); aunque en el sistema ERR-PRyC todos son igual de relevantes. Por su parte, en la ruta *Infección VIH* (32), las interacciones en complejos seguidas de la carga negativa son las características fundamentales en el sistema ERR-PDR, mientras que en el ERR-PRyC no se detecta la relevancia de la carga negativa (ver figura G.2).

7.4.3. Cobertura y Diversidad en la Extensión de Reactome

En esta sección se evalúa la extensión de las rutas originales con proteínas no anotadas, tras la aplicación de los sistemas de predicción presentados.

Al aplicar el procedimiento descrito en la sección 7.2.6 a las proteínas sin anotación en Reactome, y no redundantes con las anotadas (ver sección 7.2.3), el sistema ERR-PRyC extiende 18 rutas y el ERR-PDR 28, a pesar de que sus AUCs sean más bajos, como muestra la tabla 7.1. Las 37 rutas originales tras la eliminación de redundancia se componen de 2.762 proteínas, siendo distintas entre todas las rutas 1.654 proteínas. Recordando la importancia de la diversidad entre las reglas que expanden la misma ruta (ver sección 7.2.7), se debe remarcar que el sistema ERR-PDR logra aplicar varias reglas en 15 rutas, comparado con las escasas 5 rutas por parte del sistema ERR-PRyC. Por lo tanto, el sistema ERR-PDR consigue una mayor cobertura de expansión y una alta variabilidad molecular, de las proteínas que añade a la misma ruta, para más de la mitad de aquellas que extiende.

Cada una de las proteínas que predicen los sistemas ERR comparten propiedades de secuencia con una o unas pocas proteínas de la ruta original, no con todas, ni en todas sus propiedades. Este comportamiento concuerda con la definición real de las rutas biológicas, que realizan una función común a nivel de proceso, pero sus propiedades a nivel de secuencia son diferentes.

Tabla 7.1: Evaluación numérica de la extensión de Reactome por sistema ERR-PRyC y ERR-PDR.

Sistema	**AUPRC**	**AUROC**	**nºrutas** (total/>1regla)	**nºproteínas añadidas** (total/distintas)
ERR-PRyC	0,1695	0,7028	18 / 5	249 / 218
ERR-PDR	0,1337	0,6914	28 / 15	383 / 329

En el anexo D se pueden consultar las tablas D.1 y D.2, para el sistema ERR-PRyC y ERR-PDR respectivamente, donde se detalla el número de proteínas predichas frente al tamaño original de cada ruta, entre otros valores.

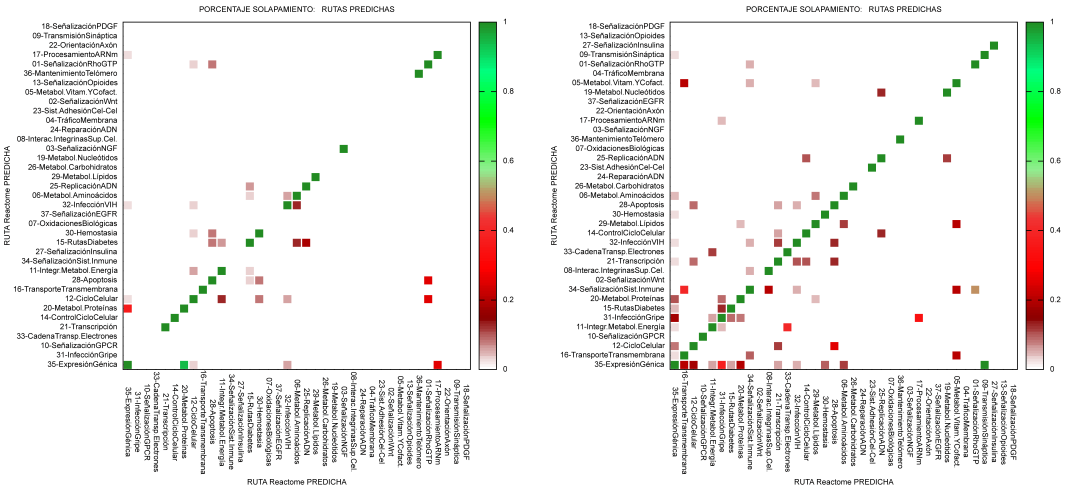
7.4.4. Solapamiento entre Rutas

En este apartado se analiza la diversidad entre proteínas que extienden distintas rutas. El objetivo es conocer si el sistema ERR predice proteínas específicas para cada ruta, o si se trata

de proteínas de características y funciones genéricas presentes en varias rutas.

En primer lugar, se puede calcular el valor absoluto de las proteínas que extienden más de una ruta. A partir de la última columna de la tabla 7.1 se extrae que un 87,50 % y un 85,90 % de las proteínas añadidas a las rutas son diferentes, para el sistema ERR-PRyC y el ERR-PDR, respectivamente. Esto quiere decir que existe menos de un 15 % del conjunto de proteínas añadidas que expande más de una ruta, en ambos sistemas.

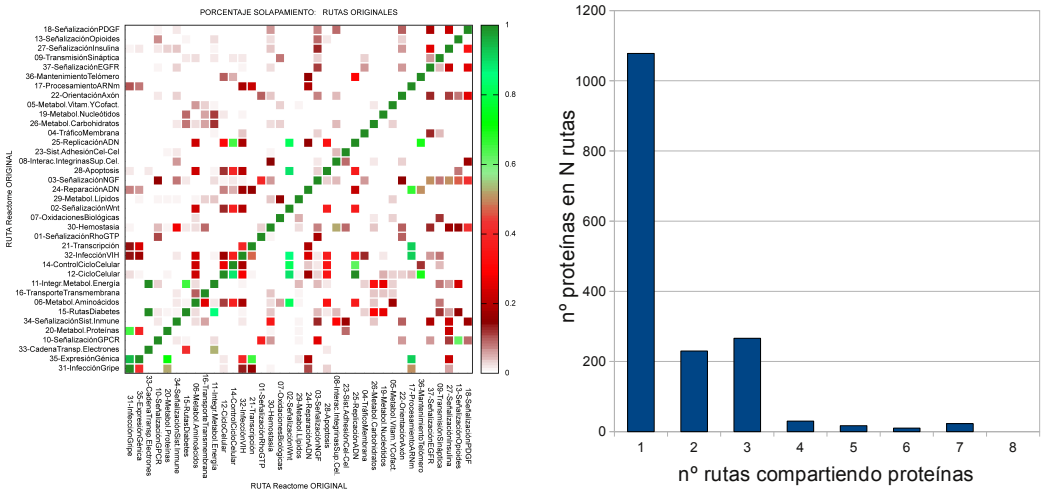
Si ahora se analiza el solapamiento de forma relativa, y se calcula la media del porcentaje de proteínas solapadas con otra ruta, siguiendo un enfoque todas contra todas las rutas, el solapamiento es 1,58 % para el sistema ERR-PRyC y 2,67 % para el ERR-PDR. En ambos casos es muy bajo, como se verifica también en las representaciones gráficas de la figura 7.11, donde el solapamiento se representa por los puntos coloreados fuera de la diagonal principal, y como se observa ahí existen muy pocos puntos diferentes al blanco. Se puede observar un solapamiento ligeramente mayor en el sistema ERR-PDR que en el ERR-PRyC. Aunque teniendo en cuenta que éste es capaz de extender diez rutas más, aumentando así la probabilidad de coincidencia, el incremento de solapamiento no es muy significativo.



(a) Sistema ERR-PRyC. (b) Sistema ERR-PDR.

Figura 7.11: Porcentaje de solapamiento entre rutas. (a) sistema ERR-PRyC y (b) sistema ERR-PDR. Las rutas están ordenadas por AUPRC creciente según cada sistema, de izquierda a derecha en el *eje x* y de abajo a arriba en el *eje y*. Cada celda representa el porcentaje (según el código de colores mostrado a la derecha) de proteínas añadidas a la ruta del *eje x* que también extienden la ruta correspondiente del *eje y*. Situación ideal (sin solapamiento): sólo diagonal en verde, resto blanco.

Es importante destacar este reducido porcentaje de proteínas comunes entre diferentes rutas en las extensiones de ERR, dado el gran nivel de solapamiento existente entre proteínas de las rutas originales, como muestran los gráficos de la figura 7.12. Es decir, se consigue predecir proteínas diferentes para cada clase o ruta, aunque las rutas originales no sean disjuntas.



(a) Porcentaje de solapamiento.

(b) Solapamiento absoluto.

Figura 7.12: Solapamiento entre rutas originales. (a) Porcentaje de solapamiento: cada celda representa el porcentaje de proteínas de la ruta del *eje x* que también están en la ruta correspondiente del *eje y*. (b) Solapamiento absoluto: el *eje y* indica el nº de proteínas que están a la vez en el nº de rutas que indica el *eje x*.

7.4.5. Similitud Semántica en la Extensión de Reactome

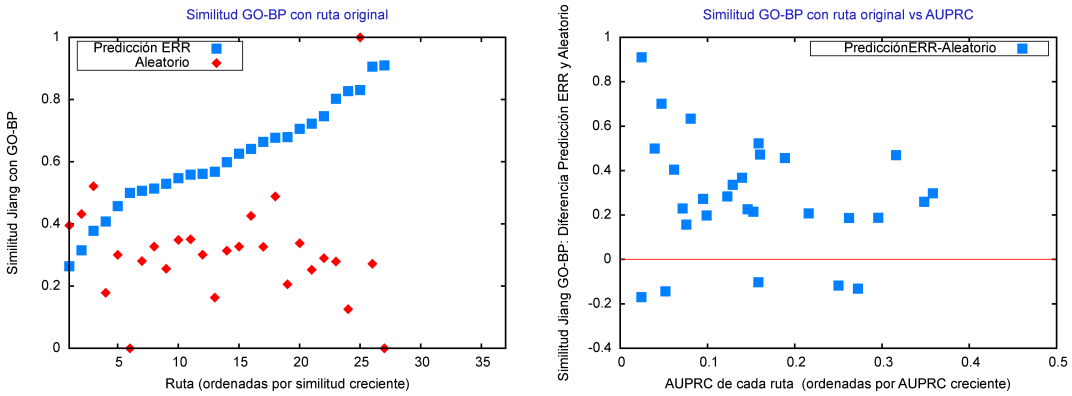
Una aproximación sencilla para conocer si las proteínas añadidas son biológicamente similares a las de la ruta que extienden, es llevar a cabo un análisis de similitud semántica entre las anotaciones funcionales de las proteínas en una base de datos externa (ver descripción del concepto de anotación funcional y derivados en B.1, 2.4.1 y B.2).

Dado que todas las proteínas en una ruta biológica están relacionadas funcionalmente a un nivel de proceso, se utilizan los términos de anotación definidos en la ontología Proceso Biológico de GO [Ashburner et al., 2000], *GO-BP* (extraídos de la versión 56 de Ensembl [Hubbard et al., 2009]). Se escogen todos los códigos de evidencia excepto *ISS* (del inglés, *Inferred from Sequence or Structural Similarity*), para evitar relaciones circulares o indirectas, derivadas de anotaciones inferidas por similitud de secuencia o de estructura. En este caso, no existe un problema de falta de ortogonalidad entre los términos de *GO-BP* y los de Reactome, porque las proteínas añadidas por el sistema ERR (para las que se hace el análisis de similitud semántica) no tienen anotaciones en Reactome, y por tanto no puede haber solapamiento con los términos de *GO-BP*. Se usa la medida de similitud de Jiang y Conrath [Jiang and Conrath, 1997], que adapta las medidas de la teoría de la información para establecer una distancia semántica entre términos de GO (ver sección 4.2.2), teniendo en cuenta la compleja estructura de grafo acíclico dirigido de esta ontología.

Con dicha medida de similitud, se comparan las proteínas originales de la ruta con las proteínas añadidas mediante predicción (dichas proteínas se identifican a partir de aquí en el texto como *proteínas predichas*), en términos de similitud semántica funcional a nivel de proceso. Se calcula la media del máximo de similitud de cada proteína (del inglés, *best-match average* [Pesquita et al., 2009]) de todas las combinaciones por pares de proteínas, obteniendo así un valor de similitud semántica para cada ruta. Además se comparan estas similitudes con aquellas calculadas entre las proteínas de la ruta original y una extensión aleatoria del mismo

tamaño, siguiendo el ejemplo de un trabajo previo [Glaab et al., 2010]. Se puede ver un análisis con distintas posibilidades para agregar similitudes en función del objetivo en la sección 4.2.2.

La figura 7.13(a) muestra que, en la mayoría de las 28 rutas extendidas, las proteínas predichas presentan mayor similitud semántica funcional con las proteínas de la ruta original que las proteínas seleccionadas aleatoriamente. Sin embargo, las rutas con bajo AUPRC no se corresponden completamente con las que tienen peor similitud funcional con la ruta original frente al aleatorio (ver figura 7.13(b)). Por lo que no hay correlación entre la fiabilidad de la predicción y la similitud semántica.



(a) Orden por similitud creciente.

(b) Orden por AUPRC creciente.

Figura 7.13: Similitud de anotación funcional entre proteínas de la ruta original y proteínas añadidas (por predicción y aleatoriamente). Sistema ERR-PDR. Las rutas sin extensión no se representan. (a) Rutas ordenadas por similitud creciente en el grupo de predicciones. Cada punto representa la similitud absoluta de las proteínas añadidas a la ruta original. (b) Rutas ordenadas por AUPRC creciente en el grupo de predicciones. Cada punto representa la diferencia de similitud a la ruta original entre las proteínas predichas y las proteínas aleatorias (*Sim.Predichas–Sim.Aleatorias*) para esa ruta. Así, la línea roja representa la inexistencia de mejora de las predicciones frente a la aleatoriedad, en términos de similitud.

En este análisis de similitud semántica, el sistema ERR-PRyC es ligeramente mejor, como se puede observar en los anexos en la figura G.3.

7.4.6. Interpretación de la Extensión basada en Aprendizaje Relacional

Esta sección explica brevemente qué tipo de proteínas extiende cada ruta según el sistema propuesto ERR, acorde a las reglas extraídas del árbol de decisión y su relación con las propiedades más frecuentes.

El análisis se centra en dos ejemplos de rutas que extiende el sistema ERR-PDR. Una regla extiende la ruta *Transporte transmembrana de moléculas pequeñas* (16) con 5 proteínas, que cumplen al menos las siguientes propiedades (ver figura 7.14): dominio transmembrana, secuencia larga de aminoácidos y nucleótidos, carga positiva, sin dominio de señal y con dos interacciones proteína-proteína (una de ellas con un proteína de alta carga negativa).

```
Regla 11:
=========
IF transmembrane_domain(A),protein_gene(A,B),gene(B,C,D,E,F),D>30447,
      protein(A,G,H,I),H<0.086957 = 0 AND
   transmembrane_domain(A) = 1 AND
   protein_gene(A,B),gene(B,C,D,E,F),D>30447,protein(A,G,H,I),G>629=1 AND
   transmembrane_domain(A),signal_domain(A) = 0 AND
   complex_interaction(A,B),not(B=A),protein(B,C,D,E),C<300,
      protein(B,F,G,H),G<0.086957 = 0 AND
   ppinteraction_pair(A,B),not(B=A),ppinteraction_pair(A,C),not(C=A),
      not(C=B),protein(C,D,E,F),F>0.133171 = 1
THEN [0.504964]
```

Figura 7.14: Regla que extiende la ruta *Transporte transmembrana de moléculas pequeñas* (16) en sistema ERR-PDR.

Por otro lado, el sistema ERR-PDR expande la ruta *Señalización de GPCR* (10) con 2 reglas o ramas diferentes del árbol de decisión, que añaden 11 y 3 proteínas respectivamente (ver figura 7.15). La primera regla describe proteínas con secuencia larga, sin dominio de señal, con varias interacciones en complejo y **sin** baja carga negativa. En contraste, la segunda regla define proteínas **con** baja carga negativa, con dominio transmembrana, sin secuencia corta y con interacción en complejo.

```
Regla 60:
=========
IF protein(A,B,C,D),D<0.072897 = 0 AND
   signal_domain(A) = 0 AND
   protein(A,B,C,D),B>629 = 1 AND
   complex_interaction(A,B),not(B=A),ppinteraction_pair(B,C),not(C=A),
      not(C=B),complex_interaction(A,D),not(D=A),not(D=B),
      not(D=C) = 0 AND
   complex_interaction(A,B),not(B=A),complex_interaction(B,C),not(C=A),
      not(C=B),signal_domain(C) = 1 AND
   complex_interaction(A,B),not(B=A),protein(B,C,D,E),E>0.133171 = 1
THEN [0.688929]

Regla 9:
========
IF protein(A,B,C,D),D<0.072897 = 1 AND
   signal_domain(A),protein(A,B,C,D),C>0.129964 = 0 AND
   protein(A,B,C,D),B<300,protein(A,E,F,G),G<0.072897 = 0 AND
   transmembrane_domain(A),protein(A,B,C,D),D<0.072897 = 1 AND
   complex_interaction(A,B),not(B=A),complex_interaction(A,C),not(C=A),
      not(C=B),protein(C,D,E,F),F>0.133171 = 1 AND
   complex_interaction(A,B),not(B=A),transmembrane_domain(B),
      complex_interaction(B,C),not(C=A),not(C=B) = 0
THEN [0.866684]
```

Figura 7.15: Reglas que extienden la ruta *Señalización de GPCR* (10) en sistema ERR-PDR.

Además, se pueden comparar los patrones que componen las reglas, con las propiedades moleculares frecuentes en el conjunto de proteínas que extiende cada ruta. Dichas propiedades se representan en la figura 7.16 como predicados lógicos y su importancia como su frecuencia de aparición en el conjunto de proteínas concreto. Sólo se seleccionan los elementos básicos (ver columnas de la figura 7.16), entre todas las posibles combinaciones de patrones frecuentes

complejos, los cuales se extraen con el algoritmo WARMR (ver sección 7.2.5). El término bajo/alto en un predicado representa un valor numérico situado en el primer/último cuartil de la distribución de valores en el conjunto completo de proteínas, respectivamente (ver sección 7.2.4).

Figura 7.16: Frecuencia de predicados simples por ruta. Sistema ERR-PDR. Los círculos representan la frecuencia en las proteínas predichas por el sistema ERR-PDR.

Los puntos de mayor frecuencia en la figura 7.16, en una ruta extendida por varias reglas, tienden a ser los correspondientes a las propiedades que se comparten entre reglas. En algún caso puede haber unos referentes a una regla y otros a otra. Aunque no todas las altas frecuencias que muestran los puntos de la figura 7.16 se corresponden con un patrón de alguna regla, y viceversa. Por ejemplo, cabe destacar que la característica diferenciadora (valor opuesto sobre la carga negativa) entre las dos reglas que muestra la figura 7.15 que extienden la misma ruta, no se refleja sobre las propiedades frecuentes (ver cuarta fila empezando por debajo en la figura 7.16). Otra muestra de la falta de equivalencia total entre altas frecuencias y reglas es la regla de la figura 7.14 para la ruta *Transporte transmembrana de moléculas pequeñas* (16), que no incluye en ninguno de sus patrones interacciones en complejos ni un alto número de transcritos, mientras que las frecuencias sí lo reflejan. Las razones de esta falta de correspondencia entre reglas y propiedades frecuentes pueden ser los patrones de una regla

que no se exige que se cumplan (con valor igual a 0 en las figuras 7.14 y 7.15), que no tienen una interpretación directa; o patrones que se satisfacen pero no se verifican en ninguna regla. No obstante, en general sí existe una coherencia razonable entre ambas interpretaciones: las reglas de decisión y las propiedades frecuentes.

7.5. Comparación con Extensión basada sólo en Similitud de Secuencia

Dado que el sistema ERR predice proteínas basándose principalmente en propiedades de la secuencia, cabe pensar en una comparación con una técnica de búsqueda de proteínas por similitud de secuencia. Es decir, buscar proteínas homólogas con la herramienta BLASTP [Altschul et al., 1997] (ver la sección B.3.1 para una definición de homología), y suponer una transferencia de las anotaciones de Reactome entre proteínas homólogas. No hay que olvidar que el sistema ERR se ha diseñado para ser aplicable en ausencia de similitud de secuencia (que en el resto de esta sección se denomina homología), por lo que las proteínas que se podrían asignar por homología se han excluido explícitamente del proceso de evaluación y aplicación.

Así, en esta sección se analiza si una proteína homóloga respecto a otra proteína anotada en la base de datos Reactome o predicha por el sistema ERR, también está anotada o se predice como perteneciente a la misma ruta, y el porqué de estos hechos.

Para decidir si dos proteínas son similares (homólogas) se realizan comparaciones de secuencias con la herramienta BLASTP [Altschul et al., 1997], y se contabilizan como homólogos aquellos pares que poseen al menos un 30 % de identidad en secuencia, y a la vez, el alineamiento cubre el 75 % de una de las secuencias. Se decide imponer la restricción de un 75 % de cobertura de secuencia, pues con un valor de 60 % se detectan pares de proteínas que sólo comparten uno o varios dominios Pfam. La estrategia de búsqueda es la siguiente: se seleccionan las proteínas predichas por ERR-PDR en los conjuntos de entrenamiento y test y se comparan con las proteínas no-anotadas en las rutas de Reactome, que se incluyen en el conjunto de aplicación o no (denominado conjunto de ‘resto de proteínas no-anotadas’). Además, las proteínas seleccionadas también se comparan respecto a otro conjunto de proteínas anotadas en las rutas de la base de datos Reactome, pero que no se incluyeron en los conjuntos entrenamiento y test por ser redundantes a ellos.

En el anexo E se presenta un resumen cuantitativo de los resultados obtenidos en la búsqueda de homólogos en dos tablas. La tabla E.1 recopila información sobre las proteínas anotadas contenidas en los conjuntos de entrenamiento y test, y la tabla E.2 sobre las anotadas que se eliminaron de dichos conjuntos por ser redundantes. Al analizar las tablas, se descubren casos donde existen:

1. Proteínas homólogas con una anotada en la ruta y la otra no.

2. Proteínas homólogas con una predicha por ERR y la otra no.

El primer caso se puede verificar comparando las columnas primera (proteínas anotadas en las rutas) y tercera (proteínas homólogas entre las no-anotadas) de cualquiera de las dos tablas resumen, E.1 ó E.2. En la figura 7.17 se muestran algunos de estos ejemplos, en diferentes rutas, seleccionando los que poseen los mayores porcentajes de identidad en secuencia.

El segundo caso se corrobora al observar que entre las columnas cuarta y quinta de las tablas E.1 y E.2 hay diferencia. Ambas columnas representan proteínas homólogas a las predichas,

```
Ruta Metabolismo de proteínas:
RL36A_HUMAN / RL36L_HUMAN (proteína ribosomal 60S L36a/tipo-L36a)
* 99.06% identidad
* Vector de propiedades:
RL36A_HUMAN | 0,1,1,0,0,1,1,0,0,0,0,0,0,0,1
RL36L_HUMAN | 0,0,1,0,0,1,1,0,0,0,0,1,0,1,0

Ruta Ciclo celular, fase mitótica:
DYR_HUMAN / DYRL1_HUMAN (Dihidrofolato reductasa/tipo-reductasa proteína 1)
* 93.44% identidad
* Vector propiedades:
DYR_HUMAN   | 1,1,1,0,0,1,0,1,0,0,0,0,0,1,0
DYRL1_HUMAN | 0,0,1,0,0,1,0,1,0,0,0,0,0,0,1

Ruta Transcripción:
TCEA1_HUMAN / TCEA2_HUMAN (Factor A elongación transcripción proteína 1/2)
* 66.78% identidad
* Vector propiedades:
TCEA1_HUMAN | 1,1,0,0,0,1,0,1,0,0,1,0,1,0,0
TCEA2_HUMAN | 1,1,1,0,0,1,0,1,0,0,0,0,0,0,1
```

Figura 7.17: Ejemplos de pares de proteínas homólogas anotadas y no-anotadas en Reactome. La primera proteína pertenece al conjunto anotadas en rutas y la segunda a las no-anotadas (ya sea predicha o no). Los componentes del vector de propiedades de la secuencia, de izquierda a derecha, son: interacción proteína-proteína, interacción en complejo, longitud de la secuencia de proteína baja y alta, carga positiva de la proteína baja y alta, carga negativa de la proteína baja y alta, dominio transmembrana, dominio de señal, dominio hélice, longitud de la secuencia génica baja y alta, y número de isoformas bajo y alto.

pero sólo las de la columna 5 se predicen por ERR, y sólo en los pares *(entrenamiento-o-test, no-anotada)*.

Para explicar las respuestas a estas cuestiones, el primer elemento que se debe considerar es que los resultados que se obtienen con una búsqueda de similitud de secuencias mediante la herramienta BLASTP, no tienen que coincidir con los obtenidos por el sistema ERR. BLASTP utiliza sólo la información de la secuencia de aminoácidos de las proteínas, mientras que ERR incluye además, otro conjunto de propiedades. Por ejemplo, el número de transcritos (isoformas), la longitud del gen que codifica la proteína, interacciones por pares con otras proteínas, y participación en complejos de proteínas.

Por ello, sólo se podrían esperar resultados similares entre ERR y BLASTP en los casos en que el porcentaje de identidad entre las secuencias que se comparan es elevado. Además, si sólo se consideran en ERR las propiedades que dependen únicamente de la secuencia de aminoácidos (longitud, cargas positiva y negativa, dominios transmembrana, de señal y en hélice), estas propiedades difieren entre las proteínas de una misma familia, cuando los porcentajes de identidad en secuencia disminuyen.

Considerando todos estos elementos, y utilizando ejemplos como los que se muestran en la figura 7.17, se observa que, en la mayoría de los casos, las proteínas "homólogas"que no se predicen es por diferencias en la existencia de interacciones proteína-proteína o participación en complejos de proteínas. Este hecho se cumple, en diferentes rutas, en los pares de proteínas (RL36A_HUMAN,RL36L_HUMAN) y (DYR_HUMAN,DYRL1_HUMAN), ya que tienen diferencias en las dos primeras componentes del vector de propiedades, correspondiente a las interacciones (ver figura 7.17).

No obstante, existen otros casos que ERR no predice, a pesar de tener anotaciones

de interacciones proteína-proteína y/o participar en complejos, porque la proteína con que interactúan difiere en características tales como: número de isoformas, longitud del gen y de la proteína, y presencia de dominios en hélice. Este hecho se cumple en el par (TCEA1_HUMAN,TCEA2_HUMAN) de la ruta *Transcripción*, cuyos vectores de propiedades aparecen en la figura 7.17.

También se observan casos extremos como por ejemplo el par de proteínas *"homólogas"* (RL36A_HUMAN,RL36L_HUMAN), de la ruta *Metabolismo de proteínas*, que poseen un 99 % de identidad en secuencia, y sin embargo RL36L_HUMAN no se predice porque no participa en un complejo en el que interaccione con una proteína de alta carga positiva. Por otro lado, cabe destacar el par *"homólogo"* (TCEA1_HUMAN,TCEA2_HUMAN) de la ruta *Transcripción*, con casi un 67 % de identidad, dónde la diferencia entre la proteína predicha y la no-predicha es muy específica. En concreto, ERR no predice la homóloga por: "no tener interacción en un complejo, con una proteína con dominio en hélice y baja carga negativa". Es decir, la justificación para no predecir depende también de las propiedades de las secuencias con las que interacciona la proteína, no sólo las de la proteína en cuestión TCEA2_HUMAN. En este caso también es destacable que la proteína TCEA2_HUMAN participa en complejos, pero ninguna de las proteínas con las que se relaciona tienen un dominio en hélice y baja carga negativa. Por otro lado, este ejemplo también verifica la utilización de la información adicional que la representación relacional permite.

Un último elemento a considerar es que se limita la búsqueda de proteínas homólogas en el conjunto de no-anotadas en Reactome, a aquellas que poseen identificador UniProt. Así, se eliminan las que sólo poseen identificadores UniParc en febrero del 2010, excepto en la quinta columna de la tabla E.1. Este criterio implica que no se detecten las cuatro proteínas predichas homólogas de proteínas anotadas también predichas.

En conclusión, por un lado, una extensión de rutas basada sólo en similitud de secuencia calculada con la herramienta BLASTP no coincidiría con la extensión del sistema ERR, porque BLASTP utiliza menos información basada en la secuencia de la que emplea el sistema ERR, a parte de las interacciones. Por otro lado, hay que destacar que existen proteínas similares en secuencia a las de la ruta original que no están anotadas en Reactome; por lo que no es excesivamente sorprendente que proteínas similares en secuencia a las predichas por ERR tampoco se predigan.

7.6. Comparación con Método de Extensión de Rutas basado sólo en Redes de Interacción

El objetivo de esta sección es comparar los sistemas de extensión de rutas desarrollados en este trabajo con otro método que también expande rutas, aunque usando sólo información de redes de interacciones moleculares.

Glaab y colaboradores [Glaab et al., 2010] ha propuesto recientemente una metodología para extender rutas biológicas y otros procesos celulares. Su método mapea el conjunto de proteínas de la ruta sobre una red de interacciones proteína-proteína y, entonces, extiende la ruta añadiendo proteínas que hagan que la ruta final aumente su conectividad y sea más compacta. La red de interacción es la única entrada de este método, y los candidatos para el procedimiento de extensión son sólo las proteínas conectadas por interacción directa a la ruta original. Cuando una proteína candidata cumple una serie de condiciones topológicas [Glaab et al., 2010] y compacta la ruta, entonces se elige para extender la ruta. Para comparar en las mismas condiciones, se ha re-implementado este método para poder aplicarlo sobre las rutas y

la red de interacción usada en el presente trabajo (ver sección 7.2.1), que es diferente de la red empleada originalmente en el método [Glaab et al., 2010].

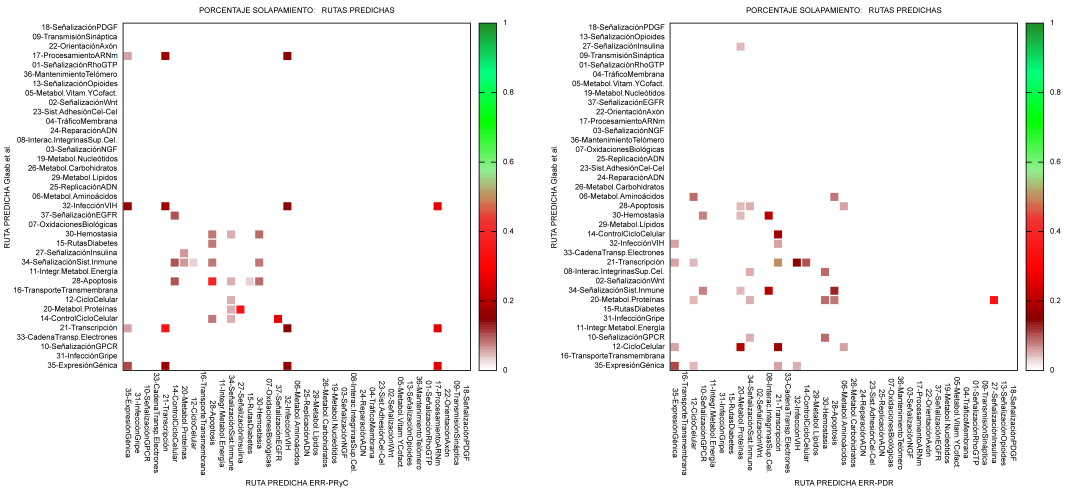
7.6.1. Análisis Cuantitativo

Usando como entrada las rutas completas (con las proteínas redundantes en términos de similitud de secuencia, según lo descrito en la sección 7.2.1), el método Glaab et al. extiende 29 de 37 rutas, con 150 proteínas directamente conectadas en total, siendo 90 de ellas diferentes (lo que supone un 60 % del total de 150 añadidas), como muestra la tabla 7.2. No se han tomado las rutas no redundantes, como en ERR, porque en ese caso las extensiones del método Glaab et al. eran muy escasas, al depender sólo de las interacciones, que se ven reducidas al eliminar la redundancia.

Tabla 7.2: Comparación numérica de la extensión de Reactome por Glaab et al. con los sistemas ERR-PRyC y ERR-PDR.

Sistema	**nºrutas extendidas**	**nºproteínas añadidas** (total/distintas/ %distintas)
Glaab et al.	29	150 / 90 / 60,00 %
ERR-PRyC	18	249 / 218 / 87,55 %
ERR-PDR	28	383 / 329 / 85,90 %

El método Glaab et al. extiende 21 rutas en común con el sistema ERR-PDR y 15 rutas con el ERR-PRyC. Sin embargo, para cada ruta concreta hay muy pocas proteínas predichas en común por Glaab et al. y por los sistemas ERR. En el sistema ERR-PDR hay 5 proteínas comunes: 2 en la ruta *Expresión génica* (35) y 3 en *Transcripción* (21). En el sistema ERR-PRyC hay 11 proteínas comunes: 2 en la ruta *Expresión génica* (35), 2 en *Transcripción* (21), 5 en *Apoptosis* (28), 1 en *Hemostasia* (30) y 1 en *Infección VIH* (32).



(a) Sistema ERR-PRyC. (b) Sistema ERR-PDR.

Figura 7.18: Porcentaje de solapamiento entre rutas. Comparación Glaab et al. con (a) sistema ERR-PRyC y (b) ERR-PDR. Las rutas están ordenadas por AUPRC creciente según cada sistema, de izquierda a derecha en el *eje x* y de abajo a arriba en el *eje y*.

Además, si se consideran las proteínas comunes añadidas por los métodos a diferentes rutas, tampoco aumenta mucho el número de coincidencias, como se puede ver en la figura 7.18. Por lo tanto, se puede concluir que ambos métodos de extensión de rutas son complementarios entre sí, pues añaden proteínas diferentes. Además, los sistemas ERR extienden con muchas más proteínas que el método Glaab et al.

7.6.2. Comparación de Similitud Semántica

En este apartado, se usa una medida de similitud semántica funcional para comparar ambos métodos de extensión de rutas. Se aplica el mismo esquema descrito previamente en la sección 7.4.5, basado en las anotaciones de Proceso Biológico de GO. La similitud semántica entre las proteínas de la ruta original y las proteínas añadidas es mayor para el método Glaab et al. que para los sistemas ERR (ver figura 7.19 para ERR-PDR y G.4 para ERR-PRyC). Las medias de similitud, calculadas sobre el número de rutas que extiende cada sistema, son: 0,700 (Glaab et al., para 29 rutas), 0,591 (ERR-PDR, para 28 rutas) y 0,589 (ERR-PRyC, para 18 rutas).

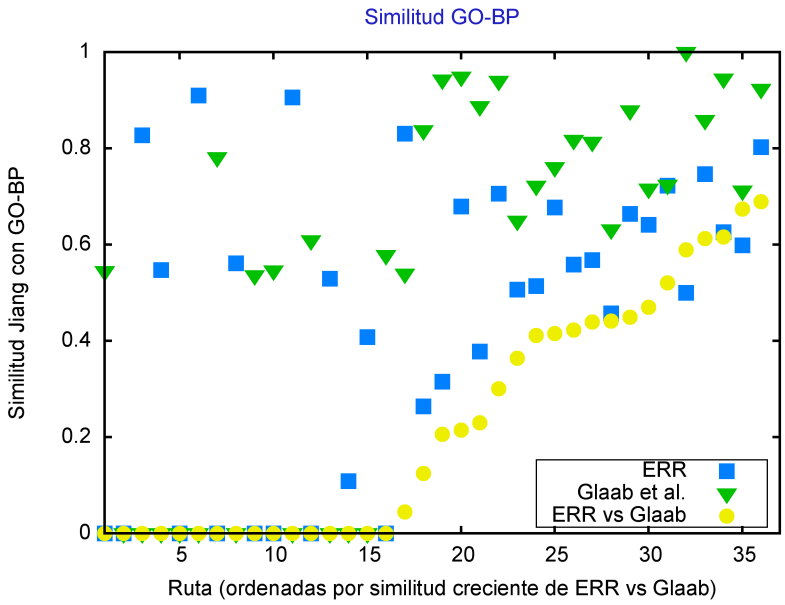


Figura 7.19: Similitud de anotación funcional entre proteínas de la ruta original y las proteínas añadidas (ERR-PDR y Glaab et al.) y entre ambos sistemas de extensión.

Además, las proteínas predichas por cada método de extensión son más similares semánticamente a la ruta original que entre ellas (0,483 y 0,412 de similitud entre Glaab et al. y el sistema ERR-PRyC y ERR-PDR, respectivamente). Este hecho verifica que los conjuntos de proteínas para ambos métodos de extensión son muy distintos entre sí. Por lo tanto, se puede concluir que las predicciones del método Glaab et al. es una extensión limitada a un área funcional próxima (es decir, proteínas muy conectadas con las rutas), mientras que las proteínas propuestas por los sistemas ERR son más distintas entre ellas y exploran más espacio funcional.

7.6.3. Comparación de Solapamiento entre Rutas

Si se analiza la cantidad de proteínas comunes añadidas a diferentes rutas, el solapamiento de las extensiones del método Glaab et al. es mayor que en los sistemas ERR: un 30 % de las proteínas extienden más de una ruta en el método Glaab et al., mientras que sólo alrededor de un 15 % en los sistemas ERR, como se observa en la última columna de la tabla 7.2. De hecho, se puede observar este mismo efecto en la figura 7.20, que representa el solapamiento de las extensiones de Glaab et al. entre sí, donde existen muchas celdas coloreadas fuera de la diagonal principal, en contraposición a la escasa existencia de estos puntos en los sistemas ERR-PRyC (figura 7.11(a)) y ERR-PDR (figura 7.11(b)).

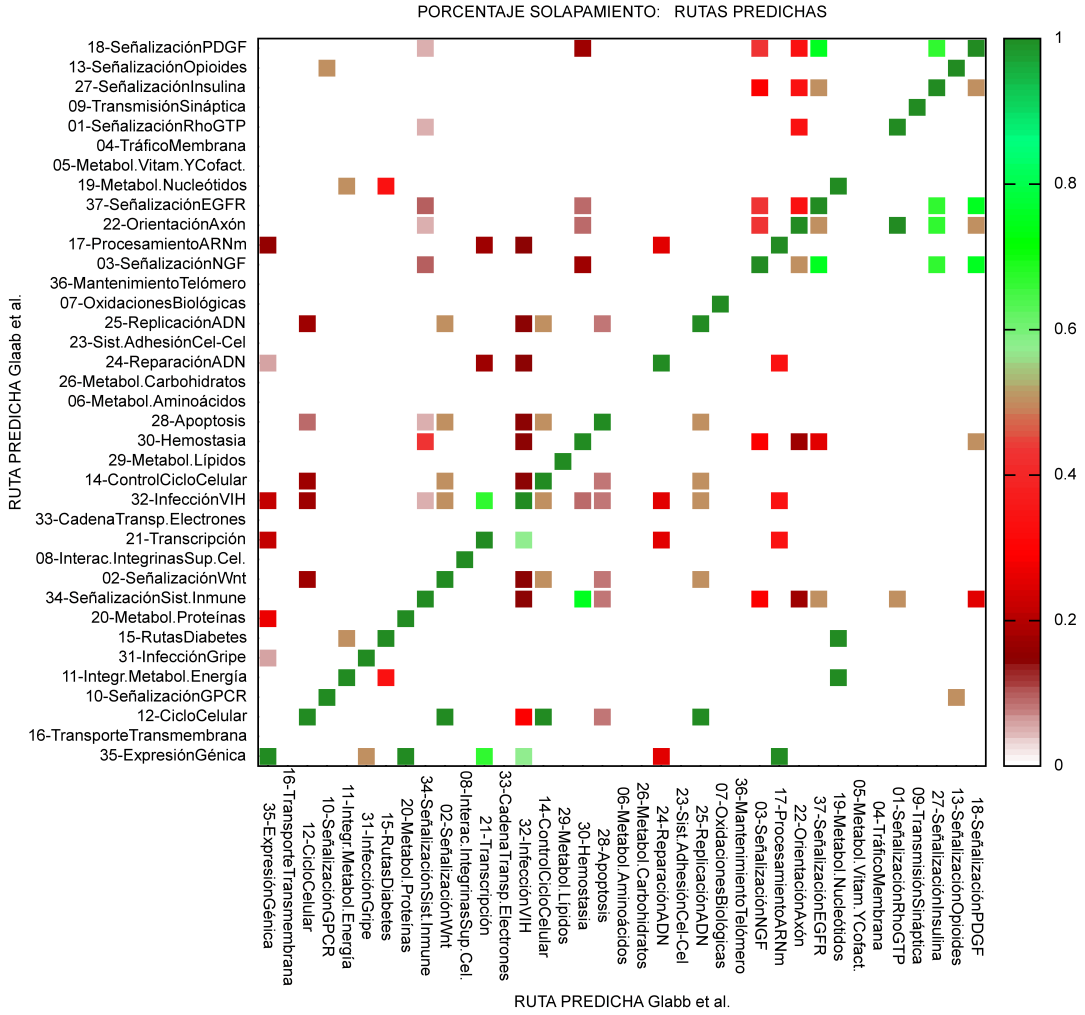


Figura 7.20: Porcentaje de solapamiento entre rutas en Glaab et al.

Si se elimina el solapamiento, es decir, las proteínas que extienden más de una ruta, ambos métodos se parecen más en términos de similitud semántica a la ruta original. Así, el método Glaab et al. sería más similar en 15 rutas y los sistemas ERR-PRyC y ERR-PDR en otras 15 rutas. Es importante tener en cuenta que, sin el solapamiento, el sistema ERR-PDR sólo deja de extender 1 ruta, quedándose en 27 de 37; mientras que Glaab et al. pierde 10 rutas, extendiendo

sólo 19 de 37.

Por lo tanto, este resultado refuerza las conclusiones de las secciones anteriores: el método Glaab et al. busca pocas relaciones obvias y cercanas, contrastando con los sistemas ERR que localizan muchas relaciones lejanas; confirmando también la complementariedad de ambos métodos de extensión de rutas.

7.6.4. Análisis de Frecuencia de Predicados

En esta sección se comparan las propiedades de las proteínas de las rutas originales con las propiedades de las proteínas añadidas por ERR-PDR y por Glaab et al.

Las propiedades y su importancia se representan con predicados lógicos y su frecuencia en el conjunto de proteínas, igual que en la figura 7.16. En este caso, en cada cruce de los *ejes X* (propiedad) e *Y* (ruta) de la figura 7.21 se representa un trío de círculos. El círculo izquierdo (rojo) representa la frecuencia en el conjunto de proteínas de la ruta original; el círculo central (azul) representa la frecuencia en el conjunto de proteínas predichas por el sistema ERR-PDR, y el círculo derecho (verde), la frecuencia en el conjunto de proteínas expandidas por el método Glaab et al.

Aunque presentan un aspecto similar, las figuras 7.21 y 7.10 son distintas, porque una propiedad puede ser frecuente en un conjunto, pero no ser relevante para el proceso de aprendizaje, y viceversa.

Al analizar la figura 7.21 (ver final del capítulo), se concluye que no hay correlación completa entre los círculos rojos, azules y verdes. Sin embargo, los círculos rojos y azules tienen más correlación que los rojos y verdes (observar que en la figura 7.21 hay más pares de círculos rojo-azul que rojo-verde). Esto quiere decir que las propiedades moleculares de las proteínas de las rutas originales se parecen más a las propiedades de las proteínas predichas por ERR-PDR que las expandidas por Glaab et al. Este hecho era esperable porque el método Glaab et al. no está basado en todas estas propiedades moleculares, sino sólo en las interacciones. Así, en las extensiones de ruta del método Glaab et al. (círculos verdes), la propiedad asociada a las interacciones (principalmente en complejos) siempre toma la frecuencia más alta de todas las propiedades. En este sentido, hay que tener en cuenta que en el cálculo de la frecuencia de predicados se consideran interacciones con cualquier proteína, perteneciente o no a la ruta. En los sistemas ERR-PRyC y ERR-PDR se considera lo mismo cuando se usa un predicado de interacción entre pares de proteínas o en complejo. Tampoco hay correlación entre los puntos azules y verdes, lo cual verifica otra vez que ambos métodos de extensión añaden proteínas diferentes, con propiedades diferentes.

Comparando los sistemas ERR-PRyC y ERR-PDR (ver figura G.5 y 7.21), el segundo presenta una densidad de círculos azules mayor que el primero, ya que el sistema ERR-PDR extiende 10 rutas más. Por otro lado, hay variaciones en la frecuencia de las propiedades cuando una ruta se extiende mediante varias reglas en vez de sólo con una, con la señal de frecuencia diluida entre más propiedades.

7.7. Relevancia Biológica de las Proteínas Predichas

La pregunta que se quiere contestar en esta sección es: *¿existen evidencias biológicas que corroboren la asociación entre las predicciones de ERR y las rutas extendidas?* Es decir, se trata de explicar la relación entre las rutas biológicas originales y las anotaciones funcionales de las proteínas que se predice que extienden cada ruta, para dar una coherencia y significado

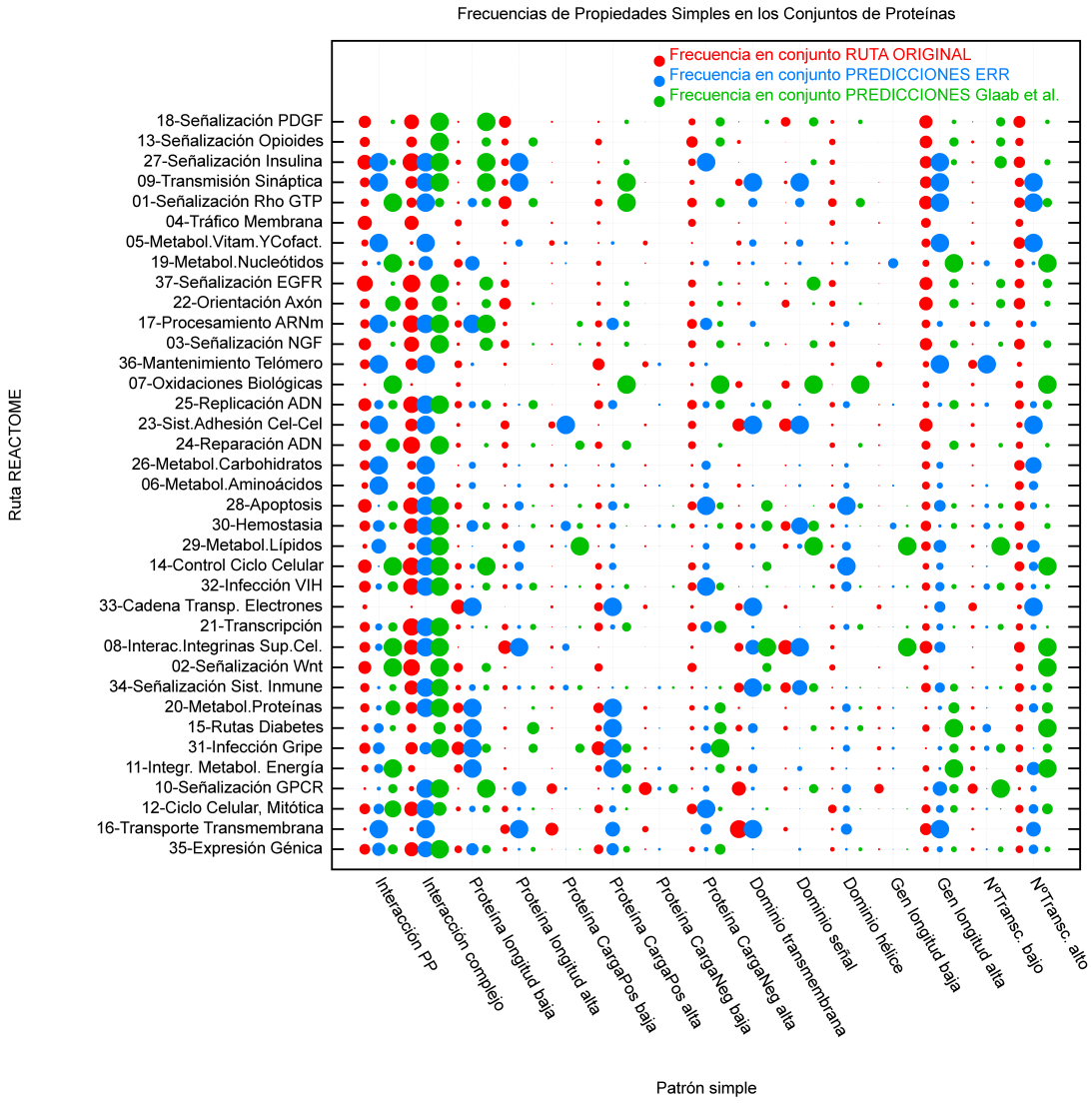


Figura 7.21: Comparación de frecuencia de predicados simples por ruta. Sistema ERR-PDR. Los círculos izquierdos/rojos representan la frecuencia en las proteínas de las rutas originales, los círculos centrales/azules la frecuencia en las proteínas predichas por el sistema ERR-PDR, y los círculos derechos/verdes la frecuencia en las proteínas expandidas por el método Glaab et al.

biológico. Se buscan anotaciones de dominios y hallazgos en la literatura para añadir evidencias biológicas que apoyen las nuevas predicciones del sistema, que justifiquen su utilidad en el área biológica.

El contenido de las siguientes sub-secciones es el que sigue. Primero se analizan las extensiones comunes entre los sistemas ERR-PRyC, ERR-PDR y Glaab et al. Segundo, se estudia el significado biológico basado en las propiedades moleculares simples de las proteínas predichas en una ruta, en subconjuntos y en proteínas independientes. Tercero, se presentan algunos ejemplos concretos de proteínas analizadas en detalle para cuatro rutas diferentes, cuya predicción está suficientemente justificada en términos biológicos, a falta de una verificación experimental.

7.7.1. Predicciones Simultáneas por Varios Sistemas

Partiendo de la idea de que predicciones comunes entre distintos métodos independientes son más fiables que las que predice un único sistema, se localizan las extensiones de rutas simultáneas entre los dos sistemas ERR propuestos en este capítulo, comparándolas también con los resultados propuestos por el método de Glaab et al. [Glaab et al., 2010]. Para ilustrar la utilidad de dichas predicciones, se analizan las anotaciones funcionales de la base de datos UniProt [Consortium, 2011] y algunas extraídas de la literatura científica. La tabla 7.3 resume los identificadores de UniProt de las proteínas predichas simultáneamente por ERR-PRyC y ERR-PDR, así como las coincidencias con las predicciones de Glaab et al.

Tabla 7.3: Lista de las rutas biológicas de Reactome y de las proteínas predichas simultáneamente por los métodos ERR-PRyC, ERR-PDR y Glaab et al. Los identificadores de las proteínas se anotan de acuerdo a la nomenclatura de la base de datos UniProt.

Id. ruta	**Nombre ruta**	**ERR-PRyC + ERR-PDR**	**Glaab + ERR-PRyC + ERR-PDR**	**Glaab + ERR-PRyC**	**Glaab + ERR-PDR**
12	Ciclo celular, fase mitótica	UCHL1_HUMAN LNP_HUMAN			
16	Transporte transmembrana de moléculas pequeñas	A6ND05_HUMAN			
20	Metabolismo de proteínas	RT14_HUMAN ZFAN5_HUMAN TBCD7_HUMAN RPP38_HUMAN A6NGZ2_HUMAN CNRG_HUMAN RT24_HUMAN POP7_HUMAN RT25_HUMAN TUSC2_HUMAN RPP29_HUMAN			
21	Transcripción	TAF2_HUMAN RPC3_HUMAN CREG1_HUMAN	TAF2_HUMAN RPC3_HUMAN	TAF2_HUMAN RPC3_HUMAN	TAF2_HUMAN RPC3_HUMAN B3KRR0_HUMAN
28	Apoptosis			SPNS1_HUMAN AVEN_HUMAN SEC20_HUMAN FKBP8_HUMAN BIK_HUMAN	
30	Hemostasia			CD22_HUMAN	
32	Infección VIH	SUPT3_HUMAN FIBP_HUMAN PURA_HUMAN MED11_HUMAN MED21_HUMAN TEBP_HUMAN		TF2H5_HUMAN	
35	Expresión génica	TBCD7_HUMAN RPP38_HUMAN A6NGZ2_HUMAN CNRG_HUMAN CRIPT_HUMAN TUSC2_HUMAN		MED19_HUMAN E9PDR7_HUMAN	PDCD4_HUMAN MED29_HUMAN

A continuación se discuten 3 escenarios diferentes, correspondientes a los tres subconjuntos de predicciones simultáneas en los que se pueden agrupar los resultados de la tabla 7.3:

1. Las predicciones coincidentes entre Glaab et al. y los sistemas ERR-PRyC y ERR-PDR
2. Las predicciones coincidentes entre Glaab et al. y sólo uno de los dos sistemas de anotación
3. Las predicciones coincidentes por los dos sistemas de anotación, pero diferentes al método de Glaab et al.

1. Glaab et al. y ERR-PRyC o ERR-PDR

En el primer caso, sólo coinciden dos proteínas: TAF2_HUMAN y RPC3_HUMAN, ambas propuestas para extender la ruta *Transcripción*. TAF2_HUMAN es la subunidad 2 del factor de iniciación de la transcripción TFIID, mientras que RPC3_HUMAN es la subunidad C3 de la ARN polimerasa III dirigida al ADN. Además, estas proteínas tienen una localización subcelular en el núcleo, donde se produce el proceso de transcripción del ADN. En conclusión, todas las anotaciones encontradas están relacionadas con la transcripción, indicando una coherencia en las predicciones.

2.a. Glaab et al. y ERR-PDR

La comparación de las predicciones del sistema ERR-PDR y Glaab et al. indica que sólo 5 proteínas coinciden: TAF2_HUMAN, RPC3_HUMAN y B3KKR0_HUMAN propuestas para extender la ruta *Transcripción*, y por su parte se sugieren MED29_HUMAN y PDCD4_HUMAN en la ruta *Expresión génica*.

En el apartado anterior, relativo al primer escenario, ya se ha explicado la inclusión de las proteínas TAF2_HUMAN y RPC3_HUMAN en la ruta de *Transcripción*. La restante B3KKR0_HUMAN es una proteína no caracterizada, la que sólo ha sido anotada con una alta similitud a la proteína ERCC-1 de reparación del ADN por escisión.

Por otro lado, MED29_HUMAN es un mediador de la subunidad 29 de transcripción de la ARN polimerasa II. Es destacable que otro miembro de este complejo, MED19_HUMAN también se predice por el sistema ERR-PRyC en la misma ruta (ver siguiente sección). PDCD4_HUMAN, la proteína 4 de la muerte celular programada, inhibe la iniciación de la traducción por unión con el factor de iniciación eucariótico 4A (eIF4A), y también inhibe la actividad helicasa de eIF4A. Aunque la relevancia biológica de estos hallazgos requieren una mayor investigación, las anotaciones de UniProt de estas proteínas y la predicción simultánea por dos métodos independientes, incrementa la fiabilidad de estos resultados.

2.b. Glaab et al. y ERR-PRyC

Por su parte, el sistema ERR-PRyC y el método Glaab et al. predicen nuevas proteínas en 5 rutas, como muestra la penúltima columna de la tabla 7.3.

TAF2_HUMAN y RPC3_HUMAN en la ruta *Transcripción*, MED19_HUMAN y E9PDR7_HUMAN en la ruta *Expresión génica*, TF2H5_HUMAN en *Infección VIH*, CD22_HUMAN en *Hemostasia*, y SPNS1_HUMAN, AVEN_HUMAN, BIK_HUMAN, SEC20_HUMAN y FKBP8_HUMAN en la ruta *Apoptosis*.

Las proteínas de la ruta *Transcripción* ya se han comentado en el apartado del escenario 1.

MED19_HUMAN es un mediador de la subunidad 19 de transcripción de la ARN polimerasa II, y un co-activador involucrado en la transcripción regulada de casi todos los genes dependientes de la ARN polimerasa II. El complejo mediador está compuesto por MED1, MED4, MED6, MED7, MED8, MED9, MED10, MED11, MED12, MED13, MED13L, MED14, MED15, MED16, MED17, MED18, MED19, MED20, MED21, MED22, MED23, MED24, MED25, MED26, MED27, MED29, MED30, MED31, CCNC, CDK8 y CDC2L6/CDK11. La otra proteína predicha para la *Expresión génica*, E9PDR7_HUMAN, es una proteína no caracterizada.

En la ruta *Infección VIH*, TF2H5_HUMAN, la subunidad 5 del factor de trascripción general IIH (TFIIH) está involucrada en la reparación del ADN por escisión de nucleótidos, y cuando forma un complejo con CAK, participa en la transcripción del ARN por la ARN polimerasa II. Merece la pena resaltar que, en esta ruta, los dos sistemas de anotación, ERR-PRyC y ERR-PDR, predicen MED11_HUMAN y MED21_HUMAN, ambos componentes del complejo mediador descrito antes en la ruta *Expresión génica*. Por lo tanto, *Infección VIH* y *Expresión génica* son procesos conectados.

CD22_HUMAN media interacciones célula-B con célula-B, y enlaza proteínas que en su estructura de oligosacárido contienen ácido siálico. Aunque no existen evidencias, su papel en la *Hemostasia* podría estar asociado al mecanismo de agregación de plaquetas. De hecho, la cascada de migración de células-B está modulada por plaquetas [Li, 2008].

Acerca de la ruta *Apoptosis*, SPNS1_HUMAN podría estar involucrado en la muerte celular necrótica o autofágica; AVEN_HUMAN protege contra la apoptosis mediado por Apaf-1; BIK_HUMAN acelera la muerte celular programada; SEC20_HUMAN está implicada en la supresión de la muerte celular, y la forma activa de FKBP8_HUMAN podría por lo tanto jugar un papel en la regulación de la apoptosis.

3. ERR-PRyC y ERR-PDR

Para el tercer escenario, es interesante observar que se predicen 5 proteínas simultáneamente por los dos sistemas de anotación propuestos en esta tesis en dos rutas diferentes, *Metabolismo de proteínas* y *Expresión génica*. TBCD7_HUMAN podría actuar como una proteína de activación del GTP para la familia de proteínas Rab; RPP38_HUMAN es una subunidad p38 de la proteína RNasaP; A6NGZ2_HUMAN es una proteína no caracterizada; CNRG_HUMAN, la subunidad-gamma de la fosfodiesterasa cGMP en la retina humana, participa en procesos de transmisión y amplificación de la señal visual; y TUSC2_HUMAN podría funcionar como un supresor de tumores. Al contrario que en los dos escenarios previos, aquí hay anotaciones funcionales de las proteínas predichas muy diversas. Por lo que sabemos, hasta ahora estas proteínas no se han enlazado a estos procesos.

También en este tercer escenario, se puede señalar que los sistemas ERR-PRyC y ERR-PDR añaden en común una proteína más a la ruta *Transcripción*, a parte de las otras dos que compartían con Glaab et al. Se trata de CREG1_HUMAN, que aunque es una proteína secretada por la célula, en sus anotaciones funcionales de UniProt argumenta explícitamente que podría contribuir al control de la transcripción del crecimiento y diferenciación de la célula.

En el caso de otras rutas donde existen predicciones de proteínas que los extienden, y que se realizaron simultáneamente por los dos sistemas ERR, existe poca información en la literatura que permita argumentar la inclusión o no de las proteínas predichas en las rutas biológicas estudiadas.

Por otro lado, está ampliamente documentado en la literatura científica que la participación de algunas proteínas en múltiples procesos sugiere que existen comunicaciones extensas entre

diferentes procesos celulares. El análisis realizado en este trabajo revela que *Ciclo celular, fase mitótica*, *Expresión génica* y *Metabolismo de aminoácidos* son las rutas con el mayor número de proteínas predichas que conectan otros procesos celulares. 68 proteínas predichas se predicen en al menos dos procesos, de las cuales sólo 10 proteínas están en tres rutas al mismo tiempo.

7.7.2. Relación con Propiedades Moleculares Simples

En esta sección, el análisis de relevancia biológica se basa en las propiedades moleculares simples que usa el predictor. Dichas propiedades, procedentes principalmente de la secuencia, se representan mediante los predicados lógicos usados (ver la figura 7.2).

Se realizan dos análisis: primero se analiza la frecuencia de los predicados en subconjuntos de proteínas y, segundo, se estudia la presencia o ausencia de cada predicado por proteína independiente.

Por Subconjuntos de Proteínas

En los sistemas ERR propuestos, en común con otros trabajos previos [Jensen et al., 2002b, 2003a; Bendtsen et al., 2004] que usan combinaciones complejas de propiedades sencillas, es difícil interpretar los resultados según dichas propiedades. No obstante, aunque un análisis genérico y exhaustivo no sea posible, un investigador interesado en un aspecto concreto de una ruta o proteína podría analizar el fragmento específico de la figura 7.21 relacionado con ello, por ejemplo, como se hace a continuación.

Este análisis es una muestra de la utilidad del enfoque, por lo que sólo se centra en algunas extensiones del sistema con mayor diversidad en reglas por ruta (ERR-PDR). En concreto, se discuten las predicciones coincidentes entre el método Glaab et al. y el sistema ERR-PDR.

Como se ha comentado previamente y se observa en la tabla 7.3, entre el sistema ERR-PDR y Glaab et al. sólo coinciden 5 proteínas: TAF2_HUMAN, RPC3_HUMAN y B3KKR0_HUMAN en la ruta *Transcripción*, y MED29_HUMAN y PDCD4_HUMAN en la ruta *Expresión génica*. Un análisis detallado de la frecuencia de los predicados en esta predicción, indica que estas cinco proteínas comparten una frecuencia alta del predicado `complex_interaction`, al igual que más del 90 % observado en las 131 proteínas anotadas originalmente en la ruta *Transcripción*, y más de un 75 % en las 379 proteínas en la ruta *Expresión génica*. Este hecho está de acuerdo con la estrategia de complejos multi-proteína para la regulación de funciones celulares [Cramer et al., 2000]. Otros predicados con baja frecuencia, por ejemplo `gene_transcriptCount_low` en la ruta *Transcripción* y `protein_negCharge_high` en la ruta *Expresión génica*, ayudan a diferenciar estas cinco proteínas de otras predichas y de aquellas anotadas en las rutas, como se observa en la figura 7.22.

Por Proteínas Independientes

La hipótesis de partida del sistema ERR es que las proteínas de una ruta pueden ser distintas a nivel molecular (ver sección 7.1). Por lo tanto, no se espera que todas las proteínas originalmente anotadas en una ruta tengan sus propiedades de secuencia en común, ni las proteínas predichas tampoco.

Para clarificar este punto, se incluye un análisis que consiste en comparar individualmente las propiedades de todas las proteínas de una ruta; tanto las de la ruta original como las proteínas

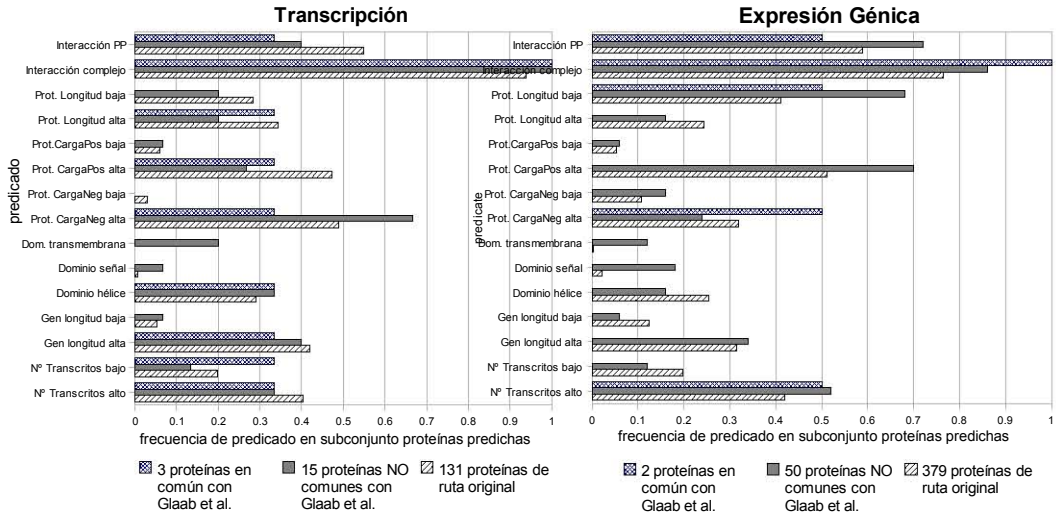


Figura 7.22: Frecuencia de predicados en subconjuntos de proteínas predichas: Rutas *Transcripción* y *Expresión génica*.

no anotadas en Reactome, pero añadidas a la ruta por el sistema ERR-PDR o por el método de Glaab et al.

Para ello, tomando como atributos las propiedades moleculares simples, representadas en predicados lógicos, se aplica un algoritmo de agrupamiento o *clustering* sobre las diferentes proteínas (originales, predichas por ERR y predichas por Glaab et al.). Se realiza un *clustering* jerárquico de aglomeración, construyendo un árbol binario de abajo hacia arriba, de las hojas a la raíz. En el árbol, proteínas con un ancestro común cercano son más similares que las que tienen un ancestro común lejano. Como medida de distancia entre elementos se usa el coeficiente de correlación de Pearson, típico al agrupar proteínas o genes. Esta medida se aplica sobre las propiedades utilizadas para representar cada proteína (simples de secuencia y presencia de interacciones). Así, se obtiene un mapa coloreado con las proteínas agrupadas por semejanza en perfiles de propiedades simples. Como las propiedades son predicados lógicos, sólo va a haber una escala de dos colores: amarillo si se satisface el predicado lógico y rojo si no se satisface.

Algunos ejemplos de estos mapas, coloreados y agrupando proteínas horizontalmente, se muestran en las figuras F.1, 7.24, 7.23 y F.2, para las rutas *Cadena de transporte de electrones*, *Replicación del ADN*, *Mantenimiento del telómero* e *Interacciones de las integrinas en la superficie celular*, respectivamente. Dos de las figuras se encuentran en esta sección como ejemplo, y las otras dos en el anexo F. Se han elegido estas rutas de Reactome por ser las mismas con propuestas reales de extensión con proteínas "de novo", analizadas en detalle en la sección 7.7.3. Todas las proteínas mencionadas en dicha sección se pueden localizar en estos mapas coloreados.

En general, las figuras ilustran que dentro de la misma ruta se pueden encontrar varios perfiles de propiedades, muy distintos entre sí. Cada uno sólo agrupa unas pocas proteínas, y no todas las de una ruta, verificando la diversidad de las rutas en términos moleculares. También se puede observar que las proteínas predichas (con el sufijo $>> ERR <<$) no se agrupan entre ellas sino que están más próximas a proteínas de la ruta original, incluso pueden

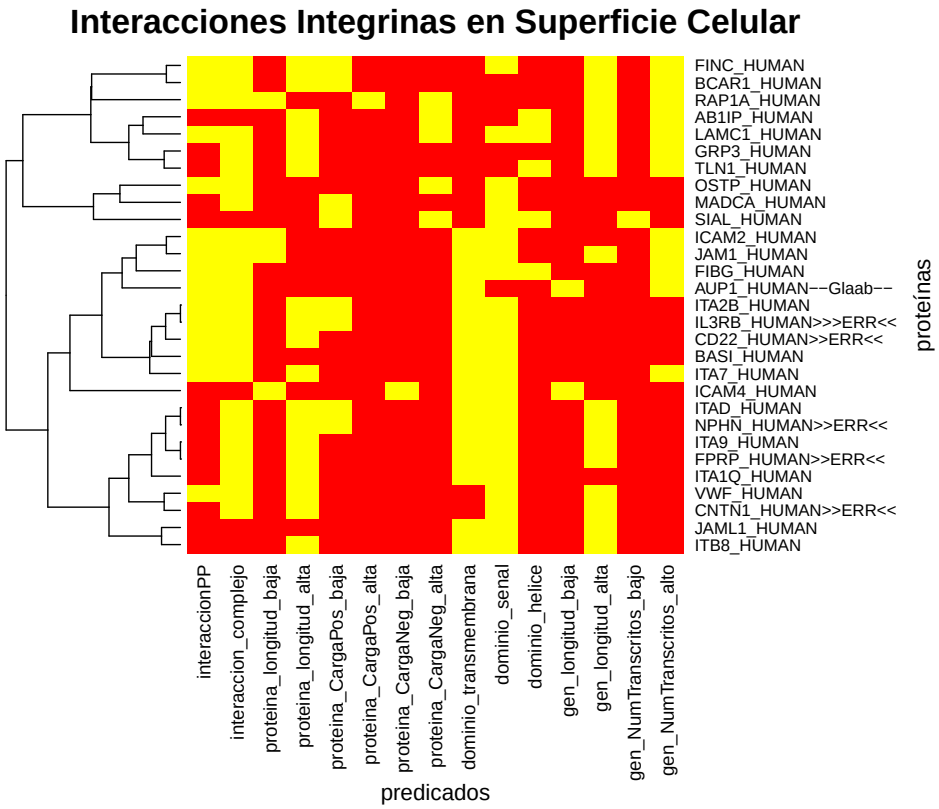


Figura 7.23: Mapa de agrupación de proteínas por propiedades simples. Se incluyen proteínas de la ruta original, y predichas por ERR o por Glaab et al. (con los sufijos $>> ERR <<$ y $--Glaab--$ en las etiquetas de las filas, respectivamente). Cada propiedad simple se representa con un predicado lógico (cada columna). Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0). Se usan los identificadores de UniProt. Ruta *Interacciones de las integrinas en la superficie celular*.

ser más parecidas que proteínas originales entre ellas.

Además, las proteínas predichas por el método de Glaab et al. (con el sufijo $--Glaab--$) presentan propiedades diferentes de las predichas por ERR, situándose ambas generalmente en regiones distantes del mapa coloreado.

En concreto en la figura 7.23 se observa que las proteínas predichas por ERR se distribuyen entre al menos dos o tres grupos diferentes de propiedades similares, siendo semejantes a distintas proteínas de la ruta original. En la figura 7.24, las proteínas predichas por ERR (y también las de Glaab et al.) están más dispersas aún, porque se usan 3 reglas de clasificación diferente para extender esta ruta. Aunque también puede influir que la ruta original sea de mayor tamaño que la anterior.

Por último, existen evidencias biológicas que relacionan funcionalmente algunas de estas proteínas predichas con las de la ruta original, lo cual se analiza en detalle en la siguiente sección.

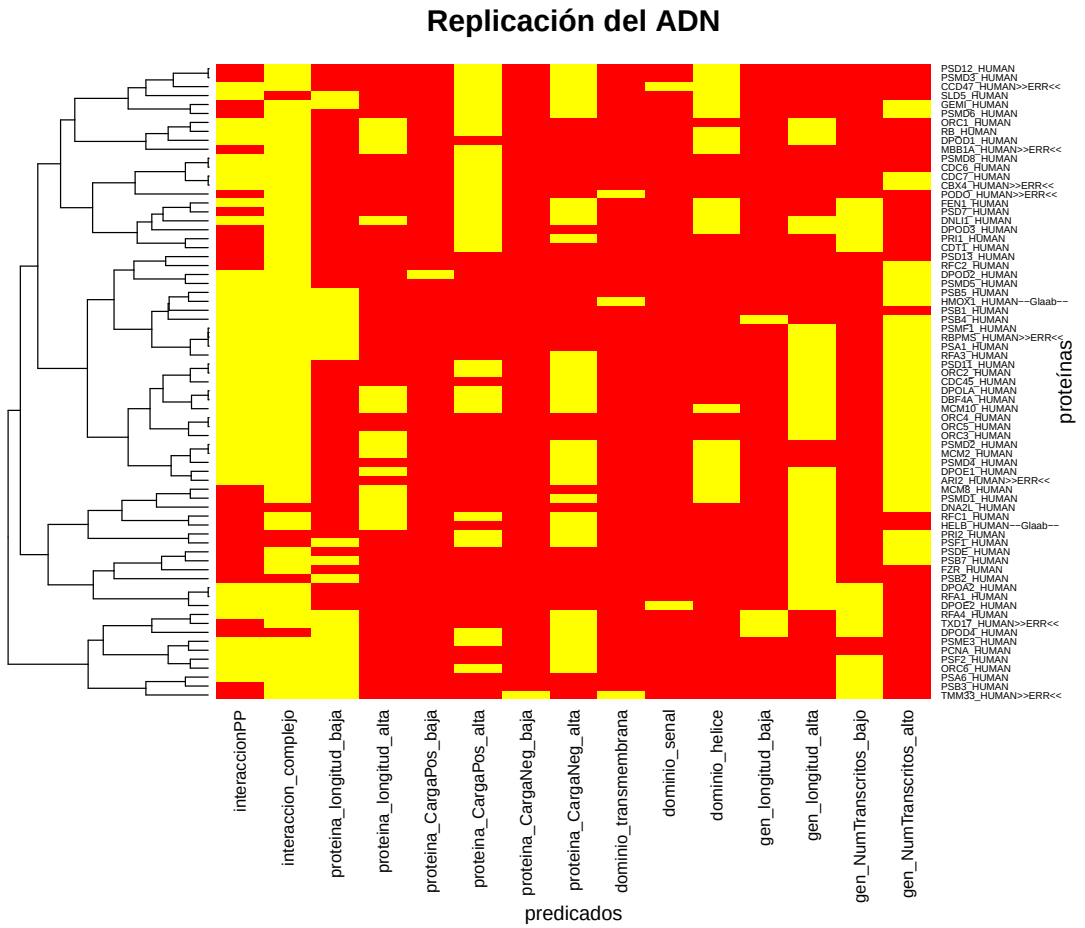


Figura 7.24: Mapa de agrupación de proteínas por propiedades simples. Se incluyen proteínas de la ruta original, y predichas por ERR o por Glaab et al. (con los sufijos $>> ERR <<$ y $--Glaab--$ en las etiquetas de las filas, respectivamente). Cada propiedad simple se representa con un predicado lógico (cada columna). Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0). Se usan los identificadores de UniProt. Ruta *Replicación del ADN*.

7.7.3. Predicciones “de novo”

En esta sección se exponen diversos casos concretos de proteínas analizadas, correspondientes a predicciones “de novo”, en las que hay una tendencia de correlación positiva entre la ruta extendida y las anotaciones funcionales en las bases de datos (UniProt) y en la literatura científica de las proteínas predichas. Como ejemplos de predicciones “de novo”del sistema ERR-PDR, se usan las rutas *Cadena de transporte de electrones*, *Replicación del ADN*, *Mantenimiento del telómero* e *Interacciones de las integrinas en la superficie celular*.

Los dos primeros ejemplos presentados incluyen una justificación breve y sencilla, mientras que para los otros dos ejemplos se plantea una propuesta de modelo de extensión de la ruta, acorde a las evidencias biológicas existentes para cada caso.

Cadena de transporte de electrones

En la ruta *Cadena de transporte de electrones*, las cinco proteínas predichas (CC078_HUMAN, CA151_HUMAN, A8MTT3_HUMAN, MANBL_HUMAN y SPAT9_HUMAN) se anotan como proteínas de membrana de paso simple, lo que está relacionado con las proteínas originales de la ruta, con la existencia del predicado `transmembrane_domain` en un 42 % de las 77 proteínas originales en esta ruta. La corta longitud de las secuencias de las proteínas es un hecho más para justificar biológicamente que las proteínas predichas son similares a las originales a nivel molecular, ya que la frecuencia del predicado `protein_length_low` es del 100 % en las proteínas predichas y del 78 % en la ruta original.

Así, para esta ruta existe una verificación biológica sencilla, porque existen anotaciones en las bases de datos referentes a dominios de membrana, y además coinciden con las frecuencias de propiedades simples de la secuencia (o predicados lógicos). Aquí se justifica también la utilidad de la representación relacional, mediante la lógica de predicados, que permite definir fácilmente términos en función de los cuales se quieren interpretar los resultados.

Replicación del ADN

La ruta de *Replicación del ADN* de Reactome, a menor nivel, se divide en tres: fases mitóticas M-M/G1, síntesis del ADN y regulación de la replicación del ADN. Así, las anotaciones de las 61 proteínas no redundantes originales de esta ruta son variadas. Entre ellas se observan 6 proteínas que se asocian al complejo de reconocimiento de los sitios de origen de la replicación (del inglés, *Origin Recognition Complex, ORC*), y 22 proteínas que están relacionadas con el proteosoma.

El ORC es un componente central para la replicación del ADN en eucariotas, localizado en el núcleo. Por su parte, el proteosoma es un complejo proteico grande presente en todas las células eucariotas y de otras especies. En eucariotas, los proteosomas están localizados en el núcleo y en el citoplasma [Peters et al., 1994]. Su función es degradar proteínas no necesarias o dañadas.

En los tres procesos comentados en los que se descompone la replicación del ADN, existen proteínas que son degradadas por el proteosoma 26S, como por ejemplo, ORC1_HUMAN, GEMI_HUMAN y CDC6_HUMAN. Este hecho explica la cantidad de componentes del proteosoma que están anotados en la ruta original. Además, como el proteosoma puede estar en el citoplasma, las proteínas de esta ruta a degradarse se tendrían que desplazar al citosol. Por ejemplo, el compuesto ORC1-ubicuitinado (ORC1_HUMAN marcada para degradación) que también se localiza en el citosol. Así, proteínas grandes localizadas en el núcleo, como el ORC1_HUMAN con 861 aminoácidos, probablemente no puedan difundir a través de la membrana nuclear para llegar al citosol. Por lo tanto, puede que alguna proteína de membrana esté implicada en su transporte, como las proteínas CCD47_HUMAN y TMM33_HUMAN, dos proteínas de membrana que predice el sistema ERR-PDR como pertenecientes a esta ruta, entre las ocho que propone. CCD47_HUMAN es una proteína de membrana de paso simple y también contiene un dominio hélice, según las anotaciones de UniProt. TMM33_HUMAN es la proteína transmembrana 33, que pertenece a la familia PER33/POM33, siendo una proteína de membrana multi-paso. Se puede observar la asociación entre las propiedades simples de secuencia de estas proteínas en el *clustering* jerárquico coloreado de la figura 7.24.

En conclusión, aunque no hay evidencias definitivas, las encontradas sugieren que algunas de las predicciones podrían tener sentido biológico.

Interacciones de las integrinas en la superficie celular

Las integrinas son los receptores que median la adhesión de la célula con la matriz extracelular, compuesta de diversas moléculas que se relacionan con la célula. En esta sección se analiza la ruta de Reactome *Interacciones de las integrinas en la superficie celular*.

Según las anotaciones en las bases de datos, de las cinco proteínas añadidas por ERR-PDR a esta ruta, cuatro son receptores de la superficie celular, para diferentes moléculas y proteínas. Además, tienen una arquitectura de membrana de paso simple. Dichas proteínas son IL3RB_HUMAN, CD22_HUMAN, NPHN_HUMAN y FPRP_HUMAN. IL3RB_HUMAN es la subunidad B de un receptor de citoquinas, CD22_HUMAN es un receptor de células-B, NPHN_HUMAN es un receptor específico de adhesión a la célula y FPRP_HUMAN es un receptor de prostaglandina F2. Por su lado, CNTN1_HUMAN (*contactin-1* o glicoproteína gp135) es una proteína situada fuera de la membrana (anclada mediante lípidos), pero relacionada con esta ruta, porque media las interacciones en la superficie celular durante el desarrollo del sistema nervioso.

La figura 7.25 representa parcialmente esta ruta, incluyendo algunas de las proteínas anotadas originalmente en Reactome, sus conexiones y las similitudes con las proteínas predichas por ERR-PDR. IL3RB_HUMAN y CD22_HUMAN, predichas por ERR-PDR, tienen un perfil de propiedades simples de secuencia similar a la proteína original de la ruta ITA2B_HUMAN (integrina alfaIIb-beta3). IL3RB_HUMAN es un receptor de alta afinidad para la interleucina-3, la interleucina-5 y el factor estimulante de colonias de granulocitos y macrófagos. CD22_HUMAN media las interacciones entre dos células B. Por otro lado, según muestra el panel B de la figura 7.25, CNTN1_HUMAN es semejante en propiedades a otra proteína diferente de la ruta original, VWF_HUMAN (multímero VWF), ambas situadas en la matriz extra-celular, fuera de la membrana. De este modo se verifica la diversidad de reglas buscada por el sistema ERR-PDR, para extender las rutas con proteínas heterogéneas a nivel molecular.

Se puede observar la cercanía de las proteínas mencionadas en la agrupación por propiedades simples de la figura 7.23.

Cabe destacar que, en contraposición a la coherencia biológica de las predicciones del sistema ERR-PDR en esta ruta, el método de Glaab et al. propone una proteína (AUP1_HUMAN) sin relación con las proteínas originales de la ruta *Interacciones de las integrinas en la superficie celular*. AUP1_HUMAN es una proteína de membrana de tipo III, no existiendo otras proteínas con esta característica en la ruta. Además, su localización es en la membrana del retículo endoplasmático, y no en la membrana celular. Por lo tanto, en este caso particular, tienen más lógica las proteínas predichas por ERR-PDR que la predicha por el método de Glaab et al.

Mantenimiento del telómero

Los telómeros son complejos proteína-ADN situados al final de los cromosomas lineales, siendo importantes para la estabilidad del genoma, al evitar la pérdida de información de los extremos y la fusión con otros cromosomas. El ADN de los telómeros humanos es una secuencia de 6 nucleótidos (TTAGGG), repetida cientos de veces. Los telómeros se van acortando en cada división celular, degradando los cromosomas. Cuando son demasiado cortos, la célula no se vuelve a replicar para evitar células erróneas, y se produce la muerte celular. Por lo tanto, para permitir las múltiples divisiones de la vida útil de las células, se necesita un mecanismo para mantener estable la longitud de los telómeros, evitando el envejecimiento y la

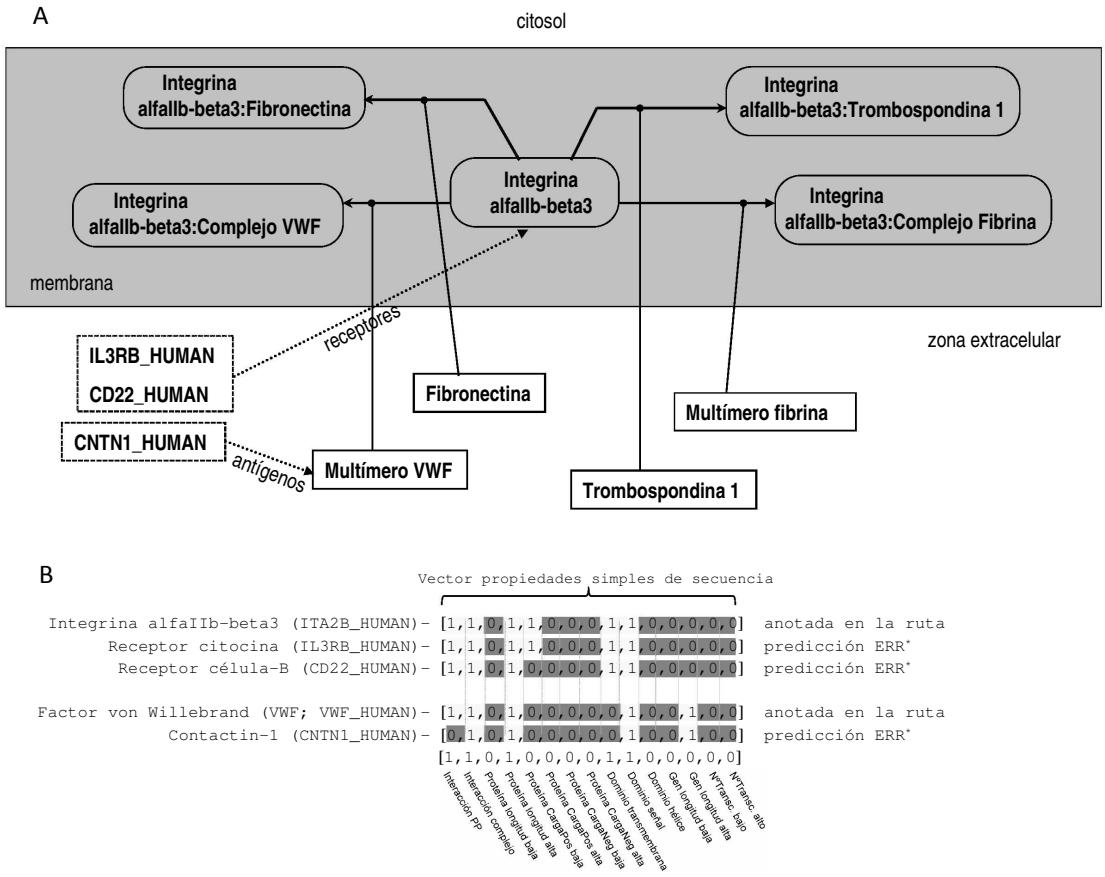


Figura 7.25: Ruta humana de *Interacciones de las integrinas en la superficie celular* de Reactome extendida por el sistema ERR-PDR. El panel A presenta un diagrama con algunas de las proteínas anotadas en la ruta originalmente, sus conexiones y tres proteínas predichas por ERR-PDR. Las líneas discontinuas representan las proteínas predichas por el sistema ERR-PDR. El panel B muestra una comparación entre los vectores de propiedades simples de secuencia, de las proteínas anotadas y de las predichas. Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0). El vector numérico de ejemplo es un vector de consenso, con la moda de los cinco vectores coloreados de arriba.

muerte celular.

El mecanismo principal para mantener los telómeros en humanos está basado en la telomerasa, una enzima que permite alargar los telómeros. La telomerasa es un complejo ribo-nucleo-proteína que incluye un dominio de transcriptasa inversa (del inglés, *TElomerase Reverse Transcriptase, TERT*) y una plantilla de ARN (del inglés, *TElomerase RNA Component, TERC*). La telomerasa usa la plantilla de ARN para añadir varias veces la secuencia TTAGGG que conforma los telómeros.

La actividad de la telomerasa es reducida o está ausente en tejidos normales con células maduras, una vez superada la fase de división celular desarrollada en las células somáticas. Sin embargo, se ha encontrado que en el 80-90 % de las células cancerígenas hay una mayor actividad de la telomerasa que en los tejidos sanos maduros [Kim et al., 1994]. Así, las células cancerosas proliferan indefinidamente y crean tumores, lo que es un punto clave de su malignidad. Es decir, dicha proliferación se debe a la elongación de sus telómeros por la

actividad continua de la telomerasa. Por lo tanto, la inhibición de la activación de la telomerasa es un enfoque novedoso para la lucha contra el cáncer [Philippi et al., 2010].

En humanos, continúa existiendo una pobre definición de los mecanismos de replicación del telómero, y se necesita más conocimiento sobre la regulación de la transcripción, de la traducción y de la post-traducción de las proteínas de enlace con el telómero [Xu, 2011]. Aún no se ha logrado comprender bien la recesión y síntesis de los telómeros, ni cómo se coordinan los pasos de elongación mediados por la telomerasa, por lo que podrían estar involucradas otras proteínas [Smogorzewska and de Lange, 2004]. Nuestro sistema ERR-PDR predice cinco proteínas para extender la ruta *Mantenimiento del telómero*, que hipotéticamente estarían implicadas en estos procesos relacionados con la estabilidad de los telómeros.

El análisis de esta ruta, en cuanto a la frecuencia de aparición de predicados (ver figura 7.21), revela que los predicados `complex_interaction` y `transcripts_low` se cumplen con una frecuencia del 100 % en las proteínas predichas (HPGDS_HUMAN, CSN4_HUMAN, PIAS4_HUMAN, DTX1_HUMAN y APBP2_HUMAN), y un 64 % y 49 %, respectivamente, en las 45 proteínas anotadas originalmente en esta ruta. Así, todas las proteínas añadidas tienen alguna interacción en un complejo. Cabe destacar que esta ruta es la única que se extiende usando el predicado `transcripts_low`. De forma que las proteínas predichas no tienen isoformas, codificando cada gen para un solo transcrito en cada caso.

Por otro lado, según las anotaciones de la base de datos InterPro, excepto HPGDS_HUMAN –una enzima bi-funcional: la prostaglandina sintetasa D hematopoyética (EC 5.3.99.2) y la S-transferasa Glutatión (EC 2.5.1.18)–, las cuatro proteínas restantes predichas, presentan patrones de enlace a ácidos nucleicos, lo cual es coherente con la definición de telómero como complejo proteína-ADN. Tales patrones son: el represor de la transcripción hélice-giro-hélice *Winged* de enlace al ADN; el patrón de secuencia SAP, un enlace al ADN tentativo encontrado en diversas proteínas nucleares involucradas en la organización cromosómica; los dominios *Zinc-finger*, ahora reconocidos como enlaces al ADN, ARN, proteínas y/o lípidos; la región repetida de péptido tetrático, la cual a pesar de mediar interacciones proteína-proteína y el ensamblaje de complejos multi-proteína en un gran rango de proteínas, adopta una disposición hélice-giro-hélice que se encuentra comúnmente en proteínas de enlace al ADN (para más detalles, ver las anotaciones de InterPro [Hunter et al., 2009] de estas proteínas).

No se conocen conexiones definidas entre la ruta *Mantenimiento del telómero* y la proteína predicha APBP2_HUMAN, la proteína 2 de enlace a proteínas amiloides, que podría desempeñar un papel en el transporte intracelular de proteínas. Sin embargo, es interesante destacar que tres de las proteínas predichas restantes (CSN4_HUMAN, PIAS4_HUMAN y DTX1_HUMAN) se han relacionado con la ruta de conjugación de la ubicuitina, según las anotaciones de UniProt. La ubicuitina (Ub) es una pequeña proteína cuya principal función es la de marcar otras proteínas para su degradación. El acoplamiento de la ubicuitina regula principalmente interacciones con otras macromoléculas, tales como enlaces proteosoma-sustrato o captación de proteínas para la cromatina [Hochstrasser, 2009]. Hay similitudes evidentes en las rutas involucradas en la activación y conjugación Ub y las proteínas tipo-ubicuitina (Ubl, del inglés *ubiquitin-like proteins*), con residuos de lisina particulares con proteínas de salida. No obstante, el mecanismo de intercambio entre proteínas SUMO (del inglés, *Small Ubiquitin-like Modifier*) y la ubicuitinación continúa sin estar claro [Bailey and O'Hare, 2005; Hochstrasser, 2009], por lo que más proteínas podrían estar implicadas en los procesos de ubicuitinación, para degradar proteínas como la telomerasa.

Aunque no existan suficientes evidencias experimentales, en la figura 7.26 se propone un papel tentativo de las tres proteínas predichas en la ruta de *Mantenimiento del telómero*

relacionadas con la ubiquitinación.

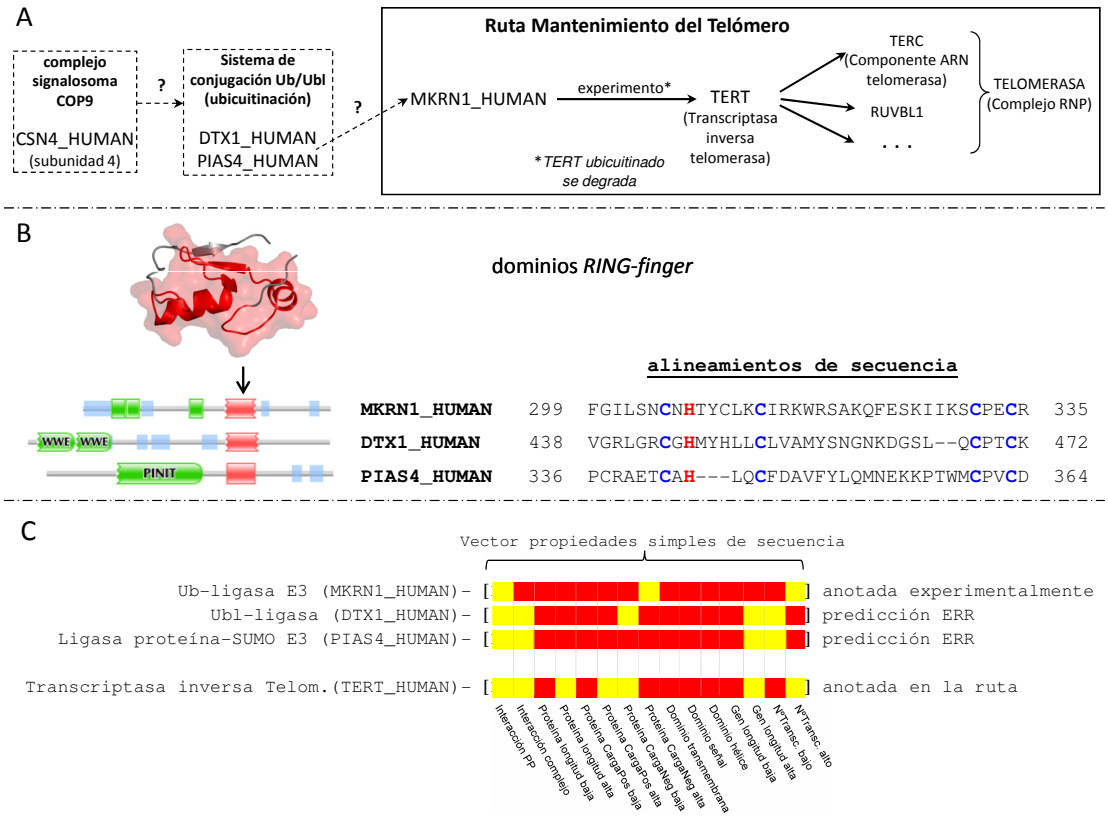


Figura 7.26: Ruta humana de *Mantenimiento del Telómero* de Reactome extendida por el sistema ERR-PDR. En el panel A, la relación tentativa de las proteínas predichas con la ruta original, así como la relación entre MKRN1 y TERT verificada experimentalmente [Kim et al., 2005] y anotada en UniProt como modificación postraduccional. Dentro del complejo RNP (ribo-nucleo-proteína) que constituye la telomerasa, TERT interacciona entre otras con TERC, la plantilla de ARN, y con RUVBL1, la proteína 1 RuvB, un componente básico de la cromatina propuesto para remodelar el complejo INO80, el cual está involucrado en la regulación de la transcripción, en la replicación del ADN y probablemente en la reparación del ADN. En el panel B, se presentan las tres proteínas propuestas para enlazar con TERT, con su composición de dominios (izquierda), incluyendo todas un dominio de tipo RING-finger (en rojo), y el alineamiento de secuencia del dominio en las tres proteínas (derecha), conservando el residuo esencial de histidina, marcado en rojo. En el panel C, se muestra una comparación de los vectores de propiedades simples de secuencia para algunas de las proteínas implicadas. Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0).

Según muestra la figura 7.26, para la ruta de *Mantenimiento del telómero* ERR-PDR predice a las proteínas CSN4_HUMAN, DTX1_HUMAN y PIAS4_HUMAN. Así, dentro de las proteínas predichas, CSN4_HUMAN es un componente del complejo signalosoma COP9 (involucrado en la regulación de la degradación de las proteínas) que es un regulador esencial de la ruta de conjugación de la ubicuitina en respuesta a daños en el ADN [Groisman et al., 2003]; PIAS4_HUMAN es una ligasa proteína SUMO E3 [Ihara et al., 2005]; y DTX1_HUMAN manifiesta una actividad Ub-ligasa *in vitro* [Takeyama et al., 2003], formando estas dos últimas parte del sistema de conjugación Ub/Ubl.

Por otro lado, se sabe que el dominio C-terminal (residuos 946-1132) de la telomerasa humana ha sido eficientemente ubiquitinado *in vivo* por la E3-ligasa MKRN1_HUMAN,

y que el dominio *RING-finger* de esta ligasa es esencial para la interacción física entre estas proteínas. De hecho, la mutación His307Glu en el dominio *RING-finger* abole su actividad de ubiquitinación [Kim et al., 2005]. Es decir, según representa el panel A de la figura 7.26, el dominio *RING-finger* de MKRN1 promueve la degradación de TERT (subunidad de la telomerasa) mediante la ubiquitinación de la misma, decrementando la actividad de la telomerasa, subsecuentemente reduciendo la longitud del telómero, y con ello facilitando la muerte celular. Las ligasas predichas por el sistema ERR-PDR (PIAS4_HUMAN y DTX1_HUMAN), poseen un dominio *RING-finger* igual que MKRN1, con un residuo conservado de Histidina (*H*), indispensable para la unión a TERT, aparte de varios residuos conservados de Cisteína *C* (marcados en rojo (H) y azul (C), respectivamente, en el alineamiento de las secuencias del panel B, en la figura 7.26). Además, el perfil de propiedades simples de secuencia mostrado en el panel C de la figura 7.26, muestra la similitud molecular entre las tres proteínas mencionadas, y a su vez distinto al de TERT, otra proteína de la ruta. Por lo tanto, si MKRN1_HUMAN toma un papel importante en la modulación de la longitud del telómero, por la existencia del dominio *RING-finger*, y comparte con PIAS4_HUMAN y DTX1_HUMAN este dominio esencial para la interacción con la telomerasa, es posible que estas proteínas sugeridas por ERR-PDR también estén implicadas en la estabilización de la longitud del telómero.

También, recientemente se ha propuesto una conexión entre el mantenimiento de la estabilidad del genoma y la conservación evolutiva de la familia de las ubicuitinas (en particular, las Ub-ligasas dirigidas a proteínas SUMO) [Nagai et al., 2011]. Esto representa una evidencia más para justificar que estas tres proteínas anotadas con la ubiquitinación puedan estar involucradas en la ruta de *Mantenimiento del telómero*.

En conclusión, aunque se necesitan más evidencias para corroborar esta hipótesis, junto con las anotaciones y los hallazgos en la literatura, la conservación del dominio RING finger entre la proteína MKRN1_HUMAN (que degrada la telomerasa) y las proteínas predichas por nuestro sistema, podría indicar que dichas proteínas son interesantes para investigaciones sobre la supresión de tumores y la prevención del envejecimiento.

7.8. Conclusiones y Discusión

Este capítulo presenta un sistema de extensión de rutas basado en un enfoque de predicción de función que confía en combinaciones de propiedades simples asociadas a cada proteína. Las predicciones se basan principalmente en características extraídas de la secuencia (incluido el número de isoformas), pero también incluye algunas propiedades relacionadas con la posición de las proteínas en la red de interacción proteína-proteína y en los complejos de proteínas, es decir, parejas de interacción con sus correspondientes propiedades. Esta información relacional hace que este sistema sea diferente de otros basados sólo en características individuales. Con estas propiedades, se buscan proteínas específicas similares a nivel molecular a alguna proteína de la ruta original, en vez de satisfacer características generales para todas las proteínas de la ruta completa, a nivel de proceso biológico.

Desde una representación relacional, el sistema de Extensión basado en Representación Relacional propuesto en esta tesis expande 28 rutas de Reactome en humanos, con 383 proteínas, dado el umbral específico elegido (definido en la sección 7.2.6). Como cada predicción de ERR tiene un valor de confianza asociado, se podría elegir un umbral más restrictivo, pero teniendo en cuenta que varias proteínas comparten el mismo valor de confianza (aquellas que se clasifican con la misma rama del árbol de decisión). El nivel de extensión

cambia de ruta a ruta, en términos de rendimiento y de las diferentes propiedades moleculares de las proteínas añadidas. Permitiendo una pérdida de precisión en el sistema, se consigue una expansión con mayor variabilidad molecular por ruta, lo cual incrementa el significado biológico de la extensión de las rutas. La figura 7.27 muestra un diagrama en el contexto de la Biología de Sistemas. En ella se representan las extensiones de rutas de ERR-PDR con alta fiabilidad, incluyendo las proteínas originales y las añadidas, realzando las principales rutas discutidas en la sección de relevancia biológica (sección 7.7).

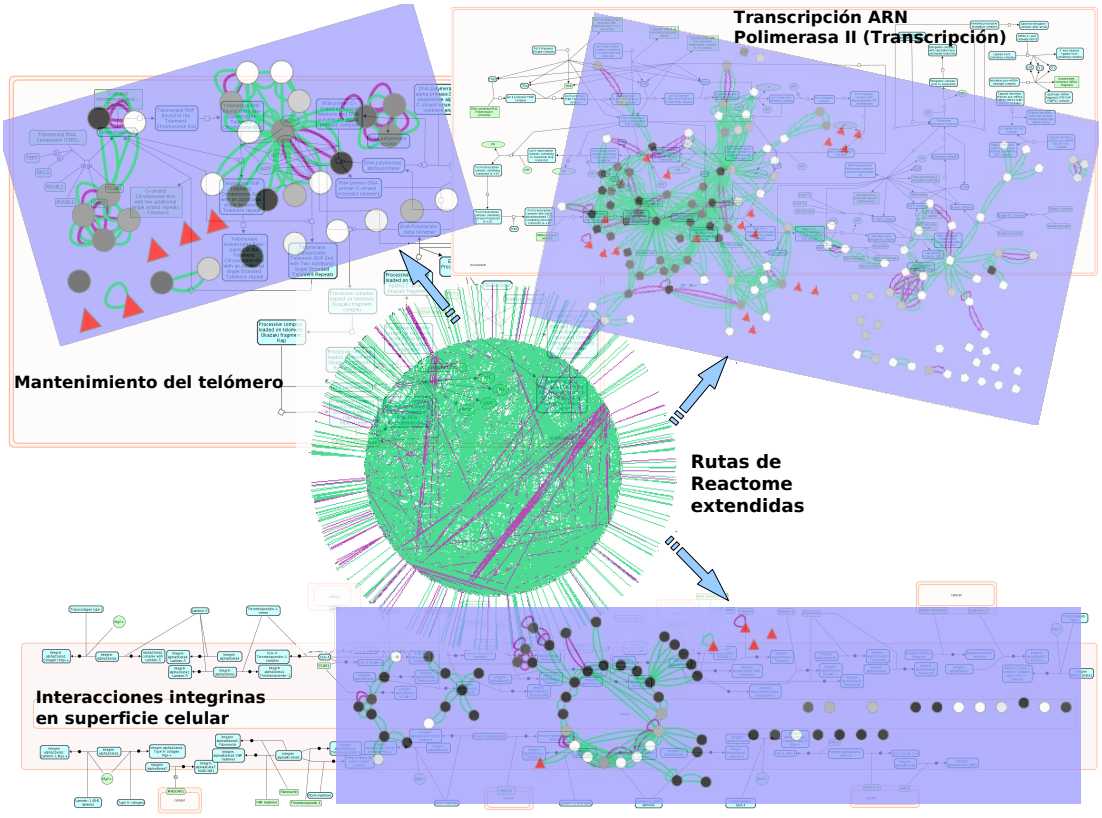


Figura 7.27: Red de interacción para las rutas extendidas por el sistema ERR. El diagrama central muestra la red de interacción con todas las proteínas de rutas extendidas con alta fiabilidad. Alrededor, se destacan tres subconjuntos de la red central, correspondientes a tres rutas individuales discutidas en la sección de relevancia biológica. Los enlaces verdes son interacciones en complejos, y los morados de interacciones proteína-proteína. Los triángulos rojos representan las proteínas añadidas por ERR, y los círculos grises las proteínas originales de la ruta. La escala de grises se corresponde con diferentes reglas de predicción, es decir, distintas combinaciones de propiedades simples.

Aunque el rendimiento global del sistema ERR no sea muy alto, se debe tener en cuenta la definición original de una ruta biológica, porque determinan los datos con los que se aprende. Dichas definiciones dependen parcialmente de las opiniones subjetivas de los expertos que diseñan las rutas [Lu et al., 2007], por lo que quizá no representan un patrón 100 % fiable (del inglés, *gold standard*). Consecuentemente, en este contexto es muy difícil alcanzar altas tasas de acierto en predicción, justificado por la dependencia de la calidad de los datos de entrenamiento [Jansen and Gerstein, 2004].

Con respecto a la relevancia de las características, los resultados de predicción señalan que las interacciones no son la única propiedad útil en el proceso de aprendizaje, ya que estas

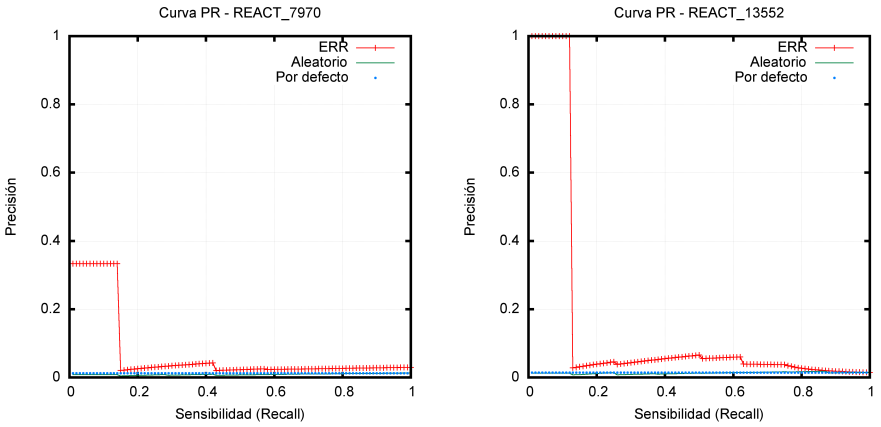
relaciones por sí solas no son capaces de lograr rendimientos de extensión tan altos como el sistema completo, a pesar de que la interacción entre proteínas tiende a ser una característica frecuente. Por lo tanto, en los problemas de extensión de rutas es necesario incluir propiedades de la secuencia en el aprendizaje, aparte de las interacciones, ya que una ruta no tiene todas sus proteínas conectadas con alguna otra de la ruta.

Como se esperaba, las extensiones de los sistemas presentados en este capítulo son diferentes a las que se alcanzan con el método de Glaab et al., un enfoque distinto basado sólo en redes de interacción. El solapamiento entre ambos sistemas (ERR y Glaab et al.) es escaso, incluso aunque ERR también usa algún conocimiento de interacciones. Además, el nuevo sistema ERR proporciona una mayor diversidad de funciones, no restringiéndose a buscar en un espacio próximo (lo que se ilustra con menos proteínas predichas en la intersección entre rutas). Como el sistema ERR se centra en las particularidades moleculares de proteínas específicas, y no en proteínas que conectan diferentes rutas, las extensiones de ERR consiguen evitar el solapamiento original entre diferentes rutas.

Como las rutas son heterogéneas a nivel molecular, se necesita un análisis por separado para cada una de ellas, como se hace a lo largo del capítulo, profundizando en algunas de ellas. Los nuevos componentes tentativos predichos por ERR proporcionan información explicativa útil para algunos de los procesos celulares estudiados en este capítulo. Son notables las proteínas predichas para las rutas *Transcripción*, *Expresión génica*, *Cadena de transporte de electrones*, *Replicación del ADN* y especialmente para *Interacciones de las integrinas en la superficie celular* y *Mantenimiento del telómero*. De forma que las anotaciones de UniProt y los hallazgos en la literatura científica, cuando se combinan con los resultados de ERR, incrementan la fiabilidad de la asignación de una proteína a una ruta.

Estos últimos resultados también confirman que un bajo AUPRC, tal como el obtenido para la ruta de *Mantenimiento del telómero* (0,0713 sobre 1), no siempre implica una mala predicción en las proteínas específicas en términos de utilidad biológica. Es decir, "falsos positivos computacionales, no son falsos positivos biológicos"[Mateos et al., 2002]. Incluso en términos de fiabilidad computacional, una mala predicción global (un bajo AUPRC), puede no serlo para proteínas específicas. Porque en la región izquierda de la curva (con baja cobertura o *recall*), la precisión de las predicciones puede ser alta, como sucede en la curva PR individual de la ruta de *Mantenimiento del telómero* (ver figura 7.28(a)). Así, tras seleccionar un umbral, la precisión en test de la/s regla/s finalmente aplicadas, puede corresponder con dicha región izquierda de la curva PR. Por lo tanto, las proteínas que ERR añade a la ruta siguen la parte de la curva de alta precisión. Incluso se alcanza la precisión máxima para las proteínas predichas en la ruta de *Interacciones de las integrinas en la superficie celular* (ver región izquierda de la figura 7.28(b)). Este hecho justifica de nuevo que las bajas curvas PR no sean determinantes en la evaluación de bondad del sistema ERR.

Finalmente, se debe tener presente que el sistema ERR, sofisticado en cierta manera, se puede usar de forma genérica para extender otras bases de datos de rutas o procesos celulares, siempre que las proteínas anotadas disponibles sean suficientes para aprender. Quizá este procedimiento podría contribuir a unificarlas. De forma más general, este procedimiento de Anotación basado en Representación y Aprendizaje Relacional se podría aplicar fácilmente a diferentes vocabularios, para anotar funcionalmente genes y proteínas no caracterizados, siguiendo las estrategias propuestas en el capítulo 8. Incluso se podrían compartir datos que ya se encuentran representados relacionalmente en la base de conocimiento.



(a) Ruta *Mantenimiento del telómero*.

(b) Ruta *Interacciones de las integrinas en la superficie celular*.

Figura 7.28: Ejemplos de curvas PR individuales con alta precisión a baja cobertura. Sistema ERR-PDR.

Capítulo 8

Otros Enfoques de Aprendizaje Automático en Bioinformática

El objetivo de este capítulo es explorar diversos enfoques en la representación del conocimiento y en las técnicas de aprendizaje, para analizar y comprender el uso de la inteligencia computacional en la resolución de problemas bioinformáticos. Especialmente se analizan los escenarios que implican aprendizaje multi-clase y multi-etiqueta (frecuentes en anotación funcional de proteínas y genes), y representación y aprendizaje relacional (por la elevada presencia de relaciones en los datos biológicos).

En definitiva, se quiere aprender de los diversos planteamientos computacionales considerados a lo largo de la tesis, que han sido la base imprescindible para conseguir encontrar una solución válida para los dos problemas elegidos de Biología Molecular con Aprendizaje Automático (capítulos 6 y 7). A partir del análisis de estos enfoques, se pretenden definir directrices generales que sirvan como guía para la aplicación de Aprendizaje Automático a otros problemas de anotación funcional en bioinformática (o incluso en algunos casos a un dominio diferente), describiendo en cada caso la estrategia propuesta según las características a afrontar del problema por resolver.

La mayor parte de los enfoques estudiados (excepto la sección 8.1) se analizan en el contexto del problema de extensión de rutas biológicas (ver capítulo 7). Debido a su mayor complejidad e interés analítico, por la existencia de muchas más relaciones entre los datos (las que se usan y las que se podrían añadir fácilmente) y por tratarse de una tarea de clasificación multi-clase y multi-etiqueta.

Este capítulo se estructura como se detalla a continuación. La sección 8.1 aplica Programación Genética, como otra técnica de inteligencia computacional para integrar atributos numéricos, en un contexto con muchos valores desconocidos con semántica biológica asociada. La sección 8.2 propone aprender con un solo clasificador en vez de usar uno independiente para cada clase, en un problema multi-clase, y también con las restricciones de un problema multi-etiqueta. La sección 8.3 analiza diferentes transformaciones de una representación relacional a proposicional, pero siempre basándose en la extracción de patrones frecuentes. La sección 8.4 estudia distintas combinaciones de representación y aprendizaje, aplicando sólo aprendizaje relacional, o sólo aprendizaje proposicional, en vez de la combinación híbrida utilizada en el capítulo 7 de la tesis. De la sección 8.5 a la 8.8 se analiza cómo afecta la inclusión de más relaciones biológicas como información de entrada para anotar funcionalmente con un método basado en representación y aprendizaje relacional. Finalmente, se expone un resumen de las conclusiones del capítulo y, como contribución de

este capítulo, una tabla recopilatoria que sugiere estrategias de aplicación de AA según las características del problema biológico.

Metodología de Evaluación

Este apartado describe la metodología de evaluación seguida en este capítulo, en el que se comparan distintas soluciones para un mismo problema. Basándose en los criterios presentados en el capítulo 4, los resultados experimentales se evalúan mediante la comparación de los sistemas propuestos con otros, como por ejemplo una solución previa, como se decide hacer en este capítulo. Esta solución previa puede ser una solución base, la mejor encontrada o no, pero con las mismas condiciones de partida que lo que se quiere comparar.

A parte de la elección del sistema con el que comparar, se deben definir las medidas de evaluación. Dado que los dos problemas tratados en esta tesis son muy diferentes, los métodos de comparación y medidas de evaluación también deben divergir. Para los enfoques de AA evaluados con el problema de predicción de asociaciones funcionales entre pares de proteínas (sección 8.1) se compara con una solución base, más simple (con menos atributos) que la presentada en el capítulo 6, pero obtenida en las mismas condiciones que el nuevo enfoque propuesto con el que se compara. Como medidas de evaluación se utilizan: para el rendimiento, la tasa de aciertos en el conjunto de test, y para la interpretación, el tamaño de la solución, su facilidad de interpretación y la frecuencia de aparición de cada operador.

Para los enfoques de AA evaluados con el problema de extensión de rutas (secciones 8.2 a 8.8), la mayoría de las veces se compara con la solución del sistema que prioriza rendimiento y cobertura (ERR-PRyC), descrito en el capítulo 7. No obstante, si el análisis influye en el número de rutas extendidas o la cantidad de reglas empleadas, se compara también con los valores del sistema que prioriza la diversidad de reglas (ERR-PDR), el cual está enfocado para obtener mejores resultados en este aspecto. De esta forma, en general se hereda la configuración completa del sistema ERR-PRyC, a excepción del parámetro afectado por el enfoque propuesto, cuyos efectos se quieren evaluar siguiendo el método científico. Para mantener una comparativa coherente y estable, no se alterna el uso de las configuraciones de los sistemas ERR-PRyC y ERR-PDR. Se puede asumir que si mejora la diversidad de reglas sobre ERR-PRyC, también lo hará sobre ERR-PDR; ya que el sistema ERR-PDR sigue una construcción incremental a partir de una configuración de ERR-PRyC aceptable, una vez que se han superado unos mínimos (en rendimiento, solapamiento, etc.).

Aunque para algún enfoque propuesto pudiera ser mejor otra configuración distinta a la de ERR-PRyC, es inviable probar todas las combinaciones posibles, además de no ser un objetivo de esta tesis. A lo sumo, si las limitaciones computacionales no permiten mantener la configuración exacta, se opta por otra más sencilla, relajando el valor en un parámetro concreto. Por ejemplo, en el sistema ERR-PRyC la profundidad en la extracción de patrones es muy elevada, y puede ser inviable aplicarlo cuando existen muchas más relaciones y combinaciones posibles de predicados (como sucede en las secciones 8.5.2, 8.6 y 8.7).

Respecto a las medidas de evaluación, se decide emplear un subconjunto básico de las ya usadas en el capítulo 7. Esporádicamente se incluye una figura adicional para evaluar un aspecto relevante en la sección específica, no analizado con las medidas básicas. Se definen como básicas AUPRC, AUROC, nº de rutas extendidas (es decir, nº de clases predichas) –total y con más de una regla–, nº de proteínas añadidas –total y porcentaje de proteínas diferentes–, similitud semántica con respecto a la ruta original y solapamiento entre las proteínas añadidas a cada ruta. Para las dos últimas medidas se utilizan representaciones gráficas definidas en el capítulo 7. Este subconjunto de medidas básicas seleccionadas se muestra en todas las

secciones, aunque en cada caso sólo se comentan las medidas que son importantes para ese análisis. Se puede observar que este grupo de medidas básicas incluyen algunas sobre el conjunto de test y otras sobre el de aplicación, porque en Biología Molecular es muy importante la interpretación de los resultados sobre el problema en cuestión, como expone el capítulo 4.

8.1. Programación Genética para Predicción de Asociaciones Funcionales entre Pares de Proteínas

Motivación-Hipótesis: ¿Se puede mejorar la interpretación de los resultados, o la predicción, o incluir más contenido semántico, usando otro enfoque de Inteligencia Artificial como es la Programación Genética? ¿Se puede conseguir una gestión de los valores desconocidos que respete la semántica biológica de los mismos?

8.1.1. Enfoque

En el capítulo 6 se afronta el problema de predicción de asociaciones funcionales entre pares de proteínas, integrando varios métodos disponibles para unificar las predicciones existentes. Se enfoca como un problema de clasificación binaria, resuelto mediante algoritmos de aprendizaje automático. El objetivo de esta sección es aplicar Programación Genética (PG) a este problema biológico debido a la flexibilidad de esta aproximación para adecuarse a las características del problema, como por ejemplo una gran cantidad de valores desconocidos [García-Jiménez et al., 2008b,a].

La PG es una técnica que evoluciona automáticamente programas de ordenador [Koza, 1992]. En esta sección, la PG se usa para obtener una ecuación equivalente a un clasificador binario, que determine si un par de proteínas dado presenta alguna asociación funcional.

Una de las razones para elegir la PG es que esta técnica permite al diseñador definir las primitivas según los requisitos del problema. Por ejemplo, para el problema de predicción AFPP se define el operador *if_desconocido (if_?)* (explicado en detalle en las secciones 8.1.2 y 8.1.4) para intentar resolver el inconveniente de los valores desconocidos, lo cual es una cuestión relevante en este dominio biológico, porque hay una gran cantidad de ellos en los conjuntos de datos. También, con la PG el criterio a optimizar se puede definir en la función de evaluación, en lugar de confiar sólo en la tasa de aciertos media que se usa típicamente en Aprendizaje Automático.

Se denomina valor desconocido al vacío de información en un atributo para alguna de las instancias. Las aproximaciones más usadas para gestionar los valores desconocidos en Aprendizaje Automático son (1) ignorar la instancia completa o (2) rellenar con el valor medio del atributo.

El primer enfoque es adecuado cuando hay pocos valores desconocidos. Sin embargo, en el problema de predicción AFPP casi todas las instancias tienen algún valor desconocido, de forma que si se ignoran todas ellas, los datos se reducen considerablemente, a menos del 0,005 % del tamaño original.

La segunda propuesta consigue una aproximación adecuada cuando hay ruido durante la recopilación de los datos, y consecuentemente algunos valores se pierden u olvidan. Pero este tampoco es el caso del problema a resolver, porque no refleja la semántica de los datos reales: la mayoría de los valores desconocidos en el problema de predicción AFPP representan datos *no-existentes* en una base de datos particular (en contraposición a perdidos u olvidados). Esto se debe a que las fuentes de datos, es decir, la salida de varios métodos computacionales

de predicción, proporcionan un valor sólo si se satisfacen todas las condiciones del método. Algunas de estas restricciones son disparar un evento o alcanzar un número mínimo de ortólogos en el alineamiento múltiple de secuencia de las proteínas del par (ver la descripción de los métodos en detalle en la sección B.4). Por lo tanto, como no se puede suponer cualquier valor medio como válido, la mejor solución es gestionar los valores desconocidos como valores especiales.

En este trabajo se propone una nueva forma de manejar los valores *no-existentes* de los atributos, que consiste en preservar como tal los desconocidos en los conjuntos de datos (representados por *'?'*). Al usar esta nueva aproximación se obtiene una representación con más sentido en términos de interpretación biológica. Otra opción para gestionar los valores desconocidos de forma especial es reemplazar los valores *no-existentes* con una marca numérica específica. Esta última opción tiene una desventaja frente a la nueva propuesta, consistente en que los valores numéricos pasan a tener dos interpretaciones semánticas diferentes: los valores reales y las marcas. No obstante, estas dos propuestas para gestionar valores desconocidos de forma especial se evalúan en la sección 8.1.4.

Además, es bien sabido que la PG sufre el problema del crecimiento desmesurado (del inglés, *bloat problem*) [Mahler et al., 2005]. Esto es, los individuos de PG tienden a crecer en tamaño sin una ganancia aparente en la evaluación. De ahí, para intentar mejorar la tasa de aciertos y la legibilidad de las ecuaciones evolucionadas por la PG, se usa el mecanismo de control del crecimiento Tarpeian, el cual sesga la evolución hacia soluciones sencillas [Poli, 2001, 2003]. También se suele esperar que el método Tarpeian acelere la evolución de las soluciones.

Programación Genética

La PG es un paradigma evolutivo que aplica algoritmos genéticos para generar automáticamente programas de ordenador [Koza, 1992]. Cada individuo de la población se representa tradicionalmente con una estructura de árbol, con terminales en las hojas y operadores o funciones en los nodos internos del árbol. La PG permite definir los operandos y operadores necesarios para cada tarea que se quiera resolver, según el dominio de aplicación. La evaluación, necesaria para la evolución genética, se determina por medio del rendimiento del individuo en la tarea concreta.

Con frecuencia, en PG se produce un crecimiento desmesurado del árbol, reduciendo en gran medida la velocidad del proceso evolutivo [Mahler et al., 2005]. Este crecimiento descontrolado presenta tres efectos negativos. Primero, soluciones difíciles de comprender, enormes y con fragmentos inútiles. Este aspecto podría ser muy importante en la predicción de interacciones y asociaciones funcionales entre proteínas, si se quiere comprender las razones de lo aprendido por la PG después de la evolución. Segundo, el proceso evolutivo se hace muy lento, porque se invierte mucho tiempo en evaluar los individuos más grandes de lo normal. Finalmente, en el contexto de los problemas de clasificación, los individuos muy grandes podrían tener una tasa de aciertos baja, porque tienden a estar sobre-entrenados.

En este apartado se aplica el método Tarpeian [Poli, 2001, 2003], una técnica bien fundada de control del crecimiento desmesurado. Brevemente, este método aborta estocásticamente algunos individuos durante el proceso evolutivo, si el tamaño de su árbol es mayor que la media de la población de la última generación (en nodos o profundidad). Así, el tamaño de la solución se limita de forma flexible y, la disminución del tamaño del árbol, mejora su interpretación. Con este método también se reduce el tiempo de ejecución, ya que los individuos abortados no se evalúan, asignándoles directamente el peor valor posible de evaluación. Además, en las

tareas de aprendizaje, reducir el tamaño del árbol es semejante a seguir la regla de *la navaja de Occam*, pudiendo mejorar la tasa de aciertos de la predicción mediantes soluciones más sencillas.

8.1.2. Configuración Experimental

Esta sección describe los elementos necesarios para aplicar PG a la resolución del problema de predicción de asociaciones funcionales proteína-proteína descrito en el capítulo 6. La herramienta de PG usada en la fase experimental es *lil-gp* 1.1 [Zongker and Punch, 1998], que está basada en los dos primeros libros de Koza [Koza, 1992, 1994].

Conjunto de Datos

Al presentar esta sección una prueba de concepto, se compara con una versión previa del predictor final descrito en el capítulo 6, como ya se ha comentado. En esta versión más simple, las fuentes de datos y la construcción de los conjuntos de datos siguen los mismos criterios que se definen en la sección 6.2. Sólo se diferencian de la configuración final en que se usan 10.000 ejemplos para el entrenamiento y 10.000 para el test, con una distribución al 50 % entre positivos y negativos, y sólo con 9 atributos por par de proteínas. Los 10 atributos de ranking de predicciones centrado en la proteína no se incluyen, y por tanto no se aplica el filtro de pares con un ranking menor a los 100 primeros.

Codificación de la Solución

En la PG es necesario definir los elementos (es decir, los terminales y los operadores) que forman parte de los árboles que representan los distintos individuos de la población.

1. Terminales

   Con estos 9 elementos se rellenan las hojas de los árboles:

   - 5 grados de asociación, proporcionados por los 5 métodos individuales de predicción, descritos en la sección 6.2.1.
   - 4 (2 por proteína) propiedades de las secuencias de proteínas: longitud y nº de ortólogos.
   - 1 terminal ERC (del inglés, *Ephemeral Random Constant*, constante aleatoria efímera) que representa a cualquier constante numérica aleatoria, la cual puede aparecer varias veces a lo largo del proceso evolutivo. Su valor se establece en el rango [0, 1].

   Un requisito típico en PG es que las operaciones sean cerradas, es decir, todos los terminales deben tener siempre un valor en cualquier instancia de entrada. Por lo tanto, la alta cantidad de valores *no-existentes* en el dominio de predicción AFPP se deben gestionar de una manera especial, como ya se ha mencionado. En un primer enfoque, dichos valores se rellenan con una marca específica: una constante numérica con un valor muy distante al del resto de características (0 ó -1, según cual sea el valor mínimo alcanzado en cada terminal). Además, todos los terminales se normalizan, para homogeneizar los resultados.

2. Operadores

   Se usan los siguientes:

   - Operadores aritméticos: suma, resta, multiplicación y división protegida (que controla la división por cero).
   - Operadores condicionales: `Si(a>=b) Entonces x; Si no y`. Sólo se considera éste, pues la comparación contraria sería redundante.
   - Operador *if_?*: `Si(k es desconocido) Entonces x; Si no y`. Es un nuevo operador específico, diseñado para este dominio. Este operador se define como segundo enfoque para gestionar los valores desconocidos o *no-existentes*, haciéndolos muy diferentes del resto. Así, cuando se usa este operador, los valores desconocidos se conservan, sin reemplazarlos con ninguna constante numérica. La forma de proteger el resto de operaciones, frente a un valor desconocido, consiste en devolver siempre como salida el valor desconocido (*'?'*), si alguno de los operandos es desconocido. Con esta implementación se valora de la misma forma que hayan aparecido uno o varios valores desconocidos, en cualquier combinación con operadores a lo largo de una secuencia concreta.

Proceso Evolutivo

Para predecir asociaciones funcionales entre dos proteínas *(p1, p2)*, se aplica el individuo evolucionado f sobre las proteínas, y se usa un umbral para determinar la clase positiva o negativa. Así, *Si* $(f \geq$ *umbral)* $\Longrightarrow$ *(p1, p2) están asociadas funcionalmente; Si no* $\Longrightarrow$ *(p1, p2) no están asociadas funcionalmente*. En todos los experimentos que se presentan en este trabajo el umbral es 0,5.

En este trabajo la función de evaluación es la tasa de aciertos, es decir, el porcentaje de instancias clasificadas correctamente, según la ecuación 8.1.

$$\textbf{evaluación} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8.1}$$

donde TP son verdaderos positivos, TN son verdaderos negativos, FP son falsos positivos y FN son falsos positivos [Fawcett, 2003].

En el proceso evolutivo, hay muchos parámetros a establecer, dando lugar a diferentes configuraciones para los experimentos. Los parámetros principales se ajustan con el rango de valores que muestra la tabla 8.1 (segunda columna), encontrando una configuración base (ver tabla 8.1, tercera columna) apropiada para los experimentos presentados en las siguientes secciones.

En el manual de la herramienta *lil-gp* [Zongker and Punch, 1998] aparece una lista completa de parámetros, con su significado y descripción detallada. Adicionalmente se añade un nuevo parámetro a la herramienta *lil-gp*: el factor Tarpeian. Se define como la probabilidad de abortar un individuo si su tamaño es mayor que la media. Los parámetros de la configuración base presentados en la tercera columna de la tabla 8.1 se obtienen empíricamente, excepto la profundidad máxima y el método de selección de individuos, que son los valores por defecto del libro de Koza previamente mencionado [Koza, 1994].

8.1.3. Comparación de PG con otras Técnicas de AA

Esta sección presenta los resultados obtenidos tras aplicar PG al problema de predicción de AFPP. Todas las configuraciones mostradas proceden de un promedio de 30 ejecuciones de

Tabla 8.1: Valores de los principales parámetros de configuración en la solución de AFPP con Programación Genética. Rango de valores (segunda columna) y valores en configuración base sin control de crecimiento (tercera columna).

Parámetro	**Rango de valores**	**Valor**
Tamaño población	1.000-25.000	1.000
Nº generaciones	15-250	50
Profundidad máxima	17	17
Nº nodos máximo	25-300	200
Operadores del árbol	+, -, *, /, $\geq$, *if_?*	+, -, *, /, $\geq$
Probabilidad operadores genéticos	cruce (0,3-0,9)	cruce (0,5)
	reproducción (0,1-0,4)	reproducción (0,1)
	mutación (0,0-0,4)	mutación (0,4)
Método selección individuos	torneo (tamaño=7)	torneo (tamaño=7)
Factor Tarpeian	0,0-0,9	0,0

PG. En la configuración base, la tasa de aciertos en test es del 60,83 % en media y de 61,44 % en la mejor ejecución encontrada, con una varianza muy baja.

La tabla 8.2 resume los resultados de varios algoritmos de Aprendizaje Automático (de la herramienta Weka [Witten and Frank, 2005]) para comparar con la PG. Todos los parámetros siguen los valores por defecto de Weka.

Tabla 8.2: Comparación cuantitativa entre Programación Genética y Aprendizaje Automático sobre el conjunto de test. Los valores están medidos en porcentaje. La fila PG presenta los resultados en media/mejor ejecución. Los algoritmos de Aprendizaje Automático usados son de diferentes tipos: ADTree, un árbol de decisión; AODE, método bayesiano; Kstar, un algoritmo de razonamiento basado en casos; MLP, una red de neuronas; PART, un método de reglas de decisión; *SimpleLogistics*, un método de regresión logística; y SMO, máquinas de vector de soporte (SVM). Ver la sección 6.2 para obtener la referencia de cada algoritmo.

Algoritmo	**Tasa de aciertos**	**Tasa de aciertos con valores desconocidos**	**Sensibilidad** (TP/TP+FN)	**Especificidad** (TN/TN+FP)
PG	60,38 / 61,44	60,67 / 61,22	58,87 / 63,54	62,62 / 59,34
ADTree	60,02	60,35	64,56	55,48
AODE	61,32	58,99	48,60	74,04
KStar	61,60	58,92	60,24	62,96
MLP	58,22	60,00	20,40	96,06
PART	61,96	58,33	60,84	63,08
SimpleLogistic	60,70	57,61	56,34	65,06
SMO	59,96	57,62	56,98	62,94

La segunda columna de la tabla 8.2 muestra que la tasa de aciertos en test de la PG es tan alta como en la mayoría de los algoritmos tradicionales de Aprendizaje Automático que se han probado. Además, las dos últimas columnas de la tabla presentan los resultados segregados en sensibilidad y especificidad. Si se interpretan estas medidas como la precisión por clase, la

primera para la clase positiva y la segunda para la negativa, se puede notar que casi todos los algoritmos consiguen predicciones aceptables en ambas clases. Las excepciones son AODE y MLP, que están sesgados hacia la clase negativa, provocando que las instancias de la clase positiva se predigan peor que la aleatoriedad. Este hecho se resuelve en la solución definitiva presentada en el capítulo 6, con 10 atributos adicionales y una distribución del 20 % de ejemplos positivos y un 80 % de negativos, alcanzando un 70,68 % de sensibilidad y un 83,80 % de especificidad.

8.1.4. Gestión de Valores Desconocidos y Simplificación de la Interpretación

Este apartado describe qué pasa cuando se añade un nuevo operador a los existentes: el operador *if_?*. Intenta gestionar el importante problema de los valores desconocidos debido al gran número de ellos en este dominio, como ya se ha comentado. También se analizan los efectos derivados de la aplicación del método Tarpeian para el control del crecimiento desmesurado.

Comparación Manteniendo Valores Desconocidos

Aquí se evalúan los dos enfoques diferentes comentados para gestionar los valores desconocidos. El primero los rellena con una marca numérica específica (configuración base). El segundo conserva los valores desconocidos en los datos, y cada algoritmo usa su propio criterio para procesarlos. Por ejemplo, la PG añade un nuevo operador (*if_?*), y los algoritmos de Weka los rellenan con la media o ignoran la instancia completa (ver la sección 8.1.1 para una explicación detallada de este aspecto).

Las columnas segunda y tercera de la tabla 8.2 muestran la tasa de aciertos en test correspondiente al primero y segundo enfoque descritos, respectivamente. Así, cuando se analiza la tasa de aciertos (segunda columna), el algoritmo PART es ligeramente mejor que PG. Sin embargo, al mirar la tercera columna, PG presenta el valor más alto de toda la columna. Esto significa que si se conservan los valores desconocidos en los conjuntos de datos, PG sobresale por encima de los algoritmos de Aprendizaje Automático.

Control del Crecimiento Desmesurado de la Solución

La tabla 8.3 y la figura 8.1 muestran cómo cambian varias medidas (tasa de aciertos en test, tamaño del árbol y tiempo de ejecución) para seis configuraciones diferentes. La configuración *Base* es la mejor encontrada, sin control de crecimiento desmesurado, cuyos parámetros se pueden consultar en la tabla 8.1. *Base sin límite* se refiere a la configuración *base* sin restringir el tamaño máximo del árbol. *if_?* es la configuración *base* pero añadiendo este nuevo operador (ver sección 8.1.2 para la descripción del operador *if_?*). Finalmente, la configuración *Tarpeian* incluye dicho método de control del crecimiento desmesurado y la característica sin límite. *if_? & sin límite* y *if_? & Tarpeian* combinan las configuraciones de ambos elementos.

En la tabla 8.3 se observa que la tasa de aciertos en test es casi constante en todas las configuraciones, alrededor de un 60,5 %; con un muy ligero descenso cuando se incluye el operador *if_?* o/y el método Tarpeian. Sin embargo, con respecto al tamaño del árbol (número de nodos) y el tiempo (ver figura 8.1), los valores para las configuraciones con el operador *if_?* o el método Tarpeian son considerablemente más bajos que los otros. Con el método Tarpeian la reducción es mayor que con el operador *if_?*, e incluso más cuando ambos se usan juntos. De

Tabla 8.3: Influencia del operador *if_?* y método Tarpeian: tasa de aciertos en test.

Id.	Configuración	Tasa de aciertos
a	Base	60,83 %
b	Base sin límite	60,93 %
c	*if_?*	60,67 %
d	*if_?* & sin límite	60,65 %
e	Tarpeian	60,43 %
f	*if_?* & Tarpeian	60,27 %

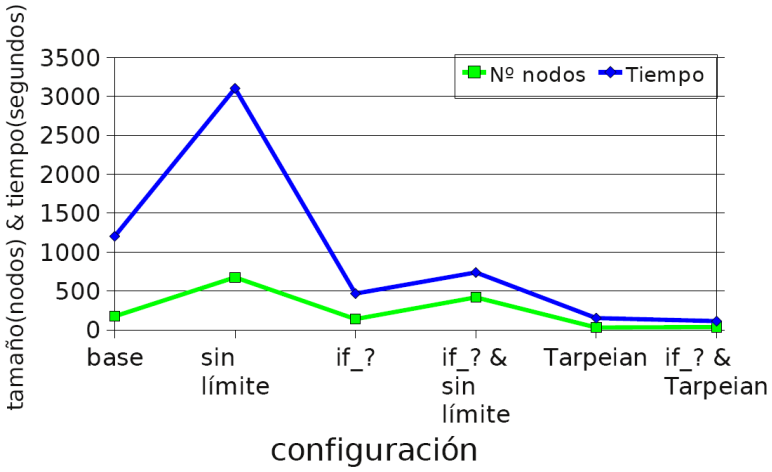


Figura 8.1: Influencia de operador *if_?* y método Tarpeian: tamaño del árbol y tiempo. El *eje y* cuantifica el tamaño (en número de nodos) y el tiempo (en segundos). La escala es la misma para ambas medidas.

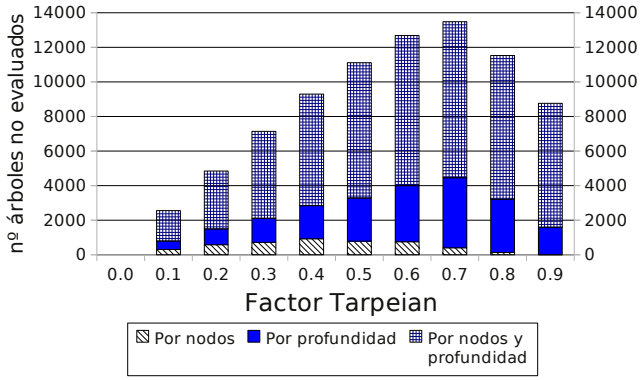
la configuración 'b' a la 'f', en media, el tamaño disminuye en más de 638 nodos y el tiempo en casi 3.000 segundos.

Además, cuando se aplica el operador *if_?* la longitud de la solución (es decir, el número de nodos del árbol) es bastante más corta que en el algoritmo PART, que es el mejor algoritmo de Aprendizaje Automático según la tasa de acierto en test (ver tabla 8.2). En la lista de decisión de PART hay 250 nodos (operandos y operadores) y en los árboles solución de PG con la configuración 'f' hay 38 nodos en media. En conclusión, el operador *if_?* y el método Tarpeian reducen el tamaño del árbol y el tiempo, disminuyendo escasamente la tasa de aciertos.

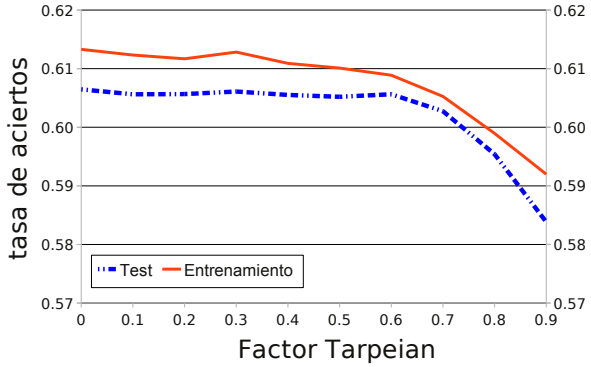
La figura 8.2(a) presenta la cantidad de árboles abortados (es decir, no evaluados) según el valor del factor Tarpeian. La configuración que se usa para generar este diagrama incluye también el operador *if_?*.

Por un lado, la figura 8.2(a) señala que el número de árboles no evaluados se incrementa con el factor Tarpeian, hasta alcanzar el valor 0,7. Cabe destacar que la mayoría de los árboles que no se evalúan, se debe a un exceso tanto en nodos como en profundidad media. Por otro lado, esta figura demuestra porqué usar el método Tarpeian implica una reducción en el tiempo de ejecución. La evaluación de los individuos es la parte de la evolución que exige más tiempo. Como el método Tarpeian aborta muchos árboles, no evaluando ninguno de ellos, entonces hace que se decremente significativamente el tiempo de ejecución.

En cuanto a la tasa de aciertos, la figura 8.2(b) indica que se mantiene casi constante para



(a) Nº árboles abortados (por diferentes criterios: nodos, profundidad y ambos).



(b) Tasa de aciertos.

Figura 8.2: Evolución al aplicar método Tarpeian con distintos factores: (a) nº de árboles abortados y (b) tasa de aciertos.

valores del factor Tarpeian del 0,0 al 0,6, y se produce un decremento considerable a partir del 0,7.

Del valor 0,7 a 0,8 del factor Tarpeian, los valores más altos alcanzados bajan mucho, tanto en cantidad de árboles abortados (ver que la columna 0,8 es la primera más baja que las anteriores en la figura 8.2(a)), como en tasa de aciertos (ver un descenso significativo del 0,7 al 0,8 en la figura 8.2(b), con menos del 60 %). 0,7 consigue un buen equilibrio entre la tasa de aciertos y la eficiencia, por lo que es el valor elegido.

Aunque la tasa de aciertos es ligeramente mejor en la configuración base con un factor Tarpeian de 0,0 (ver tabla 8.3), aplicar este método de control del crecimiento con un factor 0,7 aporta ventajas en cuanto al tamaño de la solución y el tiempo de ejecución en PG.

Resumiendo, disminuir el tamaño del árbol sólo implica una escasa pérdida en la tasa de aciertos con respecto a la configuración base. Sin embargo, los árboles obtenidos permiten una interpretación más sencilla y un proceso evolutivo mucho más rápido. Por lo tanto, se considera conveniente incluir en la propuesta de solución con PG ambos elementos, el operador *if_?* y el método Tarpeian.

Árbol de Salida

La salida de cada proceso evolutivo de un sistema de PG es un árbol, semejante al ejemplo mostrado en la figura 8.3. El ejemplo seleccionado procede de la configuración *if_? & Tarpeian*, porque es la que genera los árboles más pequeños. En concreto, el experimento que genera este árbol tiene una tasa de aciertos de un 60,44 %.

En este árbol se puede verificar que aparecen casi todos los terminales y que el operador *if_?* se usa con mucha frecuencia, como se analiza en el siguiente apartado. Si se interpreta el árbol, después de simplificar la expresión, se puede decir que: si el doble de *length_seq_min* es mayor que 0,59129 (primera línea) y *n_seqs_min* o *I2H* es desconocido (segunda línea), o también si el doble de dicho valor es menor que 0,59129 (primera línea) y *GF* o *length_seq_max* es desconocido (tercera línea), entonces la expresión comprueba si *GC* y *MT* son desconocidos también, retornando 0,74748 en ese caso; lo que significa una predicción de asociación funcional positiva para el par de proteínas dadas. Por el contrario, si ninguna de las condiciones anteriores se cumple, la predicción depende de las operaciones aritméticas específicas sobre las cantidades numéricas concretas.

Para concluir, a pesar del pequeño tamaño del árbol, la interpretación biológica es complicada. No obstante, cuando el tamaño es un poco mayor (lo que sucede en cuanto el operador *if_?* y el método Tarpeian no se aplican), ni siquiera se podría intentar interpretar el árbol.

```
ÁRBOL:                nodos: 20
 (if\_? (>= (+ length_seq_min length_seq_min) 0.59129
            (/ n_seqs_min I2H)
            (/ GF length_seq_max))
        (if\_? GC
               (if\_? MT
                      0.74748
                      n_seqs_min)
               0.59129)
        0.59129)
```

Figura 8.3: Árbol de uno de los mejores individuos usando operador *if_?* y método Tarpeian.

8.1.5. Relevancia de Operadores

En este apartado se determina cuál es la selección de operadores que realiza el proceso evolutivo, analizándolo en términos de frecuencia de aparición.

Una ventaja de la PG es que el proceso evolutivo selecciona automáticamente los terminales y operadores más relevantes para cada individuo. Por esta razón, se analiza la frecuencia con la que aparece cada uno en los árboles de salida. Los elementos más frecuentes deberían ser los más relevantes.

La figura 8.4 representa la distribución de los operadores en el conjunto de experimentos de la configuración *base* (a) y la configuración *if_? & Tarpeian* (b). La figura 8.4(a) muestra la misma proporción para cada operador. Por el contrario, cuando se añade el operador *if_?*, éste cubre la mayor proporción dentro de los operadores usados, con un 31 % cuando sólo se incluye el operador *if_?* y un 44 % si también se añade el método Tarpeian, cómo ilustra la figura 8.4(b). La razón que explique este hecho podría ser la elevada cantidad de valores desconocidos (*'?'*) en el conjunto de datos. Por lo tanto, parece importante considerar

estos valores desconocidos de forma diferente a los restantes numéricos, y aplicar cálculos especializados para los desconocidos, como los árboles solución de PG hacen frecuentemente. Además, conservar el valor *'?'* (en vez de reemplazarlo con una marca numérica) refleja mejor la situación, es más acorde con la semántica biológica real, como explica la sección 8.1.1. Se ha

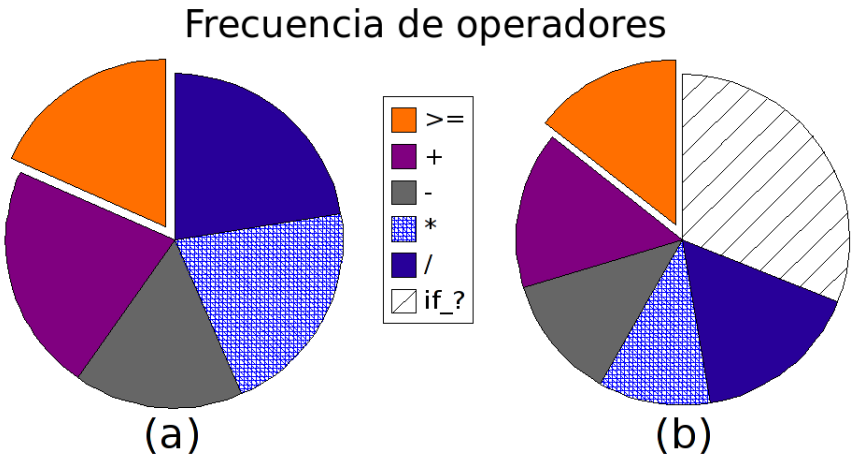


Figura 8.4: Frecuencia de operadores. (a) configuración base y (b) configuración *if_?* y Tarpeian.

aplicado la misma idea para la selección de terminales. Pero no se extrae ninguna conclusión relevante, porque todos los atributos aparecen en proporciones similares.

En general, cada característica tiene una relevancia similar, pero el operador *if_?* es el más importante, de acuerdo a las propiedades del dominio.

Conclusiones

Los resultados de la Programación Genética en términos de tasa de aciertos están en los mismos niveles que varios algoritmos de Aprendizaje Automático por defecto. Sin embargo, con la flexibilidad de la PG se consigue un manejo de los valores desconocidos o *no-existentes* mucho más cercano a la realidad biológica que con cualquiera de los otros algoritmos de AA. Además, se consigue limitar el tamaño de las soluciones (incluso más que en los algoritmos de AA), facilitando su lectura, con el uso del método Tarpeian, que también mejora la eficiencia, sin apenas pérdida en tasa de aciertos.

Los resultados se podrían mejorar más usando PG. Por ejemplo, con una función de evaluación del proceso de evolución más elaborada que la simple tasa de aciertos, dándole más peso a los aciertos positivos que son más importantes que los negativos en este dominio, lo cual se puede hacer fácilmente con PG. No obstante, en esta sección sólo se trata de demostrar la valía del uso de PG aplicado a este problema, y no experimentar con todas las configuraciones posibles. Por otro lado, se podría extender el conjuntos de terminales y funciones, usar ADFs [Koza, 1994], y otras mejoras del método Tarpeian [Mahler et al., 2005; Poli et al., 2007].

La PG es un enfoque válido e interesante para *aplicarlo a otros dominios* sobre todo cuando se necesite un diseño personalizado de los operadores. También se podría usar en otros problemas directamente el operador *if_?* aquí planteado, si se tiene la misma dificultad por la gran cantidad de valores desconocidos y representan algo diferente, porque la mayoría de algoritmos de AA no lo gestionan adecuadamente. Por otro lado, si se necesita una solución interpretable, el método Tarpeian resulta fundamental.

8.2. Aprendizaje Multi-clase. Aprendizaje Multi-etiqueta

Motivación-Hipótesis: ¿Se puede construir un modelo de predicción compuesto por un único árbol (en vez de uno por clase) que simplifique el aprendizaje y la interpretación de los resultados?

8.2.1. Enfoque

Una clasificación multi-clase, como contrapuesto a la binaria, significa que hay más de dos clases entre las que distribuir las instancias. Por su parte, una clasificación multi-etiqueta permite que a una instancia se le asocie más de una clase. El problema planteado en el capítulo 7 pertenece a ambas categorías, como suele suceder en la anotación funcional biológica. Porque las funciones moleculares o de procesos biológicos evidentemente son más de dos (multi-clase), y además un gen o producto genético no está involucrado sólo en una de ellas (multi-etiqueta).

Así, para afrontar un dominio multi-clase y/o multi-etiqueta se plantean diversas opciones, según la cardinalidad de las clases y de los valores de las mismas, y la descomposición en sub-problemas de aprendizaje que se realice.

Para resolver un aprendizaje **multi-clase** se pueden considerar dos soluciones básicas: a) dividir la salida del clasificador entre más de 2 valores, si el algoritmo de aprendizaje lo permite, como es el caso de los árboles de decisión; o b) descomponer el problema en varios clasificadores. Existen múltiples formas de descomposición en clasificadores (uno-contra-todos, uno-contra-uno, p-contra-q, etc.) [Ou and Murphey, 2007], pero en este caso sólo se considera la más sencilla y extendida (a pesar de sus limitaciones [Ou and Murphey, 2007]) para problemas con un número elevado de clases (como las 37 de la extensión de rutas): un predictor binario por clase siguiendo la estrategia uno-contra-todos.

Por otra parte, para un problema **multi-etiqueta**, las soluciones básicas que se consideran son: a) un clasificador con un vector de salida con tantas posiciones como etiquetas; o b) descomponer el problema en tantos clasificadores como etiquetas. Para la opción 'b', igual que para el problema multi-clase, se elige la opción de predictores binarios uno-contra-todos. Aunque la descomposición en binarios es igual para multi-clase y multi-etiqueta, a la hora de clasificar una nueva instancia, en el primer caso se verifican los clasificadores hasta encontrar la clase asignada; mientras que para el multi-etiqueta se deben miran las salidas de todos los clasificadores y recopilar el conjunto de clases con asignación positiva. Para un árbol de regresión (con un valor numérico en las hojas en lugar de uno nominal), en un multi-clase básico habría que seleccionar sólo la clase con valor de salida más alto, mientras que en un multi-etiqueta se establecería un umbral para cada clase, y se daría como salida el conjunto de clases que supere su umbral específico.

En la tabla 8.4 se recopila una combinatoria de posibilidades básicas, teniendo en cuenta la resolución multi-clase y multi-etiqueta comentada.

En el problema a resolver, la extensión de rutas, sólo son aplicables los casos `[N,M,?]` (ver las dos últimas filas de la tabla 8.4), porque son las opciones que resuelven los problemas multi-clase y multi-etiqueta simultáneamente, mediante combinaciones de los casos previos más simples. La principal ventaja del caso `[N,M,1]` es la sencillez tanto en aprendizaje, porque sólo se construye un clasificador, como en interpretación, ya que se predicen todas las clases a la vez en un único modelo con la salida compactada. Además, facilita un aprendizaje jerárquico (si lo hubiera) puesto que, cuando unas clases están relacionadas con otras, es más necesario realizar un aprendizaje conjunto. En el caso `[N,M,N]` puede que la predicción

Tabla 8.4: Combinatoria de nº de clases, etiquetas y aprendizajes. La columna *clases* indica cuántas clases distintas hay en el problema. La columna *etiquetas* indica cuántas clases se pueden asignar a una sola instancia. La columna *aprendizajes* indica cuántos procesos de aprendizaje diferentes se hacen, es decir, cuántos modelos de predicción se construyen.

Id.	**Clases**	**Etiquetas**	**Aprendizajes**	**Observaciones**
Valores	2, N (N=multi-clase)	1, M (M=multi-etiqueta) (M≤N)	1, N	
[2,1,1]	2	1	1	**Aprendizaje binario** *(clases=[pos, neg])*
[2,1,N]	2	1	N	No existe (*N* aprendizajes para sólo 2 clases binarias y etiqueta única).
[2,M,1]	2	M	1	**Aprendizaje multi-etiqueta único**, con salida en forma de vector de dos posiciones.
[2,M,N]	2	M	N	Aprendizaje multi-etiqueta, descompuesto en *M=2* binarios básicos [2,1,1], uno por etiqueta.
[N,1,1]	N	1	1	**Aprendizaje multi-clase único** (una etiqueta) *(clases=[a, b, c, ..., N])*
[N,1,N]	N	1	N	División del problema multi-clase en *N* binarios básicos [2,1,1], uno por clase.
[N,M,1]	N	M	1	**Aprendizaje multi-clase y multi-etiqueta único**, con salida en forma de vector de *M* posiciones.
[N,M,N]	N	M	N	División del problema multi-clase y multi-etiqueta en *M=N* binarios básicos.

proporcione unos resultados más fiables, por no mezclar clases e incluso permitir un análisis independiente por clases. No obstante, requiere un mayor procesamiento y complejidad en la construcción del modelo con N predictores diferentes, con el consecuente coste en tiempo de aprendizaje y también de interpretación, a través de un modelo de mucho mayor tamaño. Si se quisiera aplicar este enfoque a un aprendizaje jerárquico, habría que restringir manualmente las predicciones de las clases de nivel superior (más generales) según las predicciones positivas de las clases de nivel inferior (más específicas), porque los predictores que se generan por separado no se pueden considerar realmente independientes.

Desde otro punto de vista, cabe destacar que todas las combinaciones de aprendizaje de la tabla 8.4 se podrían afrontar fácilmente con una representación relacional, como la propuesta en el capítulo 5. Sólo sería necesario cambiar el objetivo de predicción, la asignación de clases a instancias y seleccionar el subconjunto de predicados lógicos correspondiente, pero el conocimiento del dominio se mantendría constante en contenido y representación distribuida en tablas. Incluso se puede usar un mismo algoritmo de aprendizaje, como los interpretables árboles de decisión, y una misma herramienta (TILDE o CLUS), siendo configurables para la mayoría de combinaciones; lo que también facilitaría una comparación de resultados. De forma opuesta, una representación proposicional requeriría más cambios, necesitando una reconstrucción de la tabla de datos según el tipo de aprendizaje.

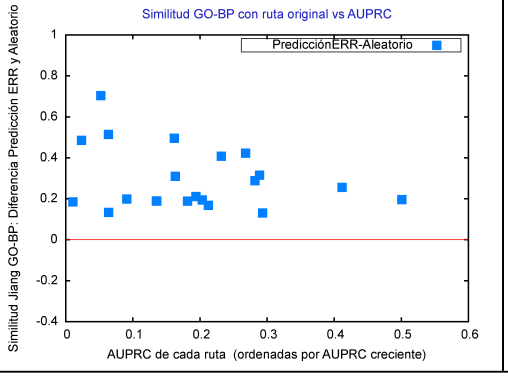
Referente a los multi-clasificadores, existe una posibilidad alternativa, que la herramienta CLUS también contempla. Se trata de generar un modelo que prediga a la vez *N* atributos diferentes. Cada uno de ellos correspondería a un ruta en el problema analizado. Pero dado

que se genera igualmente un único árbol, con la salida para cada uno de los atributos en cada hoja del árbol, el resultado es prácticamente equivalente a construir un multi-clasificador con un único atributo de salida que sea un vector.

8.2.2. Predicción con Multi-clasificador

Aunque en la solución al problema de extensión de rutas se utiliza el caso `[N,M,N]`, ya se ha introducido el uso de multi-clasificadores del tipo `[N,M,1]` para anotación funcional genómica en estudios previos [Vens et al., 2008], incluso con jerarquías funcionales, aunque sobre especies más simples que los humanos. Dichos estudios revelan que el uso de un multi-clasificador presenta un mayor rendimiento que el de un conjunto de clasificadores individuales. Por lo tanto, en esta sección se evalúa si el uso de un modelo único más simple mejora los resultados de los 37 clasificadores individuales construidos para el problema de extensión de rutas. Manteniendo la misma configuración que en el sistema ERR-PRyC, se obtienen los resultados que muestra la figura 8.5.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1559	0,6395	21 / 10	297 / 75,76 %



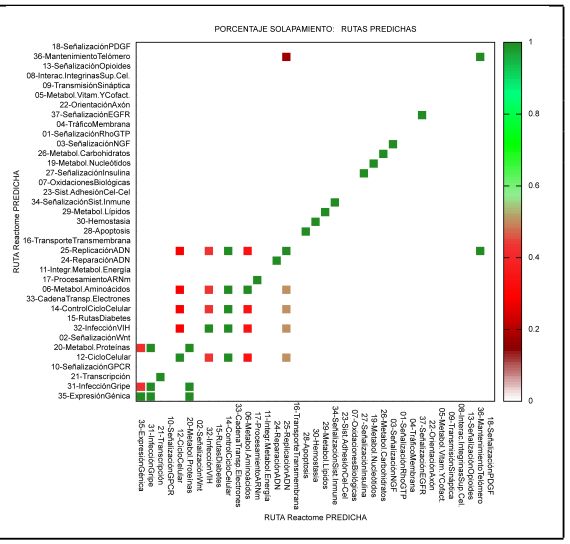


Figura 8.5: Resumen de resultados multi-clasificador.

Al comparar la figura 8.5 con la de referencia del sistema ERR-PRyC (figura C.1) se observa que el rendimiento en AUPRC es algo peor, aunque no baja hasta llegar al rendimiento del sistema ERR-PDR. No obstante, el AUROC es más bajo incluso que en ERR-PDR, lo cual indica que la predicción sobre instancias negativas es peor, es decir, se discrimina peor entre las instancias que no deben pertenecer a una clase. Tras una comprobación empírica, con varias configuraciones, el enfoque del multi-clasificador único ha alcanzado rendimientos superiores a la versión de clasificadores individuales, comparando ambos enfoques en las mismas condiciones (compartiendo dicha configuración entre la versión con multi-clasificador y la de clasificadores individuales).

Frente a ERR-PRyC se consiguen extender 3 rutas más. Pero es más relevante que se dobla el número de rutas predichas con más de una regla. No obstante, en ambos aspectos se sigue estando por debajo del sistema ERR-PDR.

Sin embargo, lo más destacado es que el multi-clasificador presenta mucho solapamiento entre rutas, incluso total entre algunas, como *Expresión génica* (35), *Infección por gripe*

(31) y *Metabolismo de proteínas* (20) (las dos primeras columnas y la sexta, de izquierda a derecha, en la parte derecha de la figura 8.5). Al compararlo con el solapamiento original de las rutas (ver figura 7.12(a)), se comprueba que el solapamiento entre estas 3 rutas existe originalmente, y el clasificador no lo resuelve en las nuevas predicciones. Pero también se observa un elevado solapamiento en otras 5 rutas con menos AUPRC que forman una especie de cuadrado concéntrico en la figura 8.5. Son *Ciclo celular* (12), *Infección por VIH* (32), *Puntos de control del ciclo celular* (14), *Metabolismo de aminoácidos* (6) y *Replicación del ADN* (25). La razón en este caso no es un solapamiento original de las rutas, sino que probablemente se predicen varias clases en un mismo nodo del árbol, con probabilidades semejantes, extendiendo con exactamente las mismas proteínas. Esta coincidencia en nodo sólo se puede producir en el enfoque de multi-clasificador único, lo que parece un inconveniente, de cara a la diversidad de predicciones entre distintas clases. Además, el total de proteínas añadidas son un 10 % menos diferentes entre sí, lo cual tampoco es favorable cuando el objetivo es una alta diversidad.

Conclusiones

Aprendiendo con un multi-clasificador único, el rendimiento en términos de AUPRC es peor que en el sistema que prioriza rendimiento y cobertura, pero mejor que en el que prioriza diversidad de reglas. Las predicciones sobre las instancias negativas son peores. Pero lo más relevante es el elevado solapamiento entre las predicciones de distintas rutas, lo cual no es admisible cuando se da preferencia a la variabilidad de proteínas añadidas.

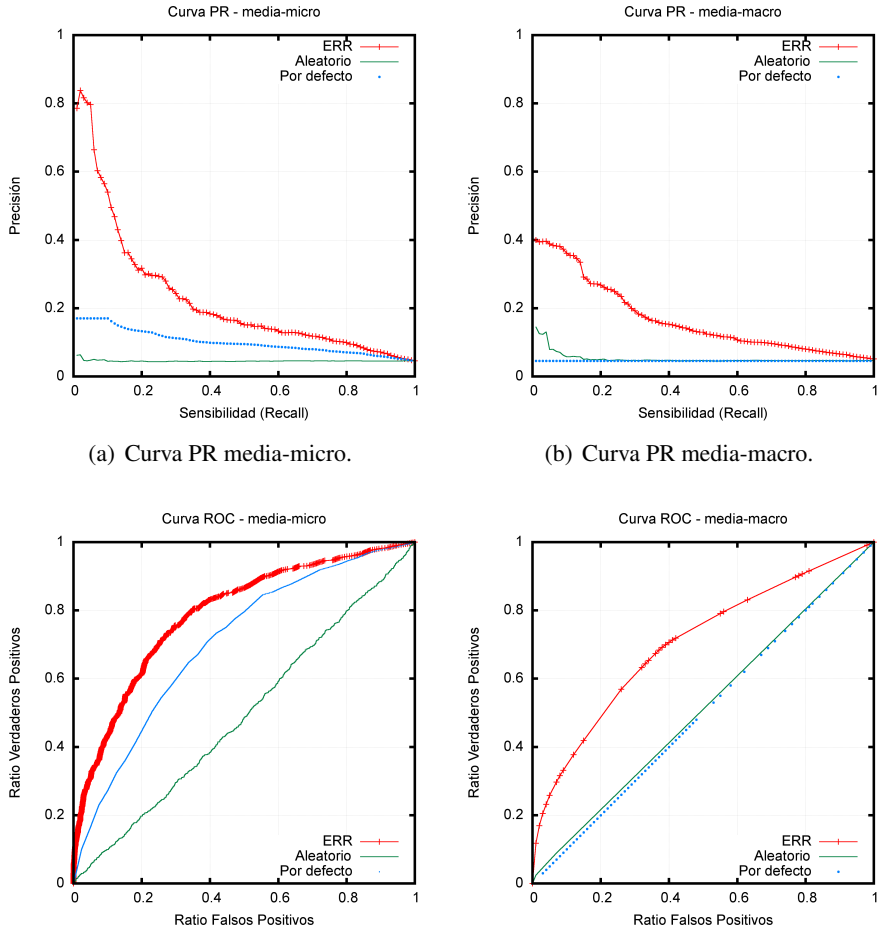
En la *aplicación a otros problemas*, si no importa que las predicciones solapen entre diferentes clases y sólo se está interesado en las predicciones positivas, construir un multi-clasificador es mejor en términos de coste computacional y de simplicidad de resultados, sin perder demasiado en rendimiento. Incluso existen trabajos que afirman que un multi-clasificador es la mejor solución, verificado sobre problemas concretos de predicción de anotación funcional de genes [Vens et al., 2008]. Ahora bien, depende del problema y del interés. Por ejemplo, si las clases están relacionadas en una jerarquía (como en el trabajo mencionado) probablemente no importe que solapen las predicciones, o incluso convenga en los casos de clases padre e hija. Por otro lado, si interesa una mayor cobertura de anotación que una elevada precisión (es decir, preferir una mayor cantidad de anotaciones frente a un conjunto más restringido y fiable) probablemente también sea más adecuado un multi-clasificador.

8.2.3. Influencia Evaluación Multi-clase

Para evaluar un problema de aprendizaje multi-clase con una medida global de rendimiento se debe elegir entre los dos métodos para promediar explicados en la sección 4.2.1: la micro-media y la macro-media. Es importante elegir la más conveniente acorde a las características del dominio para evaluar los aspectos que realmente interesan, y además poder alcanzar el objetivo buscado con el aprendizaje, sobre todo cuando se guía la búsqueda del mejor modelo con esta medida promedio, como es el caso del sistema ERR-PRyC.

En concreto, en el problema de extensión de rutas interesa hacer una predicción buena para todas las rutas, o el mayor número posible de ellas, frente a una predicción buena para el mayor número de instancias, aunque casi todas pertenezcan a un par de clases mayoritarias. Por lo tanto, se elige la media-macro que proporciona un valor medio por clase, no por instancia, sin sesgar el resultado hacia la clases más frecuentes, como hace la media-micro. Como se puede observar en las curvas de la figura 8.6 existe una diferencia notable entre la media-micro y la media-macro para el sistema ERR-PRyC, tanto en ROC como en PR, siendo la media-micro

más optimista que la media-macro. En realidad, lo que sucede es que la media-micro sesga el rendimiento global hacia el buen rendimiento de unas pocas clases mayoritarias, *ocultando* la información de rendimiento de las clases minoritarias, lo cual no interesa en este dominio.



(a) Curva PR media-micro.

(b) Curva PR media-macro.

(c) Curva ROC media-micro.

(d) Curva ROC media-macro.

Figura 8.6: Curva PR y ROC con macro-media y micro-media en ERR-PRyC.

Estas diferencias también se pueden verificar cuantitativamente, con las áreas bajo las curvas correspondientes que aparecen en la tabla 8.5. En el caso de la media-micro se calcula el área bajo la curva media (*AU(mediaPRC)*), mientras que en la media-macro se calcula la media de las áreas bajo la curva calculadas previamente de forma individual (*media(AUPRC)*).

Tabla 8.5: Áreas bajo la curva medias en ERR-PRyC. Distintos enfoques.

ERR-PRyC	Media-Micro	Media-Macro
AUPRC	0,2286	0,7893
AUROC	0,1695	0,7028

Para la *aplicación a otros problemas*, con características diferentes a la extensión de rutas, se debería usar una media-micro cuando lo que realmente interese sea el medir el mayor número

de aciertos, independientemente de la clase a la que pertenezcan. O también cuando no haya una distribución heterogénea de los ejemplos entre clases.

8.3. Extracción de Patrones Frecuentes

Motivación-Hipótesis: ¿Qué diferencias existen al extraer los patrones frecuentes de todas las clases a la vez o por separado? ¿Qué implicaciones tiene juntar los patrones frecuentes en cada clase o mantenerlos por separado, para usarlos como entrada de los clasificadores individuales?

Opciones Extracción Patrones

El modelo de predicción del sistema de ERR para expansión de rutas biológicas se descompone en dos fases, siendo la primera de ellas la extracción de patrones frecuentes. En esta sección se plantean 3 opciones posibles para esta fase, dependiendo de cómo se calculan los mismos.

1. **Opción 1** ***(opc1)***: Generar patrones frecuentes de todas las rutas a la vez, no limitándose a buscar frecuencia en cada ruta por separado. Esta opción permite usar a continuación un único clasificador o un clasificador para cada ruta.

2. **Opción 2** ***(opc2)***: Generar patrones frecuentes para cada ruta independientemente y luego juntarlos todos como entrada del clasificador. Esta opción también permite un clasificador único o uno por ruta.

3. **Opción 3** ***(opc3)***: Usar directamente para la clasificación los patrones frecuentes generados independientemente para cada ruta (extraídos como en la *opc2*), sin juntarlos con los del resto de rutas. Por lo que sólo se puede usar un árbol independiente para cada ruta.

Hasta ahora, en los trabajos de anotación funcional que se han realizado usando la combinación *'extracción de patrones frecuentes'* + *'uso de clasificador proposicional'*, sólo se ha contemplado la opción 1 [Clare et al., 2006; Vens et al., 2008].

En la opción 1, al existir muchas más instancias sobre las que calcular frecuencias frente a la opción 2, para que un mismo patrón se considere frecuente, el mínimo exigido debe ser menor. Entonces, para obtener un número suficiente de patrones, también se puede aumentar la cantidad de predicados a incluir en cada patrón (es decir, aumentar la profundidad de búsqueda de patrones). Así también se consiguen patrones base más complejos, lo que significa un aumento del número de atributos para la fase de clasificación posterior.

Por su parte, la opción 3 limita la entrada del árbol a los patrones frecuentes en esa clase particular, independientemente de los que son frecuentes en otra, o en todas. Así, se pueden seleccionar los mejores subconjuntos de patrones, que describan las propiedades relevantes de las proteínas de esa ruta. Esta opción es equivalente a usar directamente los patrones frecuentes en una ruta. Pero de forma más eficiente, porque el árbol selecciona los patrones que se tienen que verificar y cómo se tienen que combinar, pudiendo obtener varias reglas, con más probabilidad de ser diferentes al resto, y asociándoles una medida de certeza. Por el contrario, si se exigiera que se cumpla el conjunto completo de patrones, sólo se obtendría una regla (conjunción de todos los patrones frecuentes en esa ruta).

Adicionalmente existiría la posibilidad de limitar el modelo de aprendizaje al cálculo de los patrones frecuentes por clases independientes, sin incluir ningún árbol de decisión posterior. Pero en este caso habría que establecer un criterio para decidir qué patrones es obligatorio verificar para la predicción: todos los patrones, limitar a los N más frecuentes o a los que tengan una frecuencia superior a un umbral, limitar a patrones con alta frecuencia en esa clase y baja en las otras, limitar a patrones seleccionados por un árbol de decisión común (es decir, construido con un multi-clasificador), etc. No se trata de una decisión trivial, porque se debe prestar atención a no añadir redundancia en el procesamiento de los datos, realizando tareas que ya haga automáticamente algún algoritmo de aprendizaje automático utilizado. Por ejemplo, el criterio de selección de patrones a utilizar ¿no lo hace automáticamente el árbol de decisión? Por otro lado, hay restricciones como obtener más de una regla (porque con un subconjunto de patrones sólo se tendría una), y una medida de probabilidad de la bondad de la predicción, que el clasificador proporciona automáticamente. Por lo tanto, esto verifica que la segunda fase (árbol de decisión, CLUS) del método de aprendizaje no es redundante con la primera (extracción de patrones, WARMR), aunque se extraigan los patrones separados por clase, porque no se sabe cuáles elegir para componer cada regla.

Combinación con Aprendizaje Multi-clase

Si se combinan las tres opciones de extracción de patrones con el uso de un multi-clasificador o un clasificador individual por clase (analizado en la sección 8.2), resultan 5 métodos de aprendizaje alternativos, resumidos en la tabla 8.6. 3 son métodos con clasificadores individuales por clase (CI) y 2 con multi-clasificador (MC).

Tabla 8.6: Combinaciones Extracción Patrones Frecuentes y Aprendizaje Multi-clase.

Extracción Patrones	**Multi-Clasificador** MC	**Clasificadores Individuales** CI
Opción 1 en todas las clases a la vez	*opc1-MC*	*opc1-CI*
Opción 2 en cada clase por separado, **con** unión patrones	*opc2-MC*	*opc2-CI*
Opción 3 en cada clase por separado, **sin** unión patrones	—	*opc3*

A continuación se realiza una comparativa de resultados de los 5 métodos alternativos de aprendizaje aplicados al problema de extensión de rutas, con la misma configuración que el sistema ERR-PRyC (opc2-CI), excepto en lo que se refiere a extracción de patrones y enfoque multi-clase. Se utilizan las mismas medidas de comparación con el sistema ERR-PRyC definidas al inicio de este capítulo, pero distribuidas entre medidas cuantitativas (ver Tabla 8.7), solapamiento (ver Figura 8.7) y similitud semántica (ver Figura 8.8).

En primer lugar, desde una perspectiva lógica, la opción 2 debería ser la mejor, porque incluye información de otras clases para discriminar con la actual. Este resultado se confirma en términos de rendimiento en la tabla 8.7, dado que la opc2-CI se corresponde con el sistema

ERR-PRyC.

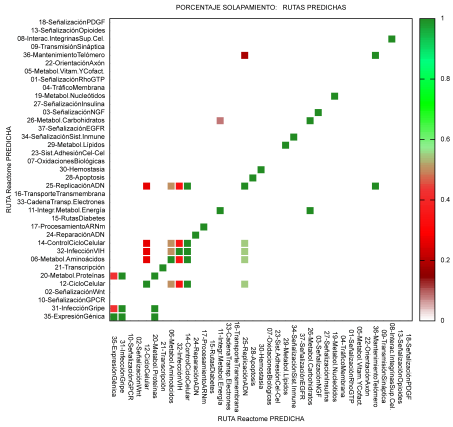
La tabla 8.7 muestra que el **rendimiento** en términos de AUPRC es más bajo en los sistemas MC que CI, sobre todo en la *opc2-MC* y en ERR-PDR. Estos dos sistemas son también los que consiguen mayor número de rutas extendidas con más de 1 regla, de lo que se concluye que se necesita ceder en rendimiento para obtener una mayor diversidad en las predicciones. Cabe destacar que en los sistemas MC presentan valores de AUROC bastante más bajos que los CI, que indica que estos clasificadores cometen más errores en los ejemplos negativos, es decir, en las proteínas que no pertenecen a las rutas.

Tabla 8.7: Evaluación numérica de la extensión de Reactome por 5 variantes según extracción patrones frecuentes y enfoque aprendizaje multi-clase.

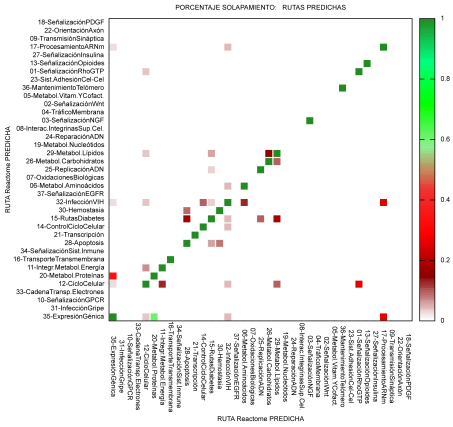
Sistema	**AUPRC**	**AUROC**	**nºrutas** (total/>1regla)	**nºproteínas añadidas** (total/ %dif.)
opc1-MC	0,1620	0,6488	21 / 9	293 / 74,74 %
opc1-CI	0,1676	0,7111	20 / 6	277 / 88,09 %
opc2-MC	0,1559	0,6395	21 / 10	297 / 75,76 %
opc2-CI (ERR-PRyC)	0,1695	0,7028	18 / 5	249 / 87,55 %
opc3-CI	0,1677	0,7073	18 / 5	259 / 88,03 %
ERR-PDR	0,1337	0,6914	28 / 15	383 / 85,90 %

Lo más relevante al observar la figura 8.7 es el mayor **solapamiento** entre las predicciones para distintas clases en los sistemas MC que en los CI. Este solapamiento se produce principalmente en las clases o rutas bien predichas (situadas en la esquina inferior izquierda de cada gráfico), siendo además de alta intensidad (observar coloración en tonos superiores de la escala derecha, como verdes). La última columna de la tabla 8.7 también confirma esta situación, con un porcentaje de proteínas añadidas diferentes bastante menor en los sistemas MC que en los CI.

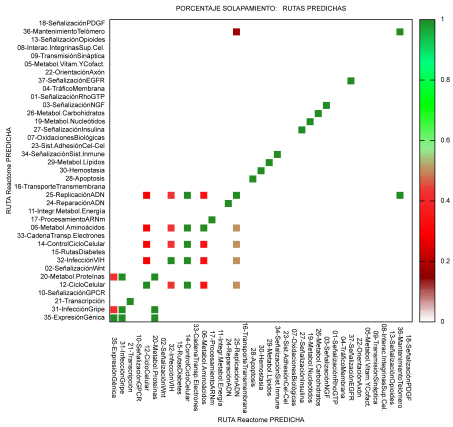
Al analizar la **similitud semántica** funcional respecto a las rutas originales de las cinco combinaciones propuestas no se observan diferencias significativas entre unos y otros en la figura 8.8. No obstante, se puede apreciar que en la opción 2 todas las rutas extendidas tienen su punto por encima de la línea, lo que representa que en absolutamente todas las predicciones del sistema son mejores que una extensión aleatoria. En el resto de sistemas, alguna ruta se extiende peor que la aleatoriedad (punto por debajo de la línea) o está más cerca de ello. Pero basándose en la configuración del sistema ERR-PRyC (con baja frecuencia y alto nivel de profundidad al extraer patrones frecuentes) cualquier sistema presenta una buena similitud semántica frente a las rutas originales. Hay que mencionar que durante la experimentación realizada sí se han encontrado algunos sistemas con perfiles de similitud semántica más cercanos a la aleatoriedad.



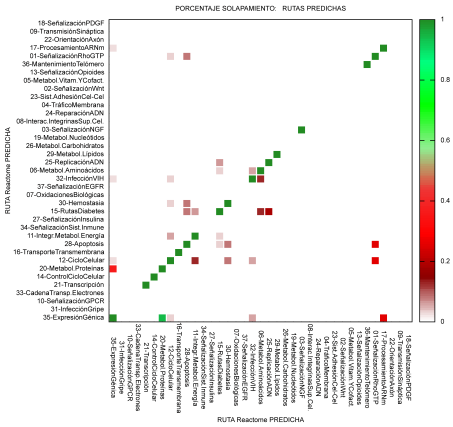
(a) Opción 1 - MC.



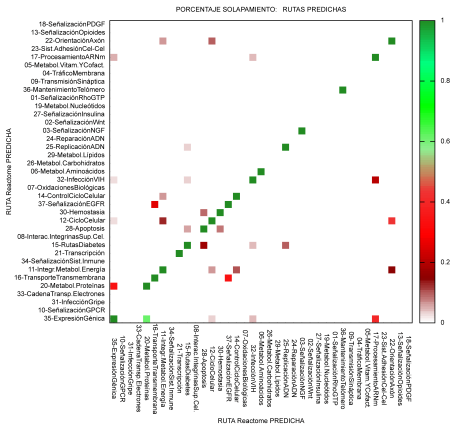
(b) Opción 1 - CI.



(c) Opción 2 - MC.

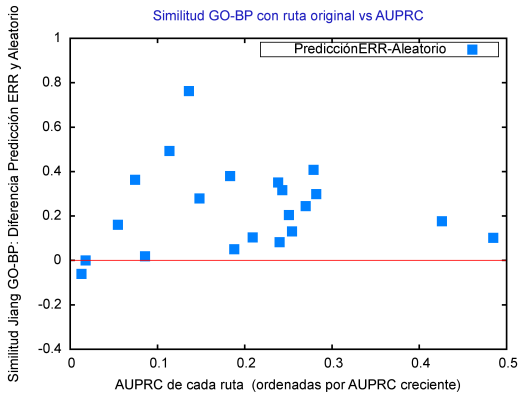


(d) Opción 2 - CI (sistema ERR-PRyC).

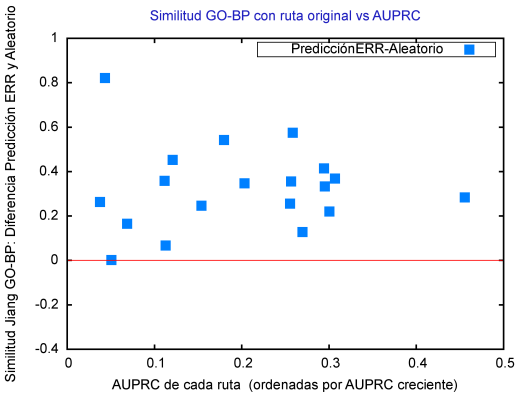


(e) Opción 3 - CI.

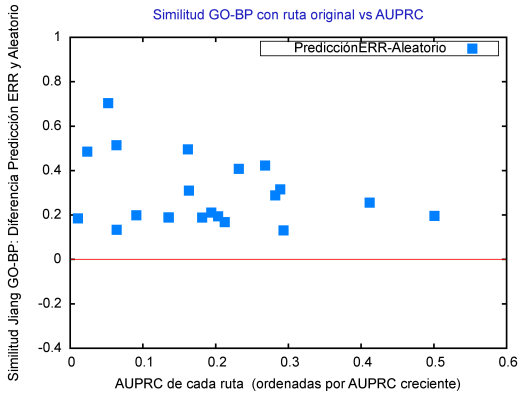
Figura 8.7: Porcentaje de solapamiento entre rutas. Comparación 5 variantes según extracción patrones frecuentes y enfoque aprendizaje multi-clase. Las rutas están ordenadas por AUPRC creciente según cada sistema, de izquierda a derecha en el *eje x* y de abajo a arriba en el *eje y*.



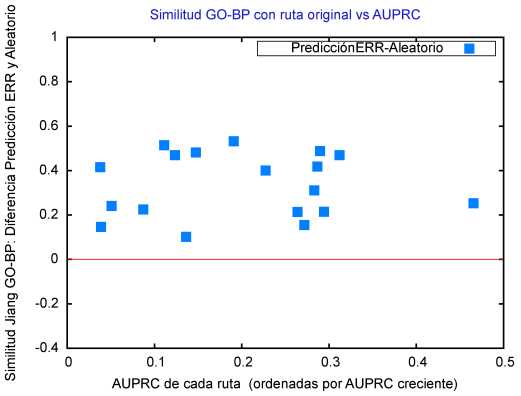
(a) Opción 1 - MC.



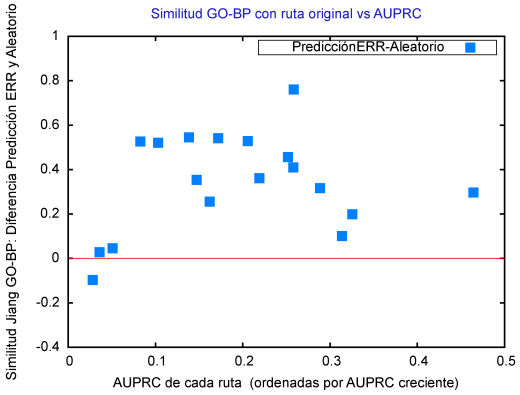
(b) Opción 1 - CI.



(c) Opción 2 - MC.



(d) Opción 2 - CI (sistema ERR-PRyC).



(e) Opción 3 - CI.

Figura 8.8: Similitud de anotación funcional entre proteínas de la ruta original y proteínas añadidas (por predicción y aleatoriamente). Comparación 5 variantes según extracción patrones frecuentes y enfoque aprendizaje multi-clase. Cada punto representa la diferencia de similitud a la ruta original entre las proteínas predichas y las proteínas aleatorias (*Sim.Predichas–Sim.Aleatorias*) para esa ruta. Así, la línea roja representa la inexistencia de mejora de las predicciones frente a la aleatoriedad, en términos de similitud. Las rutas están ordenadas por AUPRC creciente en el grupo de predicciones. Las rutas sin extensión no se representan.

Conclusiones

Del análisis de las cinco combinaciones se extraen las siguientes conclusiones:

- Las opciones 1 y 2 de extracción de patrones combinadas con un multi-clasificador (opc1-MC y opc2-MC) son prácticamente iguales en todas las evaluaciones. Tras una comprobación empírica, se puede decir que se parecen más conforme se disminuye la frecuencia y se aumenta la profundidad en la extracción de patrones frecuentes, porque estos tienden a ser los mismos.

- El rendimiento (AUPRC) es mayor al usar los mismos patrones para clasificar sobre todas las rutas (opc1 y opc2) que al dividirlos por ruta (opc3), pero esta última opción tiende a presentar un menor solapamiento y la mayor diversidad en las predicciones.

- Los métodos con clasificadores individuales permiten aumentar el número de reglas por clase y su diversidad, lo cual es de notable interés en el problema de extensión de rutas tratado.

- La lógica indica que al extraer los patrones separados por rutas (opc2 y opc3) la frecuencia mínima se debería mantener alta, para obtener patrones frecuentes sólo en cada ruta o clase particular. No obstante, parece que al bajar la frecuencia mejoran los resultados en ambos casos, aunque los patrones extraídos dejen de ser exclusivos por clase.

- El sistema ERR-PDR sigue la combinación opc3-CI (extrae los patrones por rutas separadas y mantiene dicha separación en los clasificadores individuales por ruta) con una frecuencia mínima de 0,2. La misma combinación opc3-CI con otra frecuencia mínima más baja (0,005) permite que el rendimiento sea más alto, pero la diversidad en reglas cae hasta el valor más bajo del resto de opciones.

En la *aplicación a otros problemas* para determinar qué combinación de las cinco presentadas es más adecuada, se debe tener en cuenta la cantidad de ejemplos, predicados y niveles sucesivos de relaciones que se tengan, porque influye directamente en el coste computacional de extracción de patrones por separado para cada clase, y de entrenar clasificadores independientes. En una evaluación empírica, con menos patrones extraídos (por mayor frecuencia mínima y menos profundidad en la extracción de patrones), los sistemas con multi-clasificador presentan un mayor rendimiento (en AUPRC) que los de clasificadores individuales, y más aún sobre la opción 3. Por lo que, si no se tiene mucho poder de cómputo o la cantidad de ejemplos o predicados es muy elevada, se extraerían menos patrones, donde un sistema con multi-clasificador es la mejor opción. Aunque esta elección sólo sería válida si no importa el solapamiento de predicciones entre diferentes clases, que en los multi-clasificadores se mantiene para cualquier cantidad de patrones extraídos.

Si simplemente se busca alcanzar el mayor rendimiento, independientemente del coste, es mejor la opc2-CI. Pero sobre todo extraer muchos patrones, con una frecuencia mínima muy baja y un nivel de profundidad tan alto como se pueda en la extracción de patrones. Si por el contrario, se quiere la mayor diversidad posible en las predicciones, la elección es la opc3-CI con una frecuencia mínima no muy baja, como en el sistema ERR-PDR.

Finalmente, hacer notar que no hay ningún criterio fijo que permita determinar cuál va a ser el mejor sistema para todas las medidas de evaluación.

8.4. Variación de la Representación del Conocimiento

En el problema de extensión de rutas se ha seleccionado una representación del conocimiento híbrida, partiendo de una representación relacional que posteriormente se transforma a proposicional mediante la extracción de patrones frecuentes (ver secciones 2.2.4, 7.2.4 y 7.2.5). En esta sección se analiza la influencia en el aprendizaje de otras representaciones del conocimiento posibles.

8.4.1. Representación Relacional Directa

Motivación-Hipótesis: ¿Se puede aprender con una representación relacional directa, sin necesidad de una transformación a representación proposicional intermedia? ¿Cuáles son las diferencias con el sistema de representación híbrido usado para extender rutas? ¿En qué condiciones conviene utilizar una representación relacional directa?

Con la denominación *representación relacional directa* se asume la aplicación de algoritmos de Aprendizaje Automático Relacional sobre la información distribuida en varias tablas.

Para resolver las preguntas planteadas en esta sección, se aplica un clasificador relacional para el problema de predecir la pertenencia a una ruta, y se comparan los resultados. Entre las diversas herramientas disponibles que realizan AAR (ver sección 2.2.5) se selecciona TILDE porque induce árboles de decisión, que es el mismo modelo utilizado para el problema de extensión de rutas con el que se compara.

Aunque la salida de ambos enfoques es un árbol de decisión relacional, con predicados lógicos en los nodos, la representación híbrida utilizada en el capítulo 7 permite la evaluación de una conjunción compleja de literales, mientras que la representación relacional directa sólo permite la existencia de un literal por nodo. Por otro lado, el proceso de aprendizaje con una representación relacional directa es más sencillo, construyendo en un solo paso el árbol de decisión. Se utiliza exactamente la misma representación (ver figura 7.2) y ficheros de datos (ver sección 7.2.3) que en el capítulo 7, desde donde se extraen los patrones frecuentes con el algoritmo WARMR (ver sección 7.2.5), que también es de naturaleza relacional.

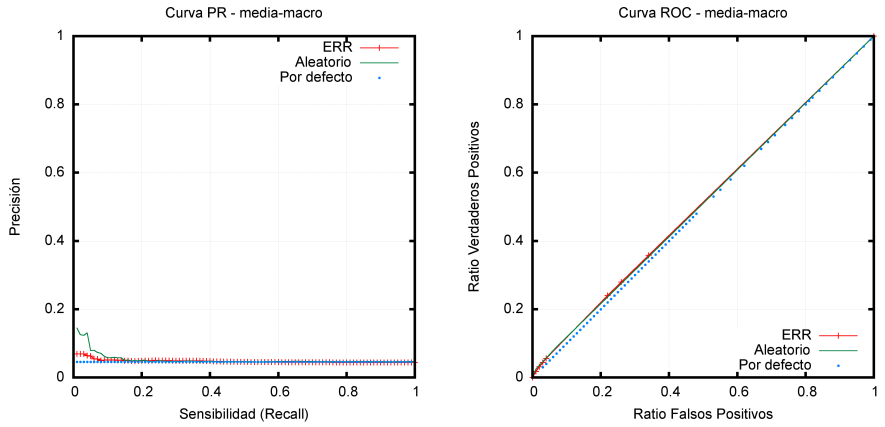
Los resultados de clasificación que arroja el algoritmo TILDE se evalúan mediante las correspondientes curvas media PR (figura 8.9(a)) y ROC (figura 8.9(b)).

Se observa que el rendimiento del uso de una representación relacional directa es equivalente al de una predicción aleatoria. Incluso, si se observan las curvas con media-micro (ver figura 8.10), según estudios previos [Vens et al., 2008], se podría decir que el sistema está sobre-entrenado, al ser peor que la clasificación por defecto.

Conclusiones

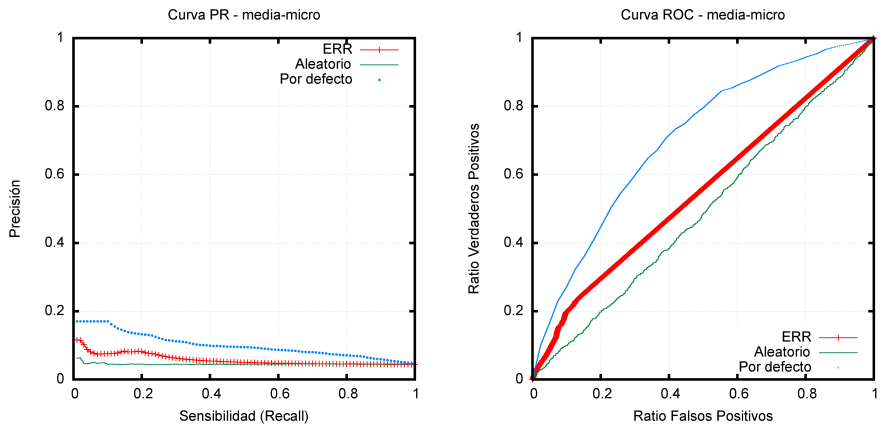
No se consigue aprender, quedándose la representación relacional directa en una clasificación aleatoria.

Tras diversas configuraciones diferentes a las del sistema de referencia del capítulo 7, se concluye que el inconveniente de la representación relacional directa radica en la evaluación de un solo predicado por nodo del árbol de decisión. Ya que en la mayoría de los casos un predicado independiente no es capaz de discriminar por si solo, sino que necesita combinarse con otros relacionados en un mismo paso de clasificación, mediante patrones frecuentes o agregados, entre otras opciones.



(a) Curvas PR. (b) Curvas ROC.

Figura 8.9: Curvas media-macro representación relacional directa: (a) curvas PR y (b) curvas ROC.



(a) Curvas PR. (b) Curvas ROC.

Figura 8.10: Curvas media-micro representación relacional directa: (a) curvas PR y (b) curvas ROC.

Respecto a la *aplicación a otros problemas*, en un dominio con menos relaciones básicas entre las instancias, donde no fueran necesarias tantas conexiones entre tablas para obtener una información que discrimine, este sencillo enfoque de aprendizaje relacional directo podría funcionar; quizá con la ayuda de un evento que evalúe un par de predicados en el mismo nodo (como permiten las técnicas denominadas en inglés *lookahead* [Blockeel and Raedt, 1998]). Pero, el uso de 3 tablas y 2 relaciones entre ellas para acceder, por ejemplo, a la información del gen correspondiente a una proteína (`protein-protein_gene-gene`) es demasiado complejo para que TILDE llegue a evaluarlo y detectarlo como diferenciador para la clasificación.

8.4.2. Representación Proposicional Directa

Motivación-Hipótesis: ¿El AA Relacional aporta ventajas frente al clásico proposicional, al permitir incluir relaciones complejas sin pérdida de semántica?

Con el término *representación proposicional directa* se quiere denominar a la representación atributo-valor más sencilla, a la que se llega sin ninguna transformación, donde sólo se representan los atributos numéricos o nominales que pueden estar contenidos en una única tabla (ver sección 2.2.4). Al aplicar esta representación al problema de extensión de rutas, se simplifica notablemente el lenguaje de representación de los datos, al definirse una columna de la tabla por cada característica numérica o nominal, quedando la siguiente lista de 11 atributos: *identificador de proteína, longitud de proteína, carga positiva, carga negativa, identificador de gen, longitud de gen, orientación del gen dentro del cromosoma, nº de transcritos, dominio transmembrana, dominio hélice y dominio de señal*. El identificador de gen, al igual que el de proteína, es un atributo nominal que funciona como 'clave' de cada ejemplo, obviándose durante el aprendizaje. Los tres últimos atributos, relativos a la presencia de uno o más dominios de cada tipo en la secuencia de proteína, son binarios. El resto toma valores numéricos.

Por otro lado, se ignora toda la información de interacciones porque implica relaciones con otras proteínas, lo que no se puede representar de forma directa. La aproximación más directa para incluir las interacciones consistiría en añadir 22.303 (*número de isoformas humanas principales – 1*) atributos, con uno por cada posible proteína diferente interaccionando. No parece una opción razonable porque muchos de estos atributos estarían vacíos la mayoría de las veces, ya que una proteína generalmente no tiene miles de interacciones, aunque la cantidad de atributos deba mantenerse constante para todos los ejemplos. Además, cada uno podría tomar 72.731 valores diferentes (número de isoformas humanas, con sus diferentes expresiones para cada gen) que implica una variabilidad excesiva, dificultando notablemente la localización de regularidades que discriminen en la clasificación. Esta inclusión de miles de atributos con miles de posibles valores también incrementaría el coste computacional, pudiendo llegar con relativa facilidad a una insuficiencia de memoria. Aparte de los inconvenientes técnicos, esta solución también implicaría una pérdida de semántica, puesto que las propiedades de las relaciones y de los elementos relacionados no quedarían representadas. Como por ejemplo, la carga de una proteína compañera de interacción o si dicha secuencia con la que se interacciona tiene algún dominio transmembrana, que son aspectos influyentes para determinar la función de una nueva proteína relacionada con otras. Parte de esta información incluida en las relaciones, se podría añadir a una representación proposicional de forma implícita utilizando agregados, como la suma de proteínas con las que se interacciona. Pero el objetivo de esta sección es evaluar los resultados que se pueden obtener con una *representación proposicional directa*, sin transformaciones intermedias a partir de una representación relacional, como la que ya se hace en la representación híbrida utilizada en el capítulo 7, con la que se compara a continuación.

Referente al modelo de clasificación, la representación proposicional genera un árbol de decisión *no* relacional, es decir, sin conjunciones lógicas en los nodos como en el caso relacional, sino con comparaciones numéricas de atributos concretos. Sólo se evalúa el aprendizaje proposicional con árboles de decisión, porque es la misma técnica que se usa en el capítulo 7 con cuyos resultados se quiere comparar, y además interesa la interpretación de los resultados. Para facilitar y hacer fiable la comparación se emplea CLUS [Blockeel et al., 1998], el mismo algoritmo y configuración de inducción de árboles que se usa en el capítulo 7 (tras la transformación a representación proposicional con patrones frecuentes), generando un árbol individual por ruta biológica.

Observando los resultados de la figura 8.11, los valores de rendimiento (en AUPRC y AUROC) están en torno a los resultados del sistema ERR-PDR, y son peores que en ERR-PRyC, ambos basados en la representación híbrida. Pero lo más destacado es que sólo es capaz de extender 3 rutas de Reactome, con ninguna variabilidad (ninguna ruta se predice con más

de 1 regla), y además con baja similitud semántica, lo que se sitúa muy lejos de los resultados de predicción alcanzados por los sistemas de representación híbrida.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1386	0,6775	3 / 0	42 / 100,00 %

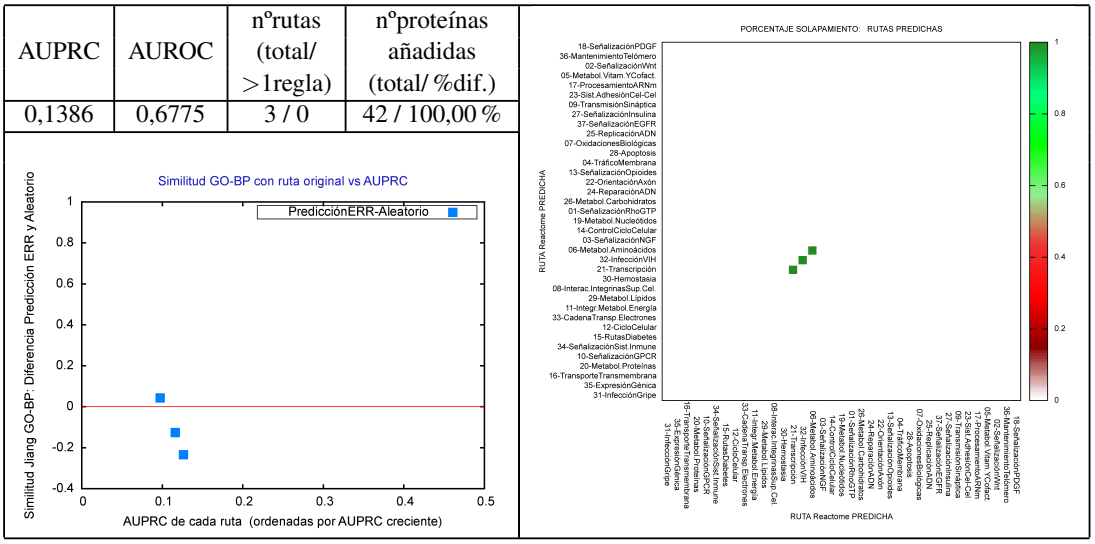


Figura 8.11: Resumen de resultados representación proposicional directa.

Además, la interpretación de los resultados también es más compleja con la representación proposicional, porque los árboles presentan muchas más comparaciones numéricas con cualquier umbral, y muchas sobre el mismo atributo, como muestra la figura 8.12.

```
geneLength > 122131.0 (10%)
+--yes: negCharge > 0.136499 (24.3%)
|       +--yes: geneLength > 134587.0 (92.6%)
|       |       +--yes: geneLength > 558075.0 (12%)
|       |       |       +--yes: [0.50519]: 3
...
```

Figura 8.12: Fragmento de árbol con representación proposicional directa.

Por otro lado, cabe destacar que con la representación relacional una regla que extiende una ruta en el sistema ERR-PDR contiene el fragmento:

```
complex_interaction(A,B),not(B=A),transmembrane_domain(B)
```

Es decir, que realmente, al contrario que en una representación proposicional, en una relacional se accede a las propiedades de las secuencias con las que se interacciona y no sólo de la `proteína A` sobre la que se predice. Incluso podría aparecer el literal `transmembrane_domain(A)` en otro patrón de la misma regla, indicando que ambas proteínas que interaccionan en un complejo contienen un dominio transmembrana.

Conclusiones

El rendimiento del AA proposicional puro es peor que en el sistema ERR-PRyC, que usa el AA relacional con una representación híbrida para la extensión de rutas de Reactome. Si se

compara con el sistema ERR-PDR, con el mismo enfoque de aprendizaje que ERR-PRyC, pero priorizando la diversidad, el rendimiento del AA proposicional puro es prácticamente igual. Sin embargo, la cantidad de predicciones es muy baja y sin ninguna variedad, y el modelo de clasificación menos interpretable, a pesar de la sencillez en la representación y del proceso de aprendizaje proposicional.

Referente a la *aplicación a otros problemas*, si lo único importante es el rendimiento y la cobertura, y no se tiene restringida la cantidad de instancias que se pueden predecir en cada clase, se puede simplificar el problema al uso de una representación proposicional. Pero si existen restricciones adicionales como en la extensión de rutas de Reactome, se necesita una representación del conocimiento más rica. Por otro lado, si existiera mucha información relacional, se perdería conocimiento por no poder representar dicho conocimiento, o la tabla de datos contendría información redundante. Por ejemplo, si se considerara un ejemplo diferente cada uno de los transcritos procedentes de un mismo gen, con la representación proposicional directa, se tendría información del gen duplicada en varias filas.

8.5. Influencia de la Información Relacional en la Predicción de Función

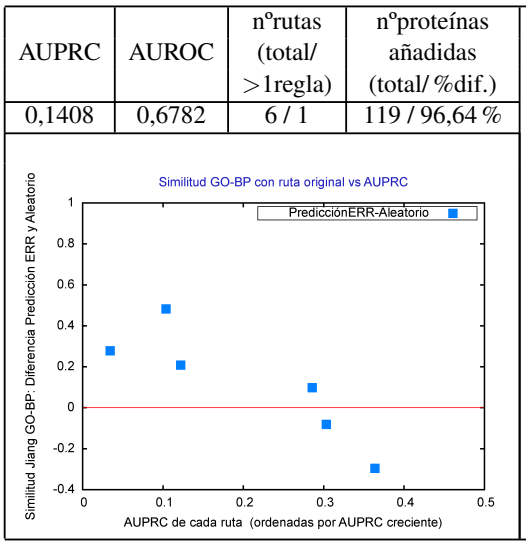
Motivación-Hipótesis: ¿Cuánto influye la información relacional (básicamente interacciones, con su información asociada) en la predicción de la anotación funcional?

8.5.1. Predicción sin Interacciones

En esta sección se analizan los resultados de extensión de rutas de Reactome sin incluir en el aprendizaje las interacciones proteína-proteína (IPP) y/o interacciones en complejos.

La figura 8.13 muestra el resumen de resultados tras eliminar ambos tipos de interacciones del aprendizaje.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1408	0,6782	6 / 1	119 / 96,64 %



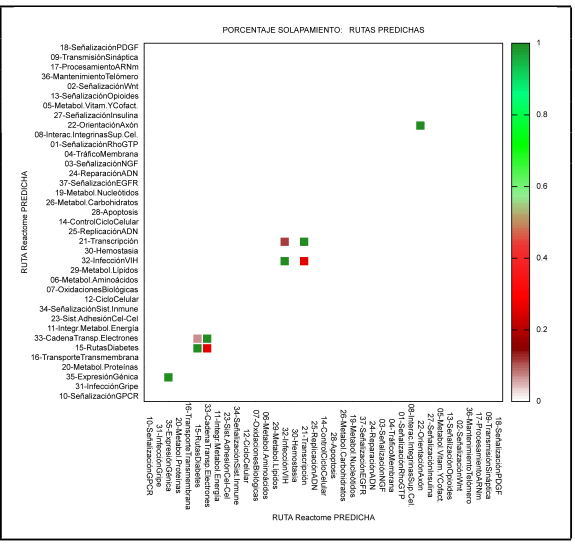


Figura 8.13: Resumen de resultados sin interacciones PP ni complejos.

En la figura 8.13 se observa que sin interacciones PP ni de complejos el rendimiento baja

casi 3 puntos, de los 17 alcanzados (sobre 100) por el sistema ERR-PRyC que sí incluye ambos tipos de interacciones. También, consigue extender pocas rutas (sólo 6), dentro de las cuales además existe solapamiento entre dos pares (ver gráfico derecho de la figura 8.13), y la similitud semántica con la ruta original de las dos rutas mejor predichas es peor que la de una extensión aleatoria (ver gráfico de la izquierda en figura 8.13).

La tabla 8.8 compara los resultados de eliminación de distinto tipo de información relacional por separado, en este caso, distinto tipo de interacciones. También se incluye una comparación con los resultados de la Representación Proposicional Directa (ver sección 8.4.2), que también representa los datos sin ninguna información relacional. El resumen de resultados completo para el aprendizaje sin interacciones PP o sin interacciones de complejos, se encuentra en la figura H.2 y H.1, respectivamente.

Tabla 8.8: Evaluación numérica de la extensión de Reactome sin interacciones.

Sistema	**AUPRC**	**AUROC**	**nºrutas** (total/>1regla)	**nºproteínas añadidas** (total/ %dif.)
ERR-PRyC	0,1695	0,7028	18 / 5	249 / 87,55 %
sin complejos	0,1559	0,7063	16 / 6	254 / 86,22 %
sin IPP	0,1331	0,6635	11 / 2	162 / 83,33 %
sin IPP ni complejos	0,1408	0,6782	6 / 1	119 / 96,64 %
repr.proposicional	0,1386	0,6775	3 / 0	42 / 100,00 %

Como era esperable, si se elimina cualquiera de los tipos de interacción los resultados empeoran, tanto en AUPRC como en rutas extendidas. Al aprender sin las interacciones de complejos (segunda fila de la tabla 8.8) la predicción empeora mucho menos que al aprender sin las interacciones PP (tercera fila), por lo que estas últimas parecen más importantes. De hecho, lo llamativo es que al eliminar sólo las interacciones PP se pierde mucho más que si se eliminan los dos tipos de interacciones (cuarta fila). Este hecho puede ser debido a que al quitar las interacciones de cualquier tipo del aprendizaje, haya alguna combinación de atributos que discrimine más que cuando se usan complejos. De esta forma además se demuestra que aunque no exista información relacional, el sistema de representación híbrido sigue aportando alguna ventaja frente a una representación proposicional directa (comparar dos últimas filas de la tabla 8.8). La mejora en rendimiento es mínima, pero en cobertura el sistema ERR-PRyC sin información relacional extiende el doble de rutas que el sistema con representación proposicional directa, y hasta una con cierta diversidad.

Conclusiones

Los resultados empeoran ligeramente sin utilizar información relacional en el aprendizaje, tanto en rendimiento como en cantidad de rutas predichas y su diversidad. La pérdida es mayor al eliminar interacciones proteína-proteína que interacciones de complejos.

A la hora de usar este conocimiento para *aplicarlo a otros problemas*, se puede decir que si la información relacional no fuera muy relevante en el aprendizaje y el principal interés está en el rendimiento del sistema, se podría simplificar el problema al uso de una representación proposicional directa, porque sobre el resto de datos la representación relacional no aporta ventajas. Sin embargo, si existe alguna limitación en la cantidad de predicciones por clase o se busca diversidad en reglas, es mejor el enfoque de aprendizaje híbrido del sistema ERR.

8.5.2. Predicción con Anotaciones de Compañeros de Interacción

En esta sección se analizan los resultados tras añadir más información relacional, al contrario que en la sección anterior, donde se eliminaba dicha información relacional. Lo que se añaden son más niveles de relaciones, en concreto, anotaciones funcionales de los compañeros de interacción, no de la proteína principal sobre la que se aplica el sistema.

Se incluyen cinco tipos diferentes de anotaciones funcionales de las proteínas: familias de proteínas (de *Pfam* [Finn et al., 2010]), dominios de proteínas (de *InterPro* [Hunter et al., 2009]), procesos biológicos, componentes celulares y funciones moleculares. Las tres últimas categorías de anotaciones se extraen de *Gene Ontology* [Ashburner et al., 2000], y sólo incluyen resultados experimentales, que ignoran las anotaciones automáticas, para evitar en lo posible sesgos inducidos por solapamientos con otras fuentes de anotación. Todas estas anotaciones se extraen de Ensembl versión 56 [Hubbard et al., 2009] a través de BioMart [Smedley et al., 2009].

Adicionalmente se añade como anotación de un compañero de interacción a qué rutas de Reactome pertenece (si pertenece a alguna); información implícitamente incluida en el método Glaab et al. (por considerar sólo proteínas interaccionando con alguna de la ruta original), pero completamente desconocida para el sistema ERR.

Las nuevas fuentes de información se representan como predicados lógicos binarios, que asocian a un identificador de proteína el identificador de la anotación correspondiente (ver figura 8.14). Dichos predicados se añaden al lenguaje de representación del conocimiento de partida para la extensión de rutas (definido en la figura 7.2).

```
pfam_domain(proteinID,pfamID).
interpro_domain(proteinID,interproID).
go_annotation_bioProcess(proteinID,goID).
go_annotation_cellComponent(proteinID,goID).
go_annotation_molFunction(proteinID,goID).
protein_in_pathway(reactID,proteinID).
```

Figura 8.14: Fragmento de lenguaje de representación del conocimiento asociado a anotaciones funcionales.

También se incluye como parte del sesgo del lenguaje las directivas que muestra la figura 8.15, para que el aprendizaje restrinja a que las anotaciones sólo se asocien a compañeros de interacción, y no a proteínas principales.

Tras incluir todas las anotaciones en un sistema con la misma configuración que ERR-PRyC, excepto para la frecuencia mínima (0,2) y la profundidad máxima (3), se obtienen los resultados que muestra la figura 8.16.

La figura 8.16 presenta una gran mejora de los resultados, con un rendimiento cercano al doble del alcanzado con el sistema de referencia ERR-PRyC. También se extienden 8 rutas más, de las cuales 3 más son con varias reglas. El solapamiento entre las predicciones de distintas rutas se reduce un poco, habiendo menos proteínas añadidas diferentes entre clases.

Hay que destacar que la extracción de patrones frecuentes está limitada frente al sistema ERR-PRyC (por razones de coste computacional) y aún así se consigue mejorar la predicción de forma tan elevada, dejando un margen de mejora adicional.

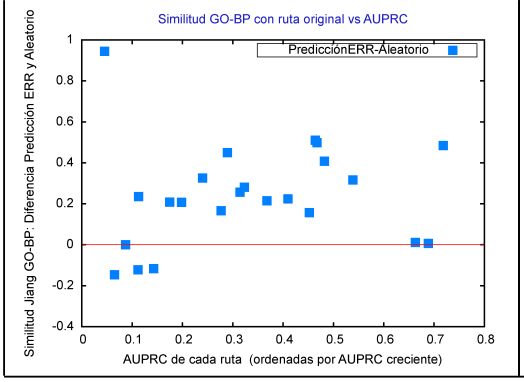
La tabla 8.9 compara los resultados de inclusión de anotaciones de compañeros de interacción, separado entre pertenencia a una ruta de Reactome, y el resto de anotaciones

```
(ppinteraction_pair(+ProtID,\X), interpro_domain(X,#)).
(ppinteraction_pair(+ProtID,\X), pfam_domain(X,#)).
(ppinteraction_pair(+ProtID,\X), go_annotation_bioProcess(X,#)).
(ppinteraction_pair(+ProtID,\X), go_annotation_cellComponent(X,#)).
(ppinteraction_pair(+ProtID,\X), go_annotation_molFunction(X,#)).
(ppinteraction_pair(+ProtID,\X), protein_in_pathway(#,X)).

(complex_interaction(+ProtID,\X), interpro_domain(X,#)).
(complex_interaction(+ProtID,\X), pfam_domain(X,#)).
(complex_interaction(+ProtID,\X), go_annotation_bioProcess(X,#)).
(complex_interaction(+ProtID,\X), go_annotation_cellComponent(X,#)).
(complex_interaction(+ProtID,\X), go_annotation_molFunction(X,#)).
(complex_interaction(+ProtID,\X), protein_in_pathway(#,X)).
```

Figura 8.15: Fragmento de sesgo del lenguaje para anotaciones de compañeros de interacción.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,3155	0,7776	23 / 8	303 / 78,22 %



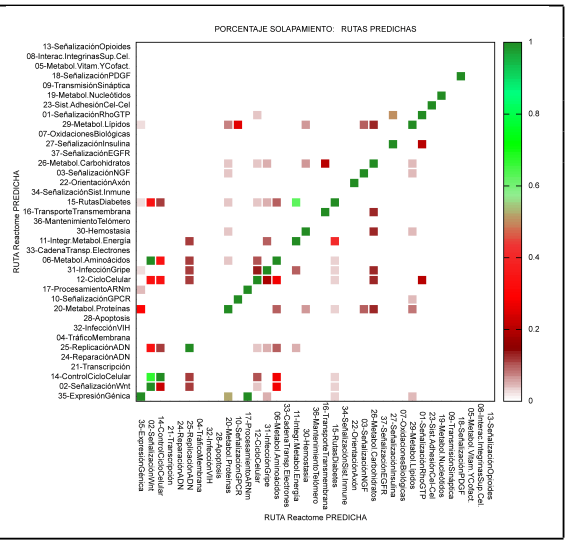


Figura 8.16: Resumen de resultados con anotación de compañeros de interacción.

funcionales de bases de datos externas (Pfam, InterPro y GO). Los resultados indican que los dos subconjuntos de anotaciones considerados permiten conseguir una gran mejora de forma independiente al otro, aunque ligeramente inferior a la conseguida con el uso de todas las anotaciones. Además, cabe destacar que el uso de anotaciones de Pfam, InterPro y GO incrementan la diversidad de las predicciones (con mayor cantidad de rutas extendidas por más de una regla), mientras que las anotaciones de Reactome mantienen dicha diversidad frente a ERR-PRyC e incluso disminuye el porcentaje de proteínas diferentes entre rutas.

Comparando desde otra perspectiva, ante esta gran cantidad de conocimiento relacional, una representación proposicional directa (ver sección 8.4.2) no puede abarcar toda esta información sin complicar en exceso el modelo de datos. Ahora no se trataría sólo de representar en una sola fila los posibles compañeros de interacción de una proteína, sino también todas las anotaciones de los mismos, y con anotaciones de diferentes bases de datos. Las anotaciones añaden una dificultad adicional, como es transformar a representación proposicional los atributos multi-valuados, es decir, aquellos con más de un posible valor a la vez para la misma instancia. Por ejemplo, las anotaciones con varios términos del

Tabla 8.9: Evaluación numérica de la extensión de Reactome con anotaciones de compañeros de interacción.

Sistema	**AUPRC**	**AUROC**	**nºrutas** (total/>1regla)	**nºproteínas añadidas** (total/ %dif.)
ERR-PRyC	0,1695	0,7028	18 / 5	249 / 87,55 %
todas las anotaciones	0,3155	0,7776	23 / 8	303 / 78,22 %
sólo Pfam, InterPro y GO	0,2881	0,7679	23 / 10	294 / 82,31 %
sólo Reactome	0,2939	0,7943	21 / 5	260 / 76,15 %

mismo vocabulario (Pfam, InterPro, GO, etc.), o pertenencia a varios grupos del mismo tipo (pertenencia a varias rutas metabólicas, varios complejos, etc.). La solución proposicional podría consistir en crear un atributo booleano por cada posible valor. Esto provocaría muchísimas celdas vacías en la tabla única de datos. Además, habría que repetir las propiedades ancladas al grupo (si las hubiera) para cada aparición de ese grupo, provocando muchos valores repetidos en la tabla. Sin embargo, con una representación relacional, varias de estas propiedades se podrían calcular con programación lógica inductiva con agregados *on-line*, o fácilmente con lógica deductiva *off-line*.

Conclusiones

Los resultados mejoran mucho con las anotaciones de los compañeros de interacción, y sin necesidad de usar información directa de la proteína sobre la que se quiere determinar la función (o pertenencia a ruta), sino de otras relacionadas que sí pueden tener anotaciones. No obstante, es cierto que se requiere más información que la simple secuencia, aunque no sea directa de la secuencia original, sino de las relacionadas.

Para la *aplicación a otros problemas*, ante la presencia de tanta información relacional no conviene aplicar una representación proposicional directa, pues se perdería mucho conocimiento, o la tabla de datos sería de gran tamaño y con mucha información redundante. Por otro lado, ante atributos multi-valuados con cientos o miles de valores diferentes (como son los de las anotaciones funcionales usados en esta sección) es importante limitar cautelosamente la cantidad de patrones frecuentes a extraer, para no desbordar al sistema de aprendizaje. Finalmente, cabe destacar que esta aproximación se puede usar también para un conjunto restringido de proteínas sin caracterizar, porque las anotaciones que se necesitan son de los compañeros de interacción.

8.6. Incremento del Conocimiento con Anotaciones de Proteínas Principales

Motivación-Hipótesis: ¿Mejora la predicción de función al incluir anotaciones asociadas a cada proteína principal como nuevos atributos/predicados?

En comparación con la sección anterior, en esta se permite acceder a la información de anotaciones de todas las proteínas, no sólo cuando una proteína es compañera de interacción de otra, lo cual restringe mucho menos el conocimiento añadido, por no estar asociado a la relación (ver figura 8.15), sino a la instancia principal (ver figura 8.17).

La inclusión de las anotaciones para cualquier proteína se puede interpretar como uso

```
(protein(\X,_,_,_), interpro_domain(X,#)).
(protein(\X,_,_,_), pfam_domain(X,#)).
(protein(\X,_,_,_), go_annotation_bioProcess(X,#)).
(protein(\X,_,_,_), go_annotation_cellComponent(X,#)).
(protein(\X,_,_,_), go_annotation_molFunction(X,#)).
```

Figura 8.17: Fragmento de sesgo del lenguaje para anotaciones de cualquier proteína.

de homología, aunque de forma indirecta. Las anotaciones se pueden considerar atributos calculados a partir de algún tipo de información de similitud con otros genes o proteínas, ya que dichas anotaciones muchas veces proceden de experimentos en otra especie, cuyos resultados se extrapolan automáticamente por homología a las bases de datos de anotación del resto de especies. Una vez que se permite el uso de anotaciones de las proteínas principales (denominándolo homología indirecta), no se restringe la inclusión de anotaciones de compañeros de interacción. Porque al permitir ya el uso de anotaciones de las proteínas principales, semánticamente da igual usar también las anotaciones de las proteínas con las que se enlaza, ya que se requiere más información aparte de la secuencia de la proteína principal, por lo que con menos influencia se puede usar más información de las proteínas relacionadas.

Como se puede observar en la figura 8.17, se usan las mismas fuentes de anotación que en la sección anterior (Pfam, InterPro y Gene Ontology), excepto Reactome que es el objetivo de aprendizaje.

En esta experimentación con anotaciones de cualquier proteína, se mantiene la misma configuración que ERR-PRyC, excepto para la frecuencia mínima (0,2) y la profundidad máxima (3), que limitan la cantidad de patrones frecuentes extraídos por razones de coste computacional, igual que al añadir nuevos predicados multi-valuados en la sección 8.5. En la tabla 8.10 se presenta el resumen de resultados de la extensión de las rutas de Reactome permitiendo el uso de diferentes subconjuntos de anotaciones. Al observar la tabla 8.10 se

Tabla 8.10: Comparación de la extensión de Reactome con anotaciones de proteínas principales (homología indirecta). Ordenación por AUPRC creciente, de arriba a abajo.

Sistema	**AUPRC**	**AUROC**	**nºrutas** (total/>1regla)	**nºproteínas añadidas** (total/ %dif.)
ERR-PRyC	0,1695	0,7028	18 / 5	249 / 87,55 %
GO-MF	0,1752	0,7325	21 / 12	278 / 83,09 %
GO-CC	0,2202	0,7721	22 / 16	185 / 80,00 %
Pfam	0,2554	0,7474	17 / 5	149 / 71,14 %
InterPro	0,2868	0,7595	22 / 9	226 / 78,76 %
InterPro y Pfam	0,2888	0,7644	23 / 8	239 / 78,24 %
GO-BP	0,3290	0,7880	15 / 7	132 / 93,94 %
GO	0,3599	0,7926	26 / 14	239 / 84,94 %
Pfam y GO	0,3973	0,7982	23 / 11	171 / 84,21 %
Pfam, InterPro y GO	0,4035	0,7914	23 / 11	195 / 82,56 %
InterPro y GO	0,4060	0,7922	23 / 14	237 / 83,97 %

puede comprobar que la inclusión de cualquier anotación mejora los resultados frente al sistema de referencia ERR-PRyC, a pesar de tener restringida la generación de patrones. La

combinación de anotaciones que proporciona un mayor rendimiento en términos de AUPRC es cuando están todas juntas, aunque con una ligera mejora si no se incluye Pfam, que quizá sea redundante con InterPro (ver 2 últimas filas de la tabla 8.10).

También se incluyen las anotaciones de cada ontología GO por separado, para verificar que GO-BP es la que más aporta individualmente con diferencia, también frente a InterPro y Pfam por separado. Se trata de un resultado esperado por ser en los procesos biológicos el nivel de anotación funcional más cercano a las rutas de Reactome que comprenden el objetivo de predicción; estando más distantes (de más a menos AUPRC) los tipos de dominios (InterPro), las familias de dominios (Pfam), los componentes celulares (GO-CC) y finalmente las funciones moleculares (GO-MF). De hecho, es necesario recordar que se necesita ortogonalidad entre los atributos de entrada y el objetivo de predicción. Muchas veces hay relaciones desconocidas o indirectas (procedentes de la homología, o alguna anotación por homología), como podría suceder en este caso entre las anotaciones GO-BP y el objetivo de predicción que son el vocabulario de Reactome, de ahí el elevado rendimiento cuando se usan estas anotaciones.

En cuanto a la diversidad de las proteínas añadidas por cada combinación de anotaciones, todas se mantienen en el mismo rango de valores que el sistema ERR-PRyC. Excepto las que sólo incluyen anotaciones de dominios (Pfam y/o InterPro), donde el porcentaje de proteínas diferentes entre rutas baja del 80 %; y cuando sólo se usan anotaciones de GO-BP, que el porcentaje sube hasta casi un 94 % de proteínas distintas.

Por otro lado, a continuación se analizan las **diferentes representaciones relacionales de los atributos multi-valuados**, como son las anotaciones utilizadas en esta sección. Como punto de partida, se decide incluir un predicado diferente para cada tipo de anotación, para diferenciar las fuentes de datos, según muestra la figura 8.14. Cada predicado tiene dos campos: el identificador de la proteína y el término de anotación según el vocabulario correspondiente. El segundo atributo es multi-valuado por poder estar anotada una misma proteína con más de un término de un mismo vocabulario. De forma que un mismo identificador de proteína puede tener muchos literales distintos de un mismo predicado (predicados instanciados con valores concretos). A la hora de extraer patrones frecuentes, el término de anotación (segundo atributo de un predicado de anotación) se puede equiparar: 1) como una variable ó 2) como una constante. En el caso de ser variable, durante la búsqueda de patrones se genera un literal por predicado; mientras que si es constante, se genera un literal por cada valor diferente del vocabulario de anotación, dando lugar a muchos más patrones. Así, en el primer caso (equiparación como variable) se podría localizar otro literal con exactamente el mismo valor, generando patrones interesantes, aunque poco frecuentes, como por ejemplo: *ppinteraction(protA,protB), interpro_domain(protA,X), interpro_domain(protB,X)*, que representa un par de proteínas interaccionando con un mismo tipo de dominio, sin necesidad de especificar cuál. Sin embargo, en el segundo caso (equiparación como constante) se pueden generar muchos más patrones con suficiente frecuencia como para ser seleccionados por el algoritmo de extracción de patrones, como por ejemplo: *complex_interaction(A,B),not(B=A), go_annotation_cellComponent(B,'GO:0005672')*. Además, como constante se especifica el valor concreto de la anotación (en el ejemplo, el componente celular), que puede ser de utilidad en la interpretación biológica. La equiparación de términos de anotación como constante presenta una gran mejora en términos de rendimiento frente a sólo equipararlo como variable. Finalmente, para aprovechar las ventajas de las dos, se decide utilizar ambas equiparaciones al incluir anotaciones.

Conclusión

Usar información procedente de homología indirecta mejora notablemente la predicción. No obstante, requiere más información (y más compleja) que sólo la secuencia de aminoácidos de la proteína nueva sobre la que predecir. Además, sin anotaciones, el modelo aprendido se puede aplicar a un conjunto más estricto (restringido) de proteínas sobre las que no se tienen anotaciones disponibles.

Para *aplicarlo a otros problemas*, si se decide usar información de homología, también se podrían incluir muchas más fuentes de datos, con conocimiento obtenido a partir de alguna similitud por homología, indirecta o directa (ver sección 8.7). Pero hay que asegurarse de si la combinación de múltiples fuentes de datos con homología aporta ventajas frente a buscar la secuencia más parecida y tomar la anotación de la misma directamente como predicción.

8.7. Predicción con Homología Directa

Motivación-Hipótesis: ¿Los resultados de predicción de anotación mejoran añadiendo datos de homología directa como entrada del sistema?

En este trabajo, se denomina predicción con homología directa, al uso explícito de relaciones de similitud entre secuencias en el aprendizaje. En vez de incluir anotaciones funcionales, como en la homología indirecta, se añade directamente qué pares de proteínas son homólogos como dato de entrada.

El sistema ERR es un método válido *en ausencia de homología*, al contrario que muchos sistemas de predicción de anotación funcional que se basan o incluyen homología como fuente de datos (ver sección B.3). Sus rendimientos son mucho mayores que los presentados en este trabajo, incluso usando enfoques de aprendizaje parecidos [Clare et al., 2006; Vens et al., 2008], y sin olvidar la dificultad adicional para ERR en la extensión de rutas por su definición no estándar y subjetiva frente a otros vocabularios de anotación.

Con la experimentación de esta sección simplemente se quiere demostrar que el método de predicción ERR-PRyC permite fácilmente su ampliación con el uso de datos de homología y que, por supuesto, el rendimiento de la predicción mejora, al igual que en otros predictores de función que usan homología. No obstante, los resultados presentados aquí probablemente sean peores que una búsqueda de similitud de secuencia con BLAST cuando realmente existe homología. Pero el alcance de esta sección no es superar a BLAST con la búsqueda del mejor sistema con homología, sino comprobar que una mínima cantidad de información de homología directa mejora el rendimiento de la predicción de ERR-PRyC.

En esta experimentación, se incluyen datos de homología directamente, pero de una forma limitada en varios sentidos, por coste computacional. Al igual que en secciones previas, la búsqueda de patrones está restringida con una frecuencia mínima mayor (0,2) y una profundidad máxima menor (3) que en la configuración del sistema ERR-PRyC. Por otro lado, sólo se incluyen las relaciones de homología con la primera secuencia homóloga de cada proteína. Finalmente, las relaciones de homología sólo se calculan dentro de la especie humana.

Las relaciones de homología se representan como una relación binaria más entre dos proteínas, al igual que las interacciones proteína-proteína y las interacciones en complejos, como se puede observar en la figura 8.18, incluyendo 69.452 nuevos predicados instanciados. Además, también se incluyen las anotaciones en Reactome sólo de los homólogos (ver figura 8.18), de forma que se tenga un predictor parecido a lo que realizaría un sistema de anotación por homología (“mirar anotación de la proteína que se parece más”): buscar la secuencia más similar (primer homólogo) y comprobar qué anotación tiene en el vocabulario objetivo

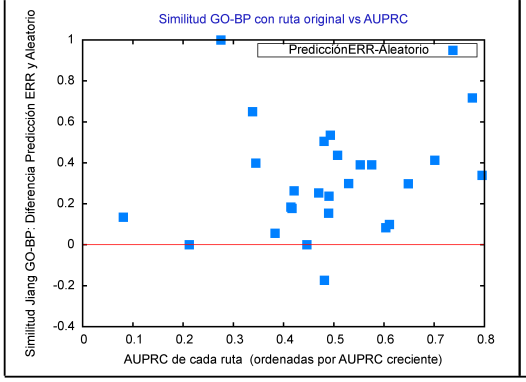
(anotación en Reactome del homólogo), para asignárselo a la nueva proteína. También se permite la búsqueda de homólogos de compañeros de interacción.

```
homolog(proteinID,proteinID).
protein_in_pathway(reactID,proteinID).

(homolog(+ProteinID,\X), protein_in_pathway(#,X)).
```

Figura 8.18: Fragmento de sesgo del lenguaje para homólogos y sus anotaciones en Reactome.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,4566	0,8505	28 / 16	299 / 82,61 %



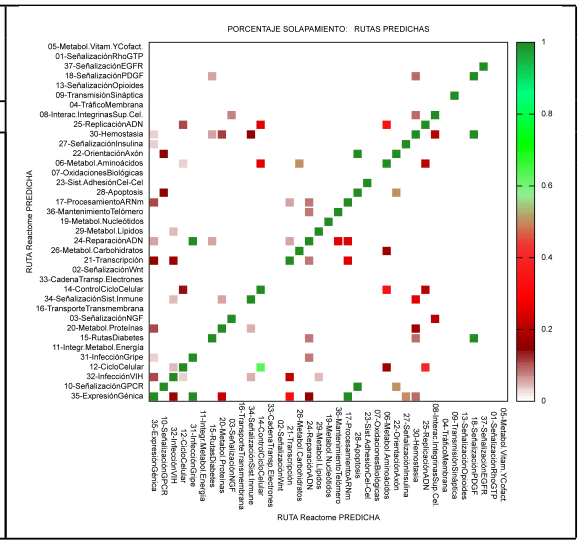


Figura 8.19: Resumen de resultados con homología directa.

A pesar de las limitaciones en el uso de homología directa, en la figura 8.19, se observa que el rendimiento (en AUPRC) del sistema con homología directa es mucho mejor que en el sistema ERR-PRyC (figura C.1), y también un poco mejor que con el uso de homología indirecta (figura 8.10). También se consiguen extender 10 rutas más que en el sistema ERR-PRyC, y más del triple con variabilidad de reglas.

Conclusión

La extensión de rutas de Reactome es mejor incluyendo relaciones de homología que cuando éstas no se usan. Aunque el sistema ERR también es válido si existe homología, no asegura ventajas sobre una búsqueda BLAST. Pero gracias a no incluir información de homólogos, el sistema es aplicable sobre un conjunto restringido de difícil anotación, cuando hay ausencia de homología, es decir, las proteínas sobre las que una búsqueda BLAST no produce resultados.

Para el uso de homología directa en la *aplicación a otros problemas* con una representación relacional y con la transformación a proposicional previa mediante patrones frecuentes, hay que tener muy en cuenta las limitaciones de memoria. Porque se produce un crecimiento muy elevado de predicados, y un aumento exponencial de los patrones frecuentes generados, requiriendo mucha potencia computacional. La cantidad de predicados crece fácilmente, por

ejemplo, al combinar homología directa e indirecta, o incluir todos los homólogos y no sólo el primero, o considerar homólogos también en otras especies. Por otro lado, el margen de mejora esperado también es muy amplio, pues se ha comprobado que con la inclusión de información de homología limitada ya se consiguen grandes incrementos en rendimiento y cobertura. No obstante, es cierto que el sistema de predicción con homología ideado debería ser capaz de superar en algún sentido las refinadas y sencillas búsquedas BLAST. Por lo tanto, la decisión del uso de homología o no depende de las capacidades computacionales que se puedan emplear para resolver el problema.

8.8. Análisis de Relaciones Indirectas entre Genes y Proteínas

Motivación-Hipótesis: ¿Cómo influyen las relaciones de homología e isoformismo entre los conjuntos de entrenamiento y test en el aprendizaje automático?

Para reducir la nomenclatura, en esta sección se simplifica el significado de proteínas homólogas o isoformas a proteínas 'parecidas'. Así, en este apartado se explican las dificultades provocadas por estas relaciones indirectas de parecido entre las proteínas que conforman los ejemplos de entrenamiento y test.

En primer lugar, hay que evitar confundir el análisis de esta sección con la predicción basada en homología (directa o indirecta) tratada en las dos secciones previas. En términos de aprendizaje automático, se podría decir que en las secciones previas la homología influye para la selección de atributos o características, mientras que en la sección actual afecta a la elección de ejemplos de entrenamiento y test. Ambos aspectos ligados al concepto de homología no se pueden obviar. Mientras que en los apartados anteriores se analiza la restricción del método para que se pueda aplicar en ausencia de homología, en este caso se analiza cómo evitar un sesgo en la evaluación de los resultados, para no dar una visión de los resultados más optimista que la realidad.

Así, se necesita una separación extremadamente cuidada de los datos de entrenamiento y test que se usan en un proceso de aprendizaje automático, sin ninguna relación de alta semejanza entre ellos, para no sobre-estimar el rendimiento de la predicción, que se evalúa sobre el conjunto de test [Hobohm et al., 1992]. De forma que si hay ejemplos muy parecidos con una muestra en el conjunto de entrenamiento y otra en el de test, se sesga la fiabilidad de la predicción hacia arriba por estos ejemplos de alta similitud. Mientras que ante un conjunto de nuevos ejemplos de aplicación, menos parecidos al entrenamiento que el test, la predicción posiblemente sería más desfavorable; como sucede frecuentemente con proteínas no anotadas, por falta de homología con las anotadas.

Las isoformas producen un efecto similar a los homólogos, incrementado incluso por el mayor parecido de las secuencias isoformas frente a las homólogas. Si las bases de datos estuvieran construidas de forma perfecta, las isoformas sólo se distanciarían por los fragmentos de secuencia que diferencian a la isoforma principal de las demás, procedentes de la fragmentación alternativa, heredando prácticamente las mismas propiedades, o subconjuntos de ellas. Esto sería equivalente a tener prácticamente el mismo ejemplo, o fragmentos del mismo, en entrenamiento y en test.

Existen diferentes formas de reducir la redundancia por homología, entre dos conjuntos o sobre un mismo conjunto, como ya se ha comentado en la sección 7.2.3. Para la reducción de isoformismo, basta con seleccionar un criterio para seleccionar la forma principal entre todas las que proceden de la expresión del mismo gen; ya que en las bases de datos se suele seleccionar la más larga, pero se está estudiando si esa debería ser realmente la isoforma

principal [Tress et al., 2008; Rodríguez et al., 2012].

Por otro lado, la reducción de redundancia en homología e isoformismo también decrementa notablemente el conjunto de ejemplos disponibles para el aprendizaje, dificultando aún más la tarea, pero haciendo los resultados más cercanos a la realidad. Comparando el tamaño de los conjuntos de datos, en el sistema ERR-PRyC, tras la reducción de homología e isoformismo quedan 1.108 proteínas para el entrenamiento y 546 para el test, muchas menos que las respectivas 3.165 y 1.615 con redundancia entre entrenamiento y test. Aún si se decide eliminar sólo la redundancia entre entrenamiento y test pero no internamente en cada conjunto, para reducir menos el tamaño de los datos, aplicando una modificación del algoritmo propuesto por Jensen et al. [Jensen et al., 2003a] quedan 1.105 proteínas en el conjunto de entrenamiento y 693 en el de test.

También se puede evaluar la influencia de incluir los ejemplos parecidos en el conjunto de entrenamiento, siguiendo el método de aprendizaje de ERR. Por ejemplo, patrones que no eran frecuentes pueden pasar a ser frecuentes y añadirse un atributo en la inducción del árbol de decisión (aunque no se sabe si ese atributo será relevante también o no en otras clases). O patrones frecuentes que se añadían al límite del umbral de frecuencia originalmente, por ser una característica muy específica, sólo visible en algunos ejemplos de esa clase; al añadir nuevos ejemplos, puede que desaparezca ese patrón como frecuente, aunque fuera útil como atributo (lo cual puede ser malo o bueno, a priori no se sabe). No obstante sería complejo determinar qué ejemplos incluir y cuáles no en el conjunto de entrenamiento, debido a la naturaleza del problema multi-clase y multi-etiqueta que no hace que una división aleatoria sea válida.

Conclusión

Cuando hay homología entre conjuntos, los resultados de rendimiento sobre el test pueden estar sobrestimados, sin mostrar la realidad en caso de que se prediga sobre ejemplos sin mucho parecido a los datos de entrenamiento. Es decir, computacionalmente se diría que no se estima sobre el peor caso. Cuando ***no*** hay parecidos entre los conjuntos de entrenamiento y test, hay muchos menos ejemplos para aprender y los resultados que se generan pueden presentar menor rendimiento, aunque están más cercanos a la realidad, ante la ausencia de fuertes similitudes con los datos de entrenamiento.

En la *aplicación a otros problemas* se debe recordar siempre que el aprendizaje automático en biología molecular se complica con las relaciones indirectas entre genes y proteínas. Se debe eliminar la redundancia (isoformas y homólogos) entre conjuntos de entrenamiento y de test, así como internamente. A lo sumo, se podría aceptar un conjunto de entrenamiento redundante (incrementando el tamaño del mismo), siempre que el test se mantenga no redundante, para evitar sobrestimar la fiabilidad de la predicción dada.

8.9. Conclusiones

En este apartado final se resumen brevemente las conclusiones de cada uno de los análisis realizados en cada sección de este capítulo.

- Para la predicción de asociaciones funcionales entre pares de proteínas también se puede aplicar Programación Genética, obteniendo una tasa de aciertos en los mismos niveles que varios algoritmos de AA por defecto, pero consiguiendo una gestión particular mejorada de los valores desconocidos.

- Al aplicar un multi-clasificador único el rendimiento puede ser mejor, dependiendo de lo que se priorice en la evaluación, pero lo más relevante es un solapamiento muy elevado entre las predicciones de distintas rutas biológicas o clases.
- Es equivalente extraer los patrones frecuentes para todas las clases a la vez o por separado pero juntándolos posteriormente en un multi-clasificador, siendo más similares conforme disminuye la frecuencia y aumenta la profundidad. El rendimiento es mayor al usar los mismos patrones para clasificar sobre todas las rutas que al dividirlos por ruta, pero esta última opción tiende a presentar un menor solapamiento y la mayor diversidad en las predicciones.
- Variando la representación del conocimiento, con una representación relacional directa, no se consigue aprender, generando una clasificación aleatoria, porque un solo predicado independiente por nodo del árbol de decisión no es capaz de discriminar, necesitando combinarse con otros. Por otro lado, con una representación proposicional directa, el rendimiento es igual o peor que con aprendizaje híbrido, con muy pocas predicciones, sin variabilidad en las mismas y con un modelo de clasificación menos interpretable.
- Sin información relacional, los resultados empeoran, tanto en rendimiento como en cantidad de rutas predichas y su diversidad. Contrariamente, los resultados mejoran bastante con información relacional adicional, como son las anotaciones de los compañeros de interacción, y sin necesidad de usar información directa de la proteína sobre la que se quiere determinar la función (o pertenencia a ruta).
- Usar información procedente de homología indirecta o directa mejora notablemente la predicción, pero requiere más información (y más compleja) que sólo la secuencia de la proteína nueva. Sin anotaciones, el modelo aprendido se puede aplicar a un conjunto más estricto (restringido) de proteínas sobre las que no se tienen ni anotaciones disponibles (homología indirecta), ni proteínas similares (homología directa).
- Si existe homología entre las proteínas de los conjuntos de entrenamiento y test, no se estima sobre el peor caso, tendiendo a ser la evaluación del rendimiento sobre el test mejor que en la realidad, en caso de que se prediga sobre proteínas sin mucho parecido a los datos de entrenamiento.

Además, cada una de las técnicas y representaciones presentados en este capítulo se pueden aplicar para resolver otros problemas con aprendizaje automático. En la tabla 8.11 se recopila el contenido de los párrafos al final de cada sección que hablan de la *aplicación a otros problemas* del enfoque planteado. Así, se obtiene un conjunto de estrategias (segunda columna), a aplicar si las características u objetivos del problema coinciden con los que se exponen (primera columna), principalmente para dominios de clasificación multi-clase y/o multi-etiqueta con información relacional. No siempre las estrategias propuestas se limitan a su uso en problemas bioinformáticos, sino que en unos casos es aplicable a cualquier otro dominio. No obstante se detallan los casos particulares de aplicación en Biología, principalmente en anotación funcional de proteínas.

Tabla 8.11: Estrategias de aplicación de enfoques de AA según características del problema.

Sec.	Características/Objetivo	Estrategia	Observaciones
8.1	Mantener semántica biológica de gran cantidad de valores desconocidos	Diseño operador particular/personalizado para gestión de desconocidos con Programación Genética	Mejor gestión de desconocidos, sin pérdida de rendimiento
8.2	Aprendizaje multi-clase y multi-etiqueta, con: ▪ bajo coste computacional ▪ simplicidad en modelo aprendido ▪ preferencia cobertura sobre precisión ▪ sin pérdida o con mejora del rendimiento ▪ sin importar alto solapamiento entre clases	Multi-clasificador	Si una o varias de las pre-condiciones no siguen el objetivo buscado, la opción es el uso de N clasificadores, individuales por cada etiqueta, que consigue predicciones más diversas y precisas
8.3	En el uso de aprendizaje híbrido: relacional (extracción patrones frecuentes) + proposicional (árbol de decisión):		Buscar estrategia equilibrada en función de la combinación de condiciones planteadas que se cumplan en el problema
	▪ Priorizar rendimiento, sin mucha capacidad computacional (para problema biológico: con muchos ejemplos, predicados y niveles sucesivos de relaciones)	- Extracción de patrones: ▪ conjunta (todas las clases a la vez) ▪ frecuencia mínima en torno al 0,2 y profundidad en 3 ó 4 niveles (centenas de patrones) - Multi-clasificador	Se produce alto solapamiento entre distintas clases, por utilizar multi-clasificador
	▪ Priorizar rendimiento, independientemente del coste computacional	- Extracción de patrones: ▪ conjunta (todas las clases a la vez) ▪ frecuencia mínima muy baja y un nivel de profundidad lo más alto posible (miles de patrones) - Clasificadores individuales por clase	

	▪ Priorizar diversidad	- Extracción de patrones: ▪ separada por clases ▪ frecuencia mínima superior a 0,2 ó 0,3 (no baja) - Clasificadores individuales por clase	
8.4	En el uso de aprendizaje simple, no híbrido:		
	▪ Sin apenas relaciones encadenadas y uso de aprendizaje más sencillo que el híbrido	Representación y aprendizaje **relacional** directo	Si se necesitan más de 2 ó 3 relaciones encadenadas para alcanzar información que discrimine, esta estrategia no es adecuada, porque sólo consulta un predicado por paso en la clasificación
	▪ Preferir representación y aprendizaje clásico ▪ Sin restricción en cantidad y diversidad de predicciones ▪ Priorizar cobertura y rendimiento ▪ Sin interés en mantener la semántica de las relaciones	Representación y aprendizaje **proposicional** directo	Desventajas: - Se permite redundancia de información - Modelos de interpretación más complejos
8.5-8.8	En el uso de información relacional:		
8.5.1	▪ No hay información relacional ▪ Sin restricción en cantidad y diversidad de predicciones	Representación y aprendizaje **proposicional** directo	
	▪ Hay información relacional	Representación relacional	Desaconsejable usar una representación proposicional directa, por pérdida de semántica e información

	▪ Hay información relacional y: • Atributos multi-valuados con cientos o miles de valores. En Biología, anotaciones en vocabularios amplios • Existencia de información adicional abundante de los elementos, relacionados en varios niveles. En Biología, por ejemplo, anotaciones múltiples de proteínas (principales o con las que se interacciona) en diversos vocabularios	- Aconsejable aprendizaje híbrido. Con extracción de patrones frecuentes (por niveles sucesivos de relaciones), pero con cautela, para no desbordar el sistema de aprendizaje supervisado posterior - Usar toda la información adicional y relaciones disponibles, limitando por restricciones computacionales	- Ventaja: mejora sobre predicción sin relaciones - Desventaja: en problemas biológicos, implica la necesidad de más información (y más compleja) que la simple secuencia
8.5.2	▪ Hay información relacional y: • Ausencia de información asociada a instancia principal sobre la que predecir. En Biología, **sin** anotaciones de proteína principal • Existencia de información adicional de elementos relacionados. En Biología, anotaciones de proteínas con las que se interacciona	*(Ídem celda superior)*	- Ventajas: - Mejora sobre predicción sin relaciones - El modelo aprendido se podría **aplicar a un conjunto restringido** de proteínas poco caracterizadas, sobre las que apenas se dispone de información, a parte de la secuencia - Desventaja: en problemas biológicos, en este escenario, no se requiere **información adicional** a la secuencia de la proteína sobre la que se quiere predecir, pero sí **de las proteínas relacionadas**
8.6	▪ Hay información relacional y: • Existencia de información asociada a instancia principal sobre la que predecir. En Biología, **con** anotaciones de proteína principal	*(Ídem celda superior)*	- Ventaja: mejora sobre predicción con información adicional de elementos **relacionados** (escenario anterior) - Desventaja: en problemas biológicos, este escenario requiere información adicional a la secuencia de la proteína principal, por lo que no se puede aplicar a proteínas sin caracterizar

8.7	▪ Hay información relacional y: • Existencia de múltiples relaciones de parecido entre elementos sobre los que predecir. En Biología, relaciones de homología entre proteínas principales	*(Ídem celda superior)*	- Ventaja: mejora sobre predicción con información adicional de elementos **principales** (escenario anterior). - Desventajas: - Igual que en el anterior, en problemas biológicos, no se puede aplicar a proteínas sin caracterizar - No se puede asegurar mejor rendimiento que la predicción de anotación funcional basada en similitud de secuencia (BLAST)
8.8	Existencia de relaciones de parecido (en Biología, proteínas homólogas o isoformas) entre ejemplos de entrenamiento y test	Construir conjuntos no redundantes, o con todos los ejemplos parecidos en un mismo subconjunto	Riesgo de evaluación más optimista de la real, sin estimar sobre el peor caso

Capítulo 9

Análisis y Discusión: Predicción en Biología de Sistemas con Aprendizaje Automático

9.1. Comparación

En este apartado se comparan los dos problemas afrontados en esta tesis realizando una discusión conjunta.

Ambos problemas versan sobre la predicción de anotación funcional encuadrados en la Biología de Sistemas (marco de las redes biológicas), con distintos enfoques a la hora de usar o predecir relaciones.

El primer problema consiste en la predicción de asociaciones funcionales entre pares de proteínas (AFPP) en el organismo modelo *E.coli* (ver capítulo 6). El segundo es la extensión de rutas biológicas en humanos, basándose en propiedades simples e interacciones puntuales (ver capítulo 7). En un caso se extiende la red con nuevas asociaciones funcionales puntuales, y en el otro se añaden elementos a un grupo de proteínas ya relacionadas. En un contexto de anotación, se trata de saber qué proteína pertenece a esas redes.

Se eligen dos problemas diferenciados que permiten entender el área de anotación funcional en Biología de Sistemas, se trabaja en ellos a lo largo de la tesis y se presenta una aproximación válida biológicamente para cada uno.

Similitudes

Se han seleccionado específicamente dos problemas que comparten similitudes en el origen relacional de los datos, y en la aproximación y objetivo biológico. Ambos tratan la extensión de redes biológicas, partiendo de unos conjuntos de proteínas y encontrando una que pertenece a uno de ellos. En la siguiente lista se mencionan los puntos comunes entre ambos problemas:

- Objetivo: Extender redes biológicas, porque las redes de interacción no están completas. O anotar proteínas con interacciones o asociaciones funcionales (puntuales o grupales).
- Representación en forma relacional del conocimiento original, puesto que en los dos problemas elegidos las interacciones y asociaciones funcionales entre los componentes son particularmente relevantes, además de ser los datos biológicos intrínsecamente relacionales.

- Uso de Aprendizaje Automático para plantear una solución, que generalice la predicción de anotación funcional en las áreas de los dos problemas elegidos (predicción de asociación funcional por pares y extensión de rutas).
- Integración de diferentes fuentes de información.
- Sistemas útiles por heterogeneidad en la definición y representación de relaciones, tanto en asociaciones funcionales por pares como en rutas. Se trata de llegar a un enfoque unificado, generalizando en ambos casos.
- Aplicables a proteínas poco caracterizadas (sin datos experimentales en asociaciones funcionales por pares o sin homología en rutas).

Diferencias

Una vez analizadas las semejanzas entre ambos problemas, en la tabla 9.1 se exponen las interesantes diferencias respecto a las relaciones, las características de los datos, y la aproximación computacional requerida; siendo la mayoría ya previsibles y comentadas en la descripción de la propuesta de tesis. Se puede observar en la tabla 9.1 en primer lugar que, respecto a los **datos**, comparados en la parte superior de la tabla, los organismos están muy distantes entre sí en el árbol filogenético de la vida, perteneciendo *E.coli* al reino de las bacterias y el ser humano al de los animales. Pero, fundamentalmente, ambos problemas están diferenciados por la información de entrada bien distinta entre sí. En la predicción de AFPP la mayoría de los datos son de alto nivel, procedentes del pre-procesamiento de las secuencias a bajo nivel por métodos o algoritmos externos. Mientras que en la extensión de rutas se parte de características básicas de las secuencias de aminoácidos, así como las interacciones proteína-proteína y de complejos, reducidas a una representación por pares. En realidad se pueden ver como dos tipos de redes conectadas entre ellas, porque parte de las asociaciones funcionales predichas en *E.coli* (las interacciones proteína-proteína y de complejos), se utilizan como entrada para extender las rutas, aunque en humanos. Partiendo ambos de una representación relacional, en el primer caso se pueden integrar varios métodos de predicción existentes y en el segundo integrar datos base, sin apenas procesar. También es importante tener en cuenta que en *E.coli* cada gen sólo se expresa en una proteína, mientras que en humanos existen *N* proteínas asociadas a un mismo gen, lo cual implica una relación más en la representación, así como una gestión de la redundancia que las proteínas isoformas añaden. Además, el conjunto de datos a alto nivel no requiere de una recopilación tan detallada como los de bajo nivel, incluyendo la búsqueda y tratamiento por bases de datos distribuidas.

En la parte central de la tabla 9.1, dedicada al **aprendizaje**, se observa que ambos problemas son diferentes en todos los sentidos, siendo mucho más complicado en el caso de la extensión de rutas (multi-clase, multi-etiqueta, combinación de varios enfoques de aprendizaje, creación del sistema desde cero, etc.). Destaca el uso de un aprendizaje proposicional simple frente a un enfoque híbrido, aunque ambos conjuntos de datos son originalmente relacionales.

Sobre el **número de atributos**, los 9 predicados de la extensión de rutas tienen de 1 a 5 argumentos cada uno, dependiendo del caso. La proporción de valores numéricos o relaciones asociadas a cada instancia de aprendizaje también varía entre ambos problemas. En la predicción de AFPP hay 19 atributos numéricos, con las relaciones implícitas en el cálculo de características derivadas, como por ejemplo la posición de las proteínas en las listas de ranking. Mientras que en la extensión de rutas, a parte de los 10 atributos numéricos o booleanos asociados a cada proteína (simplificando entre los distintos predicados), pueden existir decenas

Tabla 9.1: Comparativa diferenciadora entre predicción de AFPP y extensión de rutas. En la parte superior de la tabla se presentan las diferencias asociadas a los datos de entrada de cada problema, en la parte central las referentes al proceso de aprendizaje y predicción, y en la parte inferior aparecen algunos valores numéricos referentes a los conjuntos finales de datos.

	predicción AFPP	**extensión rutas**
DATOS DE ENTRADA		
organismo	*E.coli* (procariota)	Humano (eucariota)
origen datos	alto nivel	bajo nivel
propiedades individuales	longitud y nºortólogos	longitud,carga,nºisoformas, cromosoma,dominios,etc.
propiedades relacionales	scores y rankings (derivados de asociación)	interacciones PP y complejos, 1 gen con *N* proteínas
predicción basada en:	contenido evolutivo	combinación propiedades simples (ausencia de homología)
APRENDIZAJE Y PREDICCIÓN		
extensión red con:	enlace puntual	enlace con un grupo
definición ejemplo	par de proteínas	1 proteína
clasificación	binaria	multi-clase
nºclases/ejemplo	1	varias (multi-etiqueta)
aprendizaje	proposicional	relacional + proposicional
VALORES CUANTITATIVOS		
sistema predicción	unificar métodos previos, con datos adicionales	crear nuevo sistema desde el principio
nºatributos	19 atributos	9 predicados, con 1-5 argumentos
nºejs. entrenamiento	55.220 pares	1.108 proteínas
nºejs. test	27.610 pares	546 proteínas
nºejs. aplicación	479.582 pares	8.187 proteínas

de interacciones proteína-proteína y en complejos para una misma proteína, dado que los argumentos de estas dos relaciones de interacción son multi-valuados. Hay que hacer notar que en los 9 predicados que indica la tabla 9.1 no están incluidas las anotaciones, que se utilizan en algunas secciones del capítulo 8, que incrementarían notablemente las propiedades asociadas a cada ejemplo.

Sobre el **número de ejemplos**, las últimas filas de la tabla 9.1 muestran claramente que hay muchos más ejemplos para aprender en la predicción de AFPP que en la extensión de rutas, y también tiene un mayor campo de aplicación cuantitativamente hablando, sobre pares de proteínas de *E.coli*. Sin embargo, el modelo de extensión de rutas es aplicable más fácilmente a muchos otros problemas diferentes de anotación funcional, estando la predicción de pares más restringida en este sentido.

Por último, es importante señalar que realmente el problema de extensión de rutas es mucho más complicado, por: usar datos a bajo nivel, necesitando recopilar información de distintas bases de datos y manipular sus diferentes formatos hasta llegar a los conjuntos de datos finales; gestionar el ruido, inconsistencias y redundancias, por relaciones indirectas (homología, isoformismo o semántica similar); centrarse en un organismo más complejo; predecir a nivel de proceso con información a nivel de secuencia; clasificar entre más de dos

opciones (multi-clase); asignar más de una clase por ejemplo (multi-etiqueta); etc. También es un problema más amplio, y por ello se le dedican muchas más secciones en esta tesis.

Estudio/Discusión detallada

A continuación se detallan los aspectos específicos aprendidos al analizar y resolver los dos problemas, predicción de asociaciones funcionales entre pares de proteínas y extensión de rutas, desde diversas perspectivas.

- *¿Qué diferencia cada enfoque de los métodos existentes previamente?*

  La predicción de AFPP propuesta en el capítulo 6 integra métodos de predicción computacionales heterogéneos y permite el descubrimiento de asociaciones funcionales. Otros sistemas de predicción de asociaciones funcionales por pares integran métodos experimentales y/o sólo predicen interacciones físicas.

  Por su parte, la extensión de rutas descrita en el capítulo 7 está basada en una representación relacional que permite combinar propiedades simples de la secuencia. Unos métodos diseñan nuevas rutas partiendo desde cero, no extendiendo las ya existentes. Otros métodos comparten el enfoque de expansión de rutas, pero usando dominios o redes de interacción como información de entrada. Entre las aproximaciones que anotan funcionalmente a nivel de sistema y usando propiedades de la secuencia, dichas características no se calculan fácilmente, y no usan representación ni conocimiento relacional. Por último, otro grupo de métodos comparte un enfoque relacional (representación y aprendizaje), pero en presencia de homología y se aplica sobre organismos más simples que el humano.

  Precisamente, debido a la información de entrada, el primer sistema presenta una limitación a la hora de ser actualizado, necesitando también re-entrenar o actualizar los métodos de predicción computacionales que integra, o utilizando otros nuevos. Como prueba de concepto de la viabilidad de la idea es una buena propuesta, pero no es práctico para un mantenimiento a largo plazo. No obstante, este tipo de dependencias entre varios métodos de predicción son comunes, integrando los resultados de unos como entrada de otros, como en los trabajos de Brunak et al. ya comentados [Jensen et al., 2002b; Bendtsen et al., 2004]. Por todo ello, en el segundo sistema presentado en la tesis se evitan dichas dependencias de otros métodos de predicción, usando una información de entrada más básica, como son las secuencias.

  En resumen, ambas propuestas se diferencian en la información de entrada y en la aplicabilidad de los métodos.

- *Rendimiento alcanzado en el aprendizaje*

  El rendimiento es mucho más alto en la predicción de AFPP que en la extensión de rutas, porque el *Patrón Oro* (del inglés, *Gold Standard*) (fuentes de datos) está mucho más verificado y documentado en el primer caso, y las asociaciones funcionales por pares no tienen la influencia de la perspectiva personal del diseñador de rutas metabólicas, de señal o de regulación. No obstante, en la extensión de rutas, el rendimiento varía mucho según la ruta en cuestión, incrementando generalmente con el tamaño de la ruta, aunque con varias excepciones.

- *Algoritmo seleccionado*

En la predicción de AFPP, el algoritmo de aprendizaje seleccionado, y verificado como mejor bajo las condiciones definidas en el capítulo 6, es AODE, un algoritmo bayesiano (sub-simbólico) que exige una menor independencia entre los atributos, e ignora los valores desconocidos de forma puntual. Cabe destacar que, durante la finalización de esta tesis, se ha entrado en contacto con los autores del algoritmo AODE, principalmente Geoff Webb, profesor de investigación en la Facultad de Tecnología de la Información, de la Universidad Monash, en Melbourne, Australia. Ellos nos han proporcionado nuevas versiones mejoradas del algoritmo, denominadas A2DE y AnDE, aún no finalizadas ni publicadas por completo. Con estas nuevas versiones se quiere evaluar el enfoque consistente en ampliar ligeramente la dependencia entre atributos; en particular, incrementándola de uno (*One*) a *2* o *n* atributos, respectivamente. Al aplicar las nuevas versiones del clasificador bayesiano elegido para predecir AFPP sobre el conjunto de datos definido en esta tesis para *E.coli*, por el momento no se ha conseguido mejorar la tasa de aciertos obtenida por AODE. Estamos trabajando con los autores para resolver los problemas y mejorar los nuevos algoritmos, de forma que se logre su aplicación sobre conjuntos de datos reales.

En la extensión de rutas, según las restricciones descritas en el capítulo 7, la combinación más adecuada para resolver el problema es un aprendizaje híbrido. Éste está formado por una extracción relacional de patrones frecuentes, seguida de la inducción proposicional de árboles de decisión. Como características adicionales se requiere un clasificador independiente por ruta, poca poda, y reordenar las reglas de decisión si se busca una mayor diversidad por ruta. En este problema no se ha comparado con algoritmos alternativos al árbol de decisión porque interesa la interpretación de los resultados que aporta este enfoque simbólico, y el enfoque más cercano, la extracción de reglas de decisión, es prácticamente igual.

- *Relevancia de atributos*

  En la predicción de AFPP en *E.coli* el atributo asociado al método de conservación de genes adyacentes (GC [Dandekar et al., 1998]) es el que presenta una tasa de acierto en positivos más cercana al método unificado propuesto, a lo largo de las 5.000 primeras predicciones.

  En la extensión de rutas de Reactome en humanos ningún atributo es mejor que otro en media, sino que depende de la ruta concreta. Cabe destacar que las interacciones no presentan una mayor relevancia que el resto de atributos.

- *Comparación con métodos alternativos que resuelven una tarea semejante*

  En la predicción de AFPP se compara con la base de datos STRING [Jensen et al., 2009] dedicada a las asociaciones funcionales entre proteínas, tanto a nivel experimental como de predicción. Se determina que STRING es más adecuado para asociaciones en proteínas con información experimental conocida, y el enfoque propuesto en esta tesis para proteínas poco caracterizadas experimentalmente.

  Por su parte, la extensión de rutas propuesta en el capítulo 7 (ERR) se compara con el método de expansión definido por Glaab y colaboradores [Glaab et al., 2010], basado sólo en redes de interacción. Se verifica que existen muy pocas predicciones en común entre ambos métodos, dado que las evidencias en las que se basan difieren claramente. ERR expande las rutas de Reactome con un mayor número de proteínas, al no estar restringido a las que interaccionan explícitamente con la ruta original. Las

predicciones de Glaab et al. presentan una mayor similitud funcional semántica con las rutas originales; mientras que las predicciones de ERR presentan más variabilidad entre las distintas rutas. Es decir, las extensiones de ERR están relativamente poco solapadas entre sí, en comparación con el solapamiento original de las rutas. Por último, ERR es más parecido que Glaab et al. a las rutas originales en sus propiedades moleculares más frecuentes.

El método ERR también se compara con una predicción basada en búsqueda de similitud de secuencia (resultados de la herramienta BLASTP [Altschul et al., 1997]). En este caso, en general, no es esperable que los resultados de ambos enfoques coincidan, porque el método ERR utiliza más información que la contenida en la secuencia de aminoácidos, como es el número de transcritos (isoformas), la longitud del gen que codifica la proteína, interacciones proteína-proteína, participación en complejos de proteínas, e incluso propiedades de secuencia de proteínas compañeras de interacción, accesibles gracias a la representación relacional.

- *Uso y utilidad*

En cuanto a la utilidad de la propuesta de predicción de AFPP, durante esta tesis se demuestra que el método unificado es mejor que cualquiera de los métodos individuales, en precisión y sobre todo en cobertura. Esto hace que la puntuación que proporciona el sistema unificado de predicción se pueda usar como criterio de confianza de las asociaciones funcionales. Dicho criterio es usado por el servidor de predicciones EcID, presentando una medida única que permite seleccionar un conjunto de asociaciones más probable para proteínas pobremente caracterizadas, utilizando un solo método de predicción. Adicionalmente, el predictor unificado se puede usar para asignar una medida de confianza a las interacciones procedentes de los enfoques experimentales a gran escala. Sin embargo, esta propuesta no es válida para predecir AFPP por fuentes de datos independientes, principalmente por insuficiencia de datos en algunas de ellas.

El sistema ERR sirve para extender rutas de Reactome en humanos, proponiendo una lista de proteínas para diversas rutas sobre las que resta una verificación experimental. No obstante, sí se confirma su utilidad biológica con las anotaciones de UniProt y las extraídas de la literatura científica. Por ejemplo, sobre 2 proteínas (PIAS4_HUMAN y DTX1_HUMAN) propuestas para extender la ruta de *Mantenimiento del telómero*, quedando caracterizadas por las anotaciones de UniProt y por la existencia de un dominio *RING-finger* conservado, relacionada en la degradación de la telomerasa, la cual influye en el envejecimiento celular. Además, se pueden usar las reglas de decisión extraídas por el sistema ERR y los predicados frecuentes calculados para conocer las propiedades moleculares de las proteínas que conforman cada ruta.

Adicionalmente, recordando el contexto de la Biología de Sistemas, las predicciones obtenidas para ambos problemas (es decir, las interacciones y asociaciones funcionales puntuales o con un grupo), podrían utilizarse en el análisis de redes complejas biológicas, teniendo en cuenta las características y propiedades descritas de este tipo de redes.

9.2. Reflexión sobre Aplicación de Aprendizaje Automático en Biología

En esta sección se presenta una reflexión sobre la aplicación del Aprendizaje Automático a problemas biológicos. Se trata de una reflexión general, no sólo centrada en los dos problemas elegidos, como la comparación de la sección 9.1.

En el apartado *Estudio/Discusión detallada* de la sección 9.1 ya se discute la relación de los sistemas propuestos en esta tesis con otros métodos existentes, y en el capítulo 10 se exponen los objetivos conseguidos y las aportaciones de la tesis. Por lo tanto, en la presente sección se exponen conclusiones de carácter general y personal, a las que se ha llegado tras el desarrollo de esta tesis y el análisis del proceso llevado a cabo con el AA en Biología.

De forma general, se ha detectado una serie de ventajas que aporta el AA para la anotación en Biología Molecular. Permite desarrollar ideas complejas, como por ejemplo, la integración de varias evidencias, de origen diverso, como características extraídas de la secuencia, interacciones y anotaciones funcionales. Los resultados pueden no ser mejores que un método más sencillo sin aprendizaje, pero éste aporta una predicción alternativa y permite una integración estructurada de datos y representaciones flexibles. Aunque no resuelva por completo el problema biológico, el aprendizaje analiza propuestas de sistemas de predicción.

Desde la perspectiva de la Biología de Sistemas, los procesos o tareas que se llevan a cabo en los organismos a cualquier nivel se deben tratar como un todo, y no analizarlos a través de cada elemento independiente que participa en el sistema, porque todos ellos están relacionados por sus interacciones y asociaciones funcionales. Esta influencia de las relaciones en todos los sistemas biológicos justifica la Representación Relacional que se elige en esta tesis para abordar los problemas de Biología Molecular, en este caso centrado en la anotación funcional. Respecto a la **relevancia de la información relacional** en Biología, hay que decir que las múltiples relaciones que existen entre las moléculas parecía una ventaja muy importante inicialmente para optar por usar Aprendizaje Relacional. Pero estas mismas relaciones (muchas veces indirectas) se convierten frecuentemente en un inconveniente más que en una ventaja: por no sesgar los resultados, por evitar que no se solapen las predicciones de unas clases con las de otras, por no usarlas para un proceso complejo de aprendizaje cuando una simple búsqueda exhaustiva entre las secuencias puede dar la solución, por exigir el uso de varias medidas de evaluación continuamente, etc. Tras el desarrollo de esta tesis, aunque la representación de los datos más adecuada sea de tipo relacional, que conserva la naturaleza estructurada implícita de los datos biológicos con muchas relaciones, esto no justifica que el mejor sistema de aprendizaje sea uno relacional. Antes de aplicar el aprendizaje, hay que valorar qué transformación (parcial, total o nula) conviene realizar sobre dicha representación. Las opciones son mantener la representación relacional y usar AAR (como en la sección 8.4.1), transformar la representación a proposicional y usar AAP (como en el capítulo 6), o mantener inicialmente la representación relacional y usar un híbrido que combine el AAR y AAP (como en el capítulo 7). La elección depende de los datos específicos que se manejen, la cantidad de atributos numéricos y nominales, la diversidad de valores de los mismos, el objetivo buscado (primar la facilidad de interpretación o la precisión), etc. En conclusión, aunque el AAR parezca la mejor opción para una representación relacional, no siempre es así.

Otro punto interesante sobre el que reflexionar en esta tesis es la **generalización de los métodos computacionales** para la aplicación a otros problemas biológicos. Aunque computacionalmente se antoja viable, en términos biológicos carece de sentido definir un marco común de anotación funcional para cualquier contexto en Biología. Por ejemplo,

prácticamente el mismo esquema de AA utilizado para resolver el problema de extensión de rutas sería fácilmente aplicable a la predicción de función de grupos [García-Jiménez et al., 2009, 2010b], concluyendo que las principales dificultades no son computacionales, sino biológicas. En primer lugar, cómo definir los grupos de genes o proteínas con sentido biológico; segundo, la disponibilidad de anotaciones individuales fiables, que permitan expandir este enfoque de anotación de uno a un grupo de secuencias. El planteamiento inicial puede ser general y común, pero finalmente en los detalles de diseño se necesita particularizar en cada entorno de aplicación concreto, porque cada problema biológico tiene muchas peculiaridades dependientes del dominio que necesitan un tratamiento diferente, concreto para ese problema. No obstante, para algunas partes independientes del proceso de anotación funcional sí se pueden definir unas directrices estructuradas de recopilación, representación y aplicación de AA genéricas, para predecir función de genes, proteínas, pares o grupos de ellos.

Primero, respecto a la representación de la información, según sobre qué bases se quiera explicar la anotación, los datos de entrada serán unos u otros: propiedades de la secuencia, interacciones por pares, anotaciones, relaciones de más alto nivel, etc. En este sentido, en esta tesis se consigue una generalización en la representación, con el modelo global de representación del conocimiento propuesto en el capítulo 5, que descompone las anotaciones individuales y las anotaciones de grupo, y es una abstracción común válida para todos los problemas de anotación funcional a resolver con AA. Para cada aplicación concreta, este modelo genérico se puede instanciar fácilmente, según las indicaciones de la sección 5.4. También se pueden compartir datos representados relacionalmente según el modelo, o incluso los ya recopilados en la aplicación del capítulo 7.

A parte de la representación, el resto del proceso de AA requiere un diseño específico para cada problema biológico. Sin embargo, aunque no se pueda aplicar un método computacional predefinido, sí se pueden seguir las estrategias descritas en el capítulo 8 de esta tesis, recopiladas en la tabla 8.11, a partir de pruebas de concepto, que no pretenden definir soluciones absolutas. No obstante, con dichas recomendaciones, según las características y objetivos del problema biológico a resolver, se puede decidir por usar uno de los variados enfoques de AA planteados (multi-clasificador, extracción variada de patrones, reordenación de reglas según distintos criterios, uso de anotaciones de compañeros de interacción o de proteínas principales, representación proposicional o híbrida, etc.).

Con estas propuestas de representación y estrategias genéricas, el mismo conjunto de estrategias con el que se ha afrontado la extensión de rutas basada en representación relacional es reutilizable fácilmente, siempre que se satisfagan o adapten algunos requisitos mínimos del método. Así, se podría aplicar para extender otras bases de datos de rutas, procesos celulares, otro tipo de redes de proteínas o grupos a menor nivel (como los complejos). También se puede aplicar el enfoque para anotar funcionalmente con cualquier otro vocabulario, aunque no implique explícitamente una interacción con los elementos del grupo; por ejemplo, en la extensión de listas de genes sobre-expresados, mutados, con un mismo fenotipo, relacionados con una misma enfermedad, con una misma anotación en un vocabulario o agrupados por alguna otra razón. Se trataría de una anotación funcional a distintos niveles (con más o menos relaciones entre sus elementos).

Por otro lado, la **interpretación biológica** de los resultados puede hacer cambiar por completo el planteamiento, repetidas veces. Muchas soluciones que parecen adecuadas inicialmente, e incluso buenas en términos de rendimiento, no consiguen resolver los problemas desde la perspectiva biológica. Pero este hecho con frecuencia no se descubre hasta que se llega a la interpretación de resultados, la última fase del ciclo de minería de datos, después

de mucho trabajo avanzado en una dirección concreta. Además, los resultados realmente varían significativamente al modificar los conjuntos de datos, no por pequeños cambios en los parámetros de configuración, lo cual obliga a cambios profundos, desde el principio del proceso de aprendizaje.

Relacionado con el punto anterior, a lo largo de la tesis con frecuencia se plantea la discusión de una dicotomía entre las **soluciones preferidas computacionalmente y las preferidas biológicamente**. Es muy importante destacar que en Biología Computacional no se tiende a buscar la solución óptima, ni mejorar mínimamente el rendimiento de un método tras analizar una serie de configuraciones variadas. Sino que se pretende alcanzar una solución válida con una interpretación biológica con sentido, lo cual ya exige una gran dedicación e implica un alto grado de complejidad. En parte, el problema es la validez de los sistemas para datos o problemas cambiantes en el tiempo, lo que hace más necesario la búsqueda de métodos robustos, frente a otros un poco mejores en condiciones muy determinadas. Sin embargo, una solución más robusta que priorice principalmente una evaluación puramente computacional, puede ser muy genérica y no ser útil en términos biológicos; mientras que priorizando una evaluación biológica, la solución puede estar desviada hacia el contexto de los datos de aplicación, pero ser válida y útil para resolver el problema biológico concreto. Por lo tanto, suele resultar más adecuado e interesante abordar un problema diferente entre todos los existentes, que repetir la experimentación para una configuración distinta o un sistema de aprendizaje alternativo, con el objetivo de mejorar ligeramente el rendimiento del sistema. Es decir, se puede concluir que el proceso de aprendizaje debe estar guiado por el dominio, por la biología, los datos y los detalles; no por la metodología ni los intereses computacionales de evaluar uno u otro método con unas características concretas. Los análisis computacionales podrían ser innumerables, pero hay que limitarlos equilibradamente, buscando el rigor y sentido biológico, muy difícil de alcanzar.

Finalmente, se expone una reflexión de carácter más personal sobre la resolución de problemas en Biología Computacional con AA. Trabajar en Biología Molecular desde una perspectiva computacional resulta difícil, como cualquier otra tarea interdisciplinar, pero a la vez es una tarea interesante e intensa, con muchas áreas por investigar. Realmente se necesita más análisis computacional automatizado, mucho más que un simple almacenamiento organizado o una extracción de regularidades sencillas. Por el contrario, la aplicación de técnicas de Inteligencia Artificial a problemas reales en Biología no es trivial, ni generalizable. Además, no es factible aplicar prácticamente ninguna simplificación típica del Aprendizaje Automático, más bien al contrario: en Biología todo tienden a ser excepciones.

Capítulo 10

Conclusiones

La *tesis de esta tesis* se expone desde dos puntos de vista, el computacional y el biológico. Computacionalmente, en los problemas reales, al menos en Biología Molecular, de un ciclo clásico de descubrimiento del conocimiento con Aprendizaje Automático, se requiere dedicar un gran porcentaje de tiempo a la interpretación y análisis de resultados, más que a mejorar los parámetros de configuración del modelo computacional para incrementar unas décimas el rendimiento, que suelen ser irrelevantes en la evaluación biológica. Biológicamente, se puede decir que aún hay muchos análisis por realizar y conocimiento a extraer de todos los datos de relaciones biológicas de los que se dispone, a cualquier nivel. Por lo tanto, queda latente el potencial que ofrece la representación y el aprendizaje relacional combinado con el enfoque de la Biología de Sistemas.

10.1. Repaso de Hipótesis

Si se repasan los objetivos planteados en el capítulo 3, se puede concluir que se han logrado todos:

- Se consigue definir una representación relacional genérica del conocimiento de Biología Molecular para cualquier tarea de anotación funcional en Biología de Sistemas (en el capítulo 5), posteriormente reutilizable y/o adaptable a cada caso, verificándose el objetivo 1.

- Se construyen los cuidados conjuntos de datos (objetivo 2) (descritos en las secciones de 6.2.1 a 6.2.3 y de 7.2.1 a 7.2.4) necesarios para aplicar aprendizaje a los dos problemas biológicos seleccionados.

- Se resuelven dos problemas de anotación funcional desde la perspectiva de la Biología de Sistemas (objetivo 3), como son la predicción de asociaciones funcionales entre pares de proteínas en *E.coli* (en el capítulo 6) y la extensión de rutas biológicas en humanos (en el capítulo 7).

- Como indica el objetivo 4, se utilizan y evalúan distintas representaciones del conocimiento para la predicción de anotación funcional (codificación de operadores para Programación Genética, representaciones relacional y proposicional directa, transformación a proposicional o extracción de patrones con diferentes criterios), cuyos resultados se presentan en las secciones 8.1, 8.3 y 8.4.

- En las secciones de 8.5 a 8.8, se analiza desde distintas perspectivas la relevancia de las relaciones biológicas en el aprendizaje, es decir, la importancia de la Biología de Sistemas en la anotación funcional, dando respuesta al objetivo 5.
- Para satisfacer el objetivo 6, en los apartados *'aplicación a otros problemas'* de todas las secciones del capítulo 8 y resumido en la tabla 8.11, se propone cuándo y cómo aplicar Aprendizaje Automático en otros problemas de anotación funcional.
- Por último, el objetivo 7 también se satisface, usando componentes estándar de la Inteligencia Artificial y la Bioinformática, como: los predictores *ad-hoc* basados sólo en evidencias biológicas (I2H, MT, GC, GF y PP); algoritmos y herramientas clásicas de bioinformática (BLAST, algoritmo de reducción de homología de Hobohm, similitud semántica de Jiang y Conrath, etc); sistema de aprendizaje de árboles de decisión relacionales (TILDE); o WARMR+CLUS como combinación de algoritmos de aprendizaje híbrido (relacional y proposicional) ya utilizada. También se usan repositorios de información biológica, tanto de bajo nivel como preprocesado, para construir o procesar los conjuntos propios de datos. Es decir, se reutilizan componentes estándar existentes y se combinan según los requisitos de cada problema a resolver, sin necesidad de construir para cada tarea independiente de la tesis un sistema computacional propio, con menos potencial que los ya desarrollados y refinados a lo largo del tiempo.

10.2. Contribuciones

Esta sección expone las aportaciones que realiza esta tesis en el ámbito científico y en la comunidad de Aprendizaje Automático.

Las aportaciones en Biología son:

- Los dos conjuntos de predicciones a verificar en el laboratorio, de asociaciones funcionales entre pares de proteínas en *E.coli* y de extensión de rutas en humanos, estando algunas apoyadas por anotaciones de bases de datos de referencia y por la literatura científica, como las proteínas relacionadas con el telómero y las integrinas.
- La interpretación de las predicciones mencionadas, que puede ayudar a comprender mejor el funcionamiento de los sistemas biológicos, principalmente a través de las relaciones.
- Un método para complementar la predicción de asociaciones funcionales entre pares de proteínas que, frente a otros enfoques populares y exitosos como STRING, es adecuado para aportar conocimiento sobre el desconocido y exigente grupo de proteínas poco caracterizadas experimentalmente.
- Una medida integrada de fiabilidad de la predicción automática de asociaciones funcionales entre pares de proteínas en *E.coli*, que evita consultar los resultados de cada método por separado, si no se está interesado en los detalles.
- Un método para asignar un nivel de calidad a la predicción de interacciones proteína-proteína con técnicas de experimentación masiva, las cuales originalmente carecen de medida de confianza.
- Un método para extender rutas biológicas en humanos novedoso e interesante por basarse principalmente en características de la secuencia.

Las contribuciones científicas en el área computacional son:

- Un modelo de datos multi-relacional con el que representar el conocimiento de Biología Molecular para anotación funcional. Este modelo se puede aplicar para representar un subconjunto cualquiera de entidades y relaciones biológicas, y realizar anotación funcional en cualquier otro vocabulario, con Aprendizaje Automático o incluso otra técnica. La anotación funcional se define en un sentido amplio, abarcando también la predicción de pertenencia de una proteína a un grupo.

- Las estrategias de aplicación del AA según las características del problema biológico. Se trata del planteamiento de otras técnicas, representaciones y configuraciones en Aprendizaje Automático para ser aplicado a otros dominios de anotación funcional o también fuera de la Biología. Según los objetivos y características del problema, se sugiere el uso de una estrategia o aproximación concreta. Por ejemplo, usar una representación relacional o proposicional, un multi-clasificador o clasificadores individuales, cómo gestionar los valores desconocidos o qué variante de representación de relaciones multi-valuadas elegir, entre otras.

- La combinación de sistemas computacionales ya existentes, para la predicción de asociaciones funcionales por pares.

- Una medida de probabilidad para reordenar reglas de decisión, priorizando un criterio adicional, como es la diversidad molecular (en la figura 7.5).

- La búsqueda de las estrategias, algoritmos, técnicas de evaluación, criterios de poda, método de evasión de homología, etc. adecuadas para resolver los dos problemas afrontados.

Las contribuciones tangibles a la comunidad de Aprendizaje Automático son:

- Los dos conjuntos de datos para el aprendizaje. Los conjuntos de datos biológicos no estaban definidos a priori, por lo que ha sido necesario construirlos e ir definiendo los límites para refinarlos a lo largo de la tesis. Ambos conjuntos se pueden utilizar para evaluar otros algoritmos, de forma genérica o con un propósito específico, como la gestión de clases des-balanceadas. Por separado, el conjunto para la predicción de asociaciones funcionales entre pares de proteínas en *E.coli* es útil para evaluar la gestión eficiente de valores desconocidos. Por su parte, el conjunto de datos de rutas biológicas es un conjunto no redundante para tareas de clasificación multi-clase y multi-etiqueta, con pocas instancias y un gran desbalanceo de distribución entre clases. Este último conjunto es reutilizable más allá de una simple evaluación basada en entrenamiento y test, porque la información que contiene es más rica, y estructurada en módulos relacionales, de forma que se puede simplificar o ampliar fácilmente para su uso en aprendizaje relacional, o incluso proposicional.

- Las diferentes opciones de extracción de patrones frecuentes en el proceso de transformación de representación relacional a proposicional. En trabajos previos con aplicaciones biológicas [Clare et al., 2006; Vens et al., 2008] siempre se han extraído los patrones frecuentes en todas las clases a la vez. Sin embargo, en esta tesis se propone la posibilidad de extraer los patrones frecuentes en cada clase por separado, pudiendo luego juntarlos con los del resto de clases para realizar un aprendizaje proposicional conjunto, o llevar a cabo diferentes procesos de aprendizaje de forma individual, fomentando la diversidad.

Capítulo 11

Líneas Futuras

Existen muchas investigaciones todavía pendientes en anotación funcional en Biología Molecular. Porque es un área muy amplia, no existe ningún método genérico, ni aproximaciones particulares que resuelvan todos los problemas de anotación funcional existentes, ni siquiera en esta tesis se abarca el uso de la Biología de Sistemas en todos los sentidos posibles. Por lo tanto, se plantean algunos de los múltiples trabajos futuros que se podrían realizar en el mismo ámbito de estudio de esta tesis doctoral.

A corto o medio plazo se plantean las siguientes propuestas:

- Profundizar en el análisis concreto de las proteínas predichas, su relevancia biológica y su similitud en propiedades de secuencia.

- Ampliar el estudio de otros enfoques de diseño en la aplicación de Aprendizaje Automático a otros dominios (presentado en el capítulo 8 de esta tesis). Se podrían analizar las opciones en la búsqueda de reglas de clasificación diversas, la optimización de la poda del árbol de clasificación, la evaluación basada en similitud semántica en conjuntos de anotaciones en varios niveles, o las opciones de establecimiento del umbral de predicción.

- Usar otras representaciones híbridas más sencillas, aparte de las conjunciones de predicados relacionales como atributos proposicionales binarios. Por ejemplo, se podrían usar sólo los predicados simples, sin construir combinaciones complejas, y evaluar si existe pérdida cuantitativa o interpretativa en los resultados.

- Aplicar Aprendizaje Relacional directo al problema de predicción de asociaciones funcionales entre pares de proteínas, y compararlo con la solución proposicional presentada.

- Actualizar el predictor de asociaciones funcionales entre pares de proteínas, utilizando nuevos conjuntos de datos experimentales a pequeña escala (más numerosos y fiables), que han aparecido tras la construcción del sistema presentado en la tesis.

- Usar nuevas versiones del clasificador bayesiano elegido (AODE) para predecir asociaciones funcionales entre pares de proteínas, evaluando si el enfoque de ampliar ligeramente la dependencia entre atributos puede mejora los resultados alcanzados.

Como trabajos a largo plazo, se proponen algunas líneas futuras:

- Incluir en el proceso de aprendizaje propiedades topológicas de las redes biológicas y otro conocimiento derivado de la teoría de grafos. Analizar en el contexto de las redes las predicciones de interacciones y asociaciones funcionales obtenidas.
- Usar agregados como fuente adicional de información relacional, que recopilen características globales de un conjunto de proteínas relacionadas por algún criterio. Aunque los patrones frecuentes se pueden considerar agregados, otra alternativa sería, por ejemplo, incluir la moda de anotación en familias de dominios de las proteínas pertenecientes a una ruta biológica.
- Reutilizar un subconjunto de los datos relacionales recopilados para la predicción de rutas biológicas, y usarlos para la anotación funcional de proteínas humanas en otro vocabulario de anotación, como por ejemplo la implicación en enfermedades.
- Aplicar el modelo de representación de datos biológicos a un problema en Biología de Sistemas, con nuevas relaciones (procedentes de microarrays o tecnologías de secuenciación de nueva generación, etc.), y diseñar su resolución siguiendo las sugerencias expuestas en el capítulo 8 de esta tesis.

Apéndice A

Publicaciones

A continuación se presenta una lista de las publicaciones que están ligadas al desarrollo de esta tesis.

Título:	Sequence Features and Interactions for Relational Learning-based Human Reactome Pathways Extension
Autores:	Beatriz García-Jiménez, Tirso Pons, Araceli Sanchis y Alfonso Valencia
Publicación:	Proceedings of the 11th European Conference on Computational Biology (ECCB), issue of the Bioinformatics journal
Fecha:	*(en revisión, Abril 2012)*

Título:	Relational Learning-based Extension for Reactome Pathways with Sequence Features and Interactions
Autores:	Beatriz García-Jiménez, Tirso Pons, Araceli Sanchis y Alfonso Valencia
Publicación:	Proceedings of the 11th Spanish Symposium on Bioinformatics (JBI)
Fecha:	Enero 2012

Título:	MMRF for Proteome Annotation Applied to Human Protein Disease Prediction
Autores:	Beatriz García-Jiménez, Agapito Ledezma y Araceli Sanchis
Publicación:	Proceedings of the 20th International Conference on Inductive Logic Programming (ILP)
Fecha:	Junio 2010

Título:	S.cerevisiae Complex Function Prediction with Modular Multi-Relational Framework
Autores:	Beatriz García-Jiménez, Agapito Ledezma y Araceli Sanchis
Publicación:	Proceedings of the 23rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE)
Fecha:	Junio 2010

Título:	Inference of Functional Relations in Predicted Protein Networks with a Machine Learning Approach
Autores:	Beatriz García-Jiménez, David Juan, Iakes Ezkurdia, Eduardo Andres-León y Alfonso Valencia
Publicación:	PLoS ONE 5(4): e9969
Fecha:	Abril 2010

Título:	Modular Multi-Relational Framework for Gene Group Function Prediction
Autores:	Beatriz García-Jiménez, Agapito Ledezma y Araceli Sanchis
Publicación:	Poster. 19th International Conference on Inductive Logic Programming (ILP)
Fecha:	Julio 2009

Título:	EcID. A database for the inference of functional interactions in E.coli
Autores:	Eduardo Andres León, Iakes Ezkurdia, Beatriz García-Jiménez, Alfonso Valencia y David Juan
Publicación:	Nucleic Acids Research, Vol. 37, D629-D635
Fecha:	Enero 2009

Título:	Genetic Programming for Predicting Protein Networks
Autores:	Beatriz García-Jiménez, Ricardo Aler, Agapito Ledezma y Araceli Sanchis
Publicación:	Proceedings of the 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA)
Fecha:	Octubre 2008

Título:	Protein-Protein Functional Association Prediction Using Genetic Programming
Autores:	Beatriz García-Jiménez, Ricardo Aler, Agapito Ledezma y Araceli Sanchis
Publicación:	Proceedings of the International Conference on Genetic and Evolutionary Computation (GECCO)
Fecha:	Julio 2008

Apéndice B

Anotación Funcional del Genoma y Proteoma

En los siguientes apartados se presenta de forma breve qué es anotar funcionalmente, qué tipos de anotaciones existen, y cómo anotar (métodos).

B.1. Definición de Anotación

La anotación funcional consiste en asignar información biológica a secuencias de genes y de productos genéticos, principalmente proteínas y ARN [Stein, 2001]. Se puede hablar indistintamente de la anotación de genoma, de proteoma o de otro producto genético.

La secuenciación ahora es rápida y barata, generando el mapa genómico de muchas especies, pero sin saber cuál es la función de cada gen y proteína. En algunos casos también se conoce la estructura, pero no se sabe la función. La distancia entre el conocimiento del genoma y la anotación del mismo está creciendo en gran medida [Rost et al., 2003; Friedberg, 2006; Hawkins and Kihara, 2007]. Se necesita saber en qué tarea participa cada gen y proteína, cuáles son los implicados en cada tarea del metabolismo, para conocer dónde actuar en caso de mal-función o enfermedad. La información biológica a asignar puede ser muy diversa, según se expone en la sección 2.4.1.

B.2. Vocabularios de Anotación

Un vocabulario de anotación es un conjunto de términos definidos para describir las funciones de los genes y productos genéticos dentro de un mismo ámbito.

Los vocabularios se dividen en distintas categorías, que no coinciden necesariamente con los distintos niveles de función (descritos en la sección anterior), tendiendo a existir más de una categoría por cada nivel. Algunos ejemplos de categorías de anotación son: función molecular, localización celular, dominio de proteína, asociación con una enfermedad, participación en una ruta metabólica, etc.

Algunos vocabularios son simples listas planas de términos, otros están organizados en niveles, otros presentan alguna relación padre-hijo o de inclusión entre términos, y otros incluso están distribuidos en jerarquías. Existen muchos catálogos diferentes para la misma o diferentes categorías [Ouzounis et al., 2003]. Unos están asociados a genes y proteínas individuales, otros a conjuntos de ellos, y otros a ambos. En general, para cualquier vocabulario de anotación

existen muchas bases de datos biológicas, específicas y también solapadas entre ellas [Galperin and Fernández-Suárez, 2011].

No siendo el objetivo de este anexo presentar una lista exhaustiva, debido a la gran cantidad de vocabularios que existen, a continuación se presentan sólo las categorías de anotación a las que se hace referencia en este documento y los vocabularios correspondientes más relacionados:

- **Asociaciones funcionales entre pares de proteínas y rutas biológicas**: Estos dos tipos de relaciones también se pueden considerar anotaciones funcionales, a la vez que usarlas como entrada para predecir otras categorías de anotación. Estas categorías de anotación ya se han descrito en detalle, en la sección 2.4.3, junto con sus correspondientes vocabularios.

- **Función molecular**: Las funciones moleculares de un gen o producto genético son las tareas que hace o las habilidades que tiene. Por ejemplo, transportar, enlazar o modificar alguna molécula. El primer vocabulario de función molecular desarrollado fue *Enzyme (EC)* [Bairoch, 2000], diseñado para describir sólo la actividad enzimática. *FunCat* (*MIPS Functional Catalogue*) [Ruepp et al., 2004] extiende la idea a más proteínas y funciones a través de su catálogo de clasificación. *Gene Ontology* (GO) [Ashburner et al., 2000] es otro vocabulario de función, siendo el más usado y conocido en Biología Molecular, aunque no sea completo. Está estructurado en forma de ontología, como grafo acíclico. GO también incluye otros dos vocabularios, en ontologías separadas, sobre procesos biológicos (asociados a grupos) y localización celular (asociada a elementos individuales). Además, existen versiones reducidas (GOSlim), una genérica y otras específicas para algunas especies, incluyendo menos términos pero más generales. Estos vocabularios simplificados pueden ser útiles en tareas de clasificación con Aprendizaje Automático, técnicas para las cuales es muy difícil predecir con cientos o miles de clases diferentes.

- **Familias y dominios**: Las familias incluyen diferentes proteínas relacionadas evolutivamente, y pueden ser agrupaciones de dominios de proteína típicos en la naturaleza, que proporcionan alguna evidencia sobre la función con la que están asociadas. Por lo tanto, las familias y dominios también se consideran anotaciones funcionales. Estos patrones de secuencia se extraen de alineamientos de secuencia múltiples de segmentos, patrones de expresiones regulares o modelos ocultos de Markov (del inglés, *Hidden Markov Models, HMMs*) [Yoon, 2009]. Los resultados se almacenan en bases de datos o vocabularios de categorías de dominios y familias con distintos enfoques, como *PRINTS* [Attwood, 2002], *PROSITE* [Sigrist et al., 2010] o *Pfam* [Finn et al., 2010]. *InterPro* [Hunter et al., 2009] es un recurso integrado de familias de proteínas, dominios y sitios funcionales, que combina información de varias de las anteriores, siendo la base de datos de dominios más exhaustiva y potente. Pfam e InterPro son dos de los vocabularios con mayor cobertura de anotación del proteoma.

A pesar de la larga lista de vocabularios de anotación existente, también se podría desarrollar un método de predicción de anotación funcional sobre un vocabulario nuevo, creado para unas necesidades específicas no cubiertas con ninguno previo.

B.3. Metodologías de Anotación

Debido a la complejidad y amplitud de las funciones de genes y proteínas, la asignación de función a un producto genético no caracterizado se podría enfocar desde distintas direcciones.

Dentro de los métodos de anotación, en primer lugar, hay que diferenciar entre la anotación experimental y la predicción de anotación computacional. Las técnicas experimentales de anotación son costosas en recursos y tiempo invertido. Requieren la dedicación de expertos científicos experimentalistas para el diseño, preparación de muestras, ejecución de pruebas y análisis, utilizando químicos y materiales biológicos en un laboratorio clásico; frente a los laboratorios informáticos o de ordenadores, compuestos únicamente de elementos de silicio, menos costosos y de ejecución mucho más veloz. Las técnicas experimentales proporcionan una anotación real del gen o producto genético considerado, mientras que las técnicas computacionales sólo pueden dar una predicción, con mayor o menor fiabilidad, pero que no se puede considerar como una anotación verificada. No obstante, las técnicas computacionales se presentan como una alternativa útil en los últimos años [Pavlidis et al., 2002], debido a las limitaciones ya comentadas en las técnicas experimentales. Así, las técnicas computacionales de predicción de anotación proporcionan una serie de anotaciones predichas, las anotaciones más probables, que serán las prioritarias para ser verificadas experimentalmente en los laboratorios clásicos [Peña-Castillo et al., 2008].

Para afrontar la anotación funcional de genomas y proteomas existen multitud de métodos computacionales. Dependiendo del organismo, el nivel de función definido, el vocabulario, la evidencia utilizada, etc. se pueden utilizar muchos métodos *ad-hoc*, restringidos a un subconjunto de elementos biológicos, al igual que en otras áreas bioinformáticas. Cada método se aplica con unas determinadas características y sobre una determinada especie, cuyo planteamiento específico tendría que modificarse en gran medida si se cambia alguna de las limitaciones o consideraciones.

A continuación se presenta una posible clasificación de los distintos enfoques de predicción de anotación funcional considerados en Biología Computacional [Rost et al., 2003; Friedberg, 2006; Hawkins and Kihara, 2007]. Esta clasificación se basa en las evidencias funcionales (o información asociada al producto genético) que utilizan los métodos. Los vocabularios de anotación (sección B.2) son también un subconjunto de los posibles tipos de información asociada a un producto genético. No obstante, a la hora de hacer predicción funcional, se necesita ortogonalidad entre los atributos de entrada (evidencias funcionales) y el objetivo de predicción (vocabulario de anotación), para no sesgar la predicción con datos comunes en la entrada y la salida. Muchas veces hay relaciones indirectas y desconocidas (procedentes de la homología) que hay que evitar para no sesgar el método (ver un análisis más detallado en el capítulo 8).

B.3.1. Predicción basada en Similitud de Secuencia *(u Homología)*

La técnica básica para anotar nuevos genomas en biología computacional es el método comparativo. La genómica comparativa [Hardison, 2003] pretende descubrir las funciones de un gen por comparación de múltiples secuencias genómicas [Bandyopadhyay et al., 2007], basándose en el principio evolutivo que afirma que las funciones codificadas en el genoma se conservan a través de las especies.

Los principios fundamentales de la genómica comparativa son sencillos [Hardison, 2003]. Las características comunes entre dos especies se codifican en los fragmentos de ADN conservados. Más específicamente, las secuencias de ADN que codifican las proteínas

responsables de las funciones conservadas desde el último ancestro común de las especies, deben estar presentes en las secuencias genómicas actuales. Igualmente, se deben conservar las secuencias de ADN que controlan la expresión de los genes regulados de forma similar en dos especies relacionadas. Por el contrario, las secuencias que codifican (o controlan la expresión de) las proteínas responsables de las diferencias entre especies deben ser divergentes.

Aplicar la genómica comparativa a la predicción de anotación funcional consiste en comparar la secuencia de un nuevo gen o proteína con el resto de secuencias ya anotadas, de una misma o diferente especie, para inferir la nueva anotación a partir de las secuencias anotadas que tengan un mayor grado de similitud. Es decir, utilizar el conocimiento previo de anotaciones de secuencias completas o fragmentos, para determinar la función biológica de una secuencia desconocida.

El concepto de *homología*, desde un punto de vista evolutivo, indica que una misma característica presente en dos especies diferentes procede de un ancestro común [Lankester, 1870]. En biología molecular evolutiva se aplica el término homología para identificar la similitud o grado de identidad entre secuencias, de nucleótidos o aminoácidos. El grado de identidad permite presuponer un origen evolutivo común de las secuencias que se comparan [Saladrigas, 2006]. Se dice que dos secuencias son homólogas cuando tienen un porcentaje de identidad de secuencia elevado, en una porción relevante de la secuencia. La homología se puede separar a su vez en los conceptos de paralogía y ortología, dependiendo si las secuencias similares comparadas son de la misma o de distinta especie, respectivamente. Así, comparten función por un proceso de duplicación de la secuencia dentro del mismo organismo de una especie (parálogos) o por un proceso de especiación a partir de un ancestro común (ortólogos) a lo largo del proceso evolutivo. No obstante, no hay que confundir la identidad (absoluta) de secuencia con la similitud (relativa) de secuencia. Esta última permite más flexibilidad en la comparación y establece diferentes grados en la sustitución de un nucleótido o aminoácido por otro en la misma posición de la secuencia.

Así, el enfoque de predicción basado en similitud de secuencia consiste en buscar la secuencia más similar a una proteína dada desconocida, a la que se le asigna la anotación de la proteína similar encontrada. El algoritmo o método básico de búsqueda de secuencias similares por pares es el famoso y eficiente BLAST (del inglés, *Basic Local Alignment Search Tool*) [Altschul et al., 1990], que encuentra regiones de similitud entre miles de secuencias, tanto de nucleótidos como de aminoácidos, calculando la relevancia estadística de los emparejamientos. Existen múltiples especializaciones de BLAST [Ye et al., 2006] y métodos que añaden información adicional a la homología. Otro método precursor muy conocido es FASTA [Pearson and Lipman, 1988]. No obstante, para anotación funcional en realidad la técnica más utilizada es la evolución de las búsquedas por pares, es decir, los alineamientos múltiples de secuencia (del inglés, *Multiple Sequence Alignment, MSA*), que contienen información evolutiva que las comparaciones por pares no tienen. Existen múltiples algoritmos, enfoques y mejoras progresivas para calcular dichos alineamientos [Do and Katoh, 2008].

El enfoque basado en similitud es el más extendido para anotar funcionalmente. Pero con el paso del tiempo, la similitud no es suficiente para asignar anotación, porque empieza a anotar menos secuencias y muchas veces amplifica errores existentes. Porque si las proteínas buscadas tienen una mala anotación (por una asignación equivocada), la predicción por similitud también será errónea, propagando el error. Con el crecimiento exponencial de secuencias, también crece su diversidad, y las secuencias conocidas anotadas no sirven porque no se parecen lo suficiente a las nuevas [Friedberg, 2006]. Por lo tanto, surgen los enfoques descritos en los siguientes

apartados.

B.3.2. Predicción basada en Similitud Estructural

Se puede extender el concepto de similitud de secuencia a similitud estructural, comparando estructuras tridimensionales de proteínas (si están disponibles) en vez de secuencias lineales, pero conservando el núcleo de significación de similitud por evolución común. La estructura se conserva más que la secuencia, por lo que proteínas con poca o ninguna similitud de secuencia pueden tener similitud de estructura, lo que permite aplicar este enfoque. Estos métodos confían en el contenido de PDB (del inglés, *Protein Data Bank*) [Berman et al., 2000], donde buscan las estructuras a comparar.

B.3.3. Predicción basada en Patrones de Secuencia o Estructura

Muchas veces las funciones no dependen de la secuencia o estructura completa, sino de una pequeña región altamente conservada. Por lo tanto, en muchos casos, basta con identificar un patrón característico en la secuencia o en la estructura, que es la que se asocia con la función. En distintas proteínas existen tipos de dominios similares, que han seguido un proceso evolutivo común, que se suelen agrupar en familias. Así, se trata de buscar dominios y familias comunes para anotar una proteína. Se pueden consultar más detalles sobre estas aproximaciones en [Friedberg, 2006; Hawkins and Kihara, 2007].

B.3.4. Predicción basada en Asociación o Contexto Genómico

La filogenómica indica que hay que tener en cuenta la historia evolutiva de los posibles homólogos cuando se usa la similitud para asignar función. En la práctica, significa transferir la anotación del ortólogo más cercano, no de la secuencia más similar. Se suelen denominar métodos basados en filogenética, y se deberían aplicar cuando BLAST devuelve más de una anotación, discriminando la correcta con el uso de la historia evolutiva de los homólogos en cuestión. Son métodos que usan la secuencia, pero no usan la transferencia de función por similitud directamente.

La organización del gen es una fuente de evidencias funcionales, dentro del organismo y entre especies. Se pueden diferenciar tres asociaciones genómicas diferentes relacionadas con la función: la similitud de perfiles filogenéticos, la proximidad cromosómica y los eventos de fusión de genes o dominios. Una de estas tres evidencias o varias se utilizan en los métodos de predicción de función, haciendo corresponder los perfiles de proteínas desconocidas con los perfiles de las que ya están anotadas.

- El perfil filogenético de un gen es un vector booleano, con un 1 si tiene homólogo en un genoma (especie) dado y un 0 si no lo tiene. Así, se considera que dos proteínas con el mismo, o muy cercano perfil han evolucionado juntas, y probablemente estén asociadas funcionalmente [Pellegrini et al., 1999].

- En genomas procariotas existen más evidencias semejantes que se suelen utilizar en predicción basada en contexto genómico, porque los genes asociados funcionalmente suelen estar cercanos en el cromosoma único, para facilitar una transcripción común [Dandekar et al., 1998].

- Si dos genes pueden fusionarse dentro de un mismo gen, aunque como dos dominios diferentes, también se muestra una clara asociación funcional [Enright et al., 1999; Marcotte et al., 1999].

B.3.5. Predicción basada en Redes de Interacción

En estos métodos, a partir de un conjunto de interacciones y asociaciones funcionales experimentales o predichas, inicialmente se construye la red, y después se necesita realizar un análisis para extraer la función de los elementos de la red, para asignar función a las proteínas desconocidas de la misma, porque es razonable que compartan función con sus parejas de interacción o asociación funcional.

Los principales enfoques para este análisis [Sharan et al., 2007] son: 1) la asignación de la función de los vecinos más cercanos de la red y 2) la agrupación de elementos de la red a los que se asigna una función común. Como ejemplo avanzado del primer enfoque [Chua et al., 2006], se puede explorar toda la red, en lugar de consultar sólo las funciones de las dos o tres proteínas más próximas, y asignar diferentes pesos en función de la distancia a las proteínas anotadas y la frecuencia de la función en la red. Como ejemplo del segundo, *Prodistin* [Brun et al., 2003] realiza una agrupación jerárquica basada en una medida de distancia más elaborada que el camino más corto entre pares de proteínas, aplicándolo a varias especies, incluida la humana. Además está disponible en una aplicación web [Baudot et al., 2006].

Además, se ha demostrado que las interacciones detectadas por técnicas experimentales masivas contienen muchos falsos positivos [von Mering et al., 2002]. A la hora de predecir función, estas interacciones (y la predicción de función que derive de ellas) deberían considerarse de menos calidad, cuando generalmente se tratan todas igual.

B.3.6. Predicción basada en Co-expresión

Los microarrays examinan los patrones de expresión de cientos o miles de genes a la vez. Desde su desarrollo en 1995 [Schena et al., 1995], esta tecnología se ha convertido en un método experimental estándar en un gran rango de ámbitos de investigación biológica, empezando actualmente a sustituirse por la secuenciación de nueva generación (del inglés, *Next Generation Sequencing, NGS*) [Mardis, 2011].

De los datos de intensidad de la expresión de genes se pueden extraer grupos estadísticamente significativos que probablemente están involucrados en un proceso biológico coordinado [Chen, 2007]. Así, genes de función desconocida, que se co-expresan con genes conocidos, pueden ser anotados con la función del segundo por esta asociación. Este método es útil para predecir función de proceso biológico o celular, no función molecular o bioquímica.

B.3.7. Predicción basada en Minería de Textos

Estos métodos tratan de extraer conocimiento del análisis automático de la literatura científica biomédica, mediante técnicas de procesamiento del lenguaje natural u otras diferentes [Krallinger et al., 2010]. Con esta base, por ejemplo, cuando los identificadores de genes o proteínas aparecen verbalmente asociados en un texto, en diferentes grados, se puede realizar la transferencia de anotación entre ambos [Hoffmann and Valencia, 2004].

B.3.8. Predicción basada en Propiedades Extraídas de la Secuencia

En este caso, la entrada se limita a la secuencia de nucleótidos o aminoácidos, sobre la que se aplican cálculos sencillos y algoritmos simples, para obtener diferentes atributos. Se incluyen desde parámetros físicos y químicos (peso molecular, longitud de la secuencia, composición de nucleótidos o aminoácidos, carga eléctrica, etc.), similares a los extraídos de la herramienta *ProtParam* [Gasteiger et al., 2005], hasta modificaciones post-traducción y localización subcelular [Jensen et al., 2003b; Lee et al., 2007; Juncker et al., 2009]. Así, algunos enfoques de Aprendizaje Automático reducen la entrada a atributos numéricos [Jensen, 2002] aplicando redes de neuronas [Jensen et al., 2003a] o máquinas de vector de soporte [Lee et al., 2009], técnicas sub-simbólicas con salida poco interpretable, pero muy usadas en Biología Computacional.

La predicción de función basada **sólo** en características de la secuencia es la opción más restrictiva (en datos disponibles y condiciones a comprobar) dentro de la anotación funcional, y por tanto la más compleja y de las menos afrontadas. Es por ello que no hay que confundir métodos basados en características (procedentes sólo de la secuencia o no) con métodos en ausencia de homología.

Para que la predicción de función basada en propiedades de la secuencia pueda ser aplicable en ausencia de homología, no debería contener ninguna información de similitud con otros genes o proteínas, por homología directa o indirecta. En la mayoría de métodos híbridos, aunque se trate de predicción basada en características, se incluyen atributos calculados a partir de algún tipo de relación de homología (anotaciones de dominios, predicción de estructura secundaria usando datos de similitud de secuencia, etc). Al eliminar estas anotaciones, que muchas veces proceden de experimentos en otra especie y que se extrapolan automáticamente por homología a las bases de datos de anotación del resto de especies, se restringe notablemente la cantidad de información a usar en la predicción. Usar información procedente de esta homología indirecta (por anotaciones) mejora notablemente la predicción. Sin embargo, la predicción no sería necesaria si a partir de dicha anotación se puede deducir la clase directamente.

B.3.9. Métodos Híbridos

Estos métodos utilizan diferentes enfoques para combinar e integrar muchos tipos de fuentes de datos diferentes, tales como características de la secuencia, predicción de estructura secundaria y estructura terciaria, anotaciones de dominios y familias, rutas metabólicas, patrones de expresión, redes de interacción, etc. Algunos ejemplos son [Al-Shahrour et al., 2006] (*Babelomics*), [Clare et al., 2006], [Tetko et al., 2008] y [Linghu et al., 2009]. Cabe mencionar también el sistema *DAVID* [Dennis et al., 2003] en este apartado, porque usa como entrada varias fuentes de anotaciones; aunque no es estrictamente un método de predicción, sino que calcula las anotaciones enriquecidas en una lista de genes no caracterizada, comparada con un conjunto de genes ya anotados.

Otros aspectos de la anotación funcional

El resto de vocabularios distintos del objetivo de anotación funcional, se pueden utilizar como información de contexto para predecir la salida. Por ejemplo, si el objetivo es anotar el genoma de una especie con las enfermedades en las que puede estar involucrado cada uno de sus genes, como datos de entrada se pueden utilizar el resto de categorías de anotación de las que se tenga información disponible: rutas metabólicas y redes de interacción a las

que pertenece, tipos de dominios y familias de proteínas en las que se clasifica, funciones moleculares, localización celular, perfiles de expresión génica, etc. No obstante, hay que tener en cuenta que cualquier combinación de atributos de entrada no es válida. La información de contexto y el objetivo de predicción deben ser ortogonales. Es decir, dado que en biología todos los elementos están muy relacionados entre sí, hay que usarlos cuidadosamente, para evitar predecir algo evidente, como puede ser predecir una categoría de anotación en función de otras con las que mantiene una relación de implicación directa o indirecta. Por ejemplo, no sería útil predecir a qué complejo de proteínas pertenece una nueva proteína dada, utilizando como conocimiento de contexto las interacciones proteína-proteína, ya que los complejos se construyen a partir de dichos pares.

También hay que tener en cuenta que para distintas especies no hay la misma información disponible en las bases de datos. Puede haber menos, más, o tener que buscar en una base de datos especializada en cada especie de forma independiente (por ejemplo, *EcoCyc* [Keseler et al., 2005] para *E.coli*, *SGD* (del inglés, *Saccharomyces Genome Database*) [Cherry et al., 1998] para *Saccharomyces cerevisiae*, o *FlyBase* [Tweedie et al., 2009] para *Drosophila melanogaster*); o incluso especializada en un tipo de datos concreto en una especie, como *HPID* (del inglés, *Human Protein Interaction Database*) [Han et al., 2004] que contiene sólo datos sobre asociaciones entre proteínas humanas.

Toda esta información asociada a los productos genéticos está almacenada en múltiples bases de datos diferentes y distribuidas en toda la web. A principios de cada año la revista *Nucleic Acids Research* presenta una revisión de todas las bases de datos disponibles para biología molecular [Galperin and Fernández-Suárez, 2011]. La mayoría son de acceso público, aunque existen algunas que pueden requerir un registro previo para búsquedas más detalladas (por ejemplo, *STRING* [Jensen et al., 2009]). Los formatos de acceso a los datos son variados, sobresaliendo la interfaz web como el mayoritario, aunque también se pueden encontrar simples ficheros de texto plano o tabulados.

Aunque en general se accede directamente a cada base de datos en la que se esté interesado, para obtener los datos de forma más fiable y actualizada, también existen algunos sistemas web que integran varias fuentes. Por ejemplo, *BioMart* [Smedley et al., 2009] que integra los datos de los genes y proteínas de una de las principales fuentes de secuencias como es *Ensembl* [Flicek et al., 2010], o los recursos bioinformáticos *DAVID* [Dennis et al., 2003], que permiten obtener datos variados de un grupo de genes o de uno concreto. Estos sistemas facilitan la extracción de información, evitando tener que ir a las bases de datos independientes para cada tipo de información. Aunque si las condiciones en las cuales la herramienta recopila, almacena y presenta la información no coinciden con el planteamiento del problema, no es adecuado utilizarla para alcanzar el objetivo buscado.

B.4. Métodos de Determinación de Interacción o Asociaciones Funcionales por Pares

La forma tradicional de detectar interacciones sigue procedimientos experimentales ejecutados en laboratorios biológicos. Puesto que la aplicación de estos métodos es costosa en tiempo y recursos, en los últimos años ha surgido un interés creciente en el uso de métodos computacionales de predicción, que consiguen reducir dichos costes, permitiendo priorizar las interacciones más probables.

En esta sección se revisan los métodos de anotación específicos para interacciones y asociaciones funcionales de proteínas por pares, tanto experimentales como computacionales.

B.4.1. Métodos Experimentales

Aparte de los experimentos a pequeña escala, diseñados individualmente para identificar un conjunto pequeño de interacciones concretas, existen dos técnicas básicas experimentales de determinación de interacciones que se pueden aplicar de forma masiva [Causier, 2004]: el método de los dos híbridos en levadura (del inglés, *yeast two-hybrid*) y la co-precipitación y espectrometría de masas.

El método de los dos híbridos [Fields and kyu Song, 1989; Ito et al., 2001] utiliza la premisa descubierta por Fields y Song en levadura, consistente en que los factores de transcripción en eucariotas son modulares. Es decir, que la transcripción se produce sin necesidad de que el dominio de activación esté unido físicamente con el dominio de enlace al ADN, sino sólo conectados indirectamente a través de dos proteínas que interaccionan, unidas cada una a uno de los dominios del factor de transcripción de la levadura. El proceso de fusión de dominios para formar el factor de transcripción se produce *in vivo*, dentro del organismo de la levadura. Posteriormente se comprueba si el producto resultante de la transcripción está presente en el organismo para verificar la existencia de la interacción física.

En el método de espectrometría de masas [Gavin et al., 2002; Ho et al., 2002], se sigue un proceso en el que primero se etiquetan las proteínas problema para identificarlas posteriormente, y se deja que formen complejos físicos libremente. A continuación se separan los componentes del complejo por purificación de afinidad, se obtienen péptidos de cada componente, y éstos se identifican mediante las técnicas propias de espectrometría de masas, comparando su masa con los registros en bases de datos.

Ambas técnicas se comparan en la tabla B.1.

Tabla B.1: Comparativa entre métodos experimentales a gran escala.

Método de los híbridos (Y2H)	**Espectrometría de masas (MS)**
Detecta interacciones: - binarias - débiles y transitorias - físicas directas	Detecta interacciones: - en complejos (grandes) - abundantes y estables - físicas indirectas
In vivo	*In vitro*
Menor nº de interacciones	Mayor nº de interacciones (pero indirectas)
Escaso solapamiento entre distintas técnicas	
Muchos falsos positivos y falsos negativos	

B.4.2. Métodos Computacionales

Los métodos computacionales de predicción de interacciones y asociaciones funcionales entre proteínas se suelen clasifican en cinco categorías, según las evidencias que utilizan. En este apartado se describe un representante de cada una de estas cinco categorías, correspondientes a las técnicas que se utilizan en este trabajo [Valencia and Pazos, 2002], representadas de forma gráfica en la Figura B.1.

- **Perfiles filogenéticos (PP, *Phylogenetic Profiles*)**
  Este método se basa en la similitud de perfiles filogenéticos, examinando la presencia o ausencia de los genes en diferentes especies. De forma que si las mismas proteínas

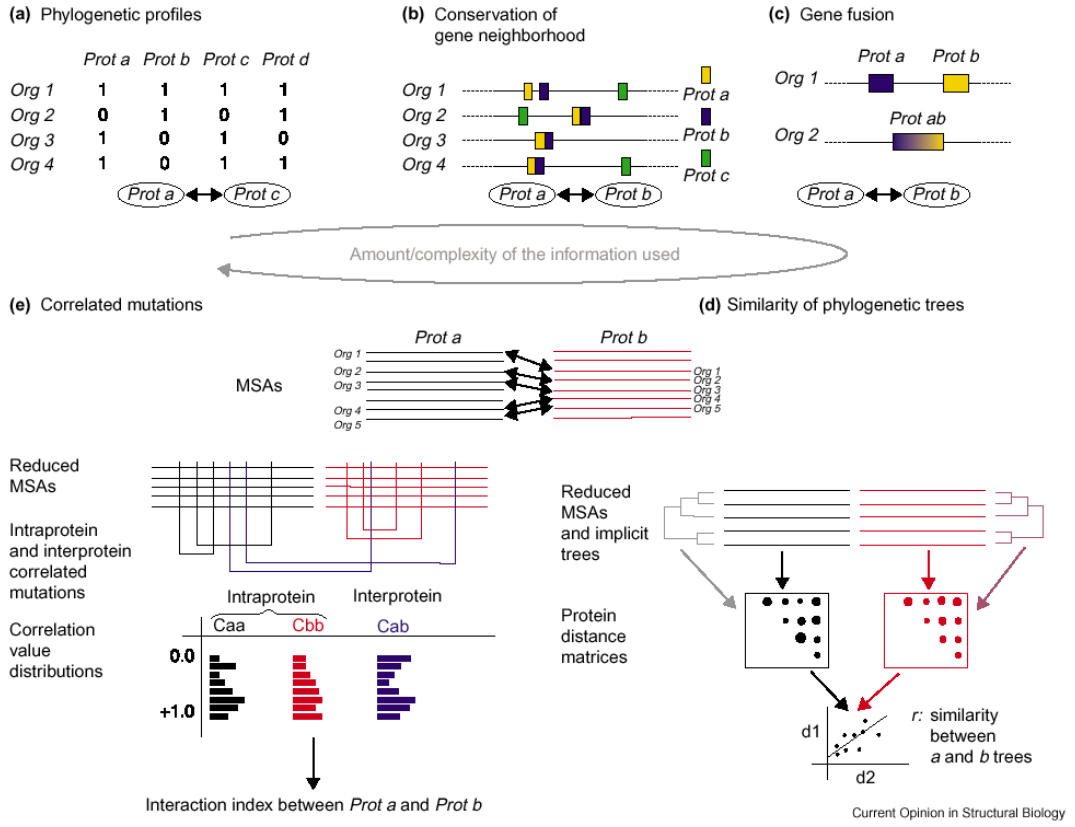


Figura B.1: Representación gráfica de los 5 métodos computacionales de predicción usados. (a) PP: Perfiles filogenéticos, (b) GC: Conservación de genes adyacentes, (c) GF: Eventos de fusión de genes, (d) MT: Similitud de árboles filogenéticos, (e) I2H: Mutaciones correlacionadas. Fuente: [Valencia and Pazos, 2002].

se mantienen constantes a lo largo de un conjunto de especies (igual o muy semejante perfil), esto es indicativo de que ambas son necesarias para realizar alguna función conjunta.

Presenta varios inconvenientes. Por un lado, no asegura la existencia de interacción física, sino que sólo se trata de asociación funcional; y por otro, este método necesita el genoma completo de todos los organismos implicados, para poder saber si un determinado gen falta realmente o no.

Un esquema de este método aparece en la Figura B.1(a), y una descripción detallada en [Pellegrini et al., 1999].

- **Conservación de genes adyacentes (GC, *Gene Context*)**
  Se considera que dos proteínas interaccionan o se asocian funcionalmente cuando sus genes están cercanos en los genomas de varios organismos. Este método se basa en que se conoce que en los genomas bacterianos los genes adyacentes a veces se expresan a la vez, y cuando estas relaciones de vecindad se conservan en diferentes especies, pueden dar lugar a proteínas con una misma función.

  El problema de este método es que sólo puede ser aplicado a organismos procariotas,

que es donde se cumple dicha propiedad.

Un esquema de este método aparece en la Figura B.1(b), y una descripción detallada en [Dandekar et al., 1998].

- **Eventos de fusión de genes (GF, *Gene Fusion*)**
  En este caso, se dice que dos proteínas de un organismo dado interaccionan o se asocian funcionalmente si dichas proteínas forman parte de una sola proteína en otra especie. Se trata de buscar los mismos dominios de proteína en distintas especies, de forma que en una aparezcan los dominios en distintas proteínas, y en otra especie aparezcan los dos dominios fusionados dentro de la misma proteína. Para ello se necesita buscar previamente secuencias semejantes en múltiples especies, utilizando técnicas de alineamiento múltiple de secuencia.

  La desventaja de este método es que los eventos de fusión de dominios no ocurren con frecuencia.

  Un esquema de este método aparece en la Figura B.1(c), y una descripción detallada en [Enright et al., 1999; Marcotte et al., 1999].

- **Similitud de árboles filogenéticos (MT, *MirrorTree*)**
  Este método (MT) y el siguiente (I2H) se basan en la co-evolución de proteínas para determinar la interacción o asociación funcional. Incluso el método PP ya descrito, se podría decir que se basa en una co-evolución *extrema* de interdependencia funcional entre las dos proteínas.

  En el primer caso se estudia la similitud de sus árboles filogenéticos (los cuales representan la historia evolutiva de una proteína), pues se conoce que las proteínas que interaccionan o se asocian funcionalmente tienen árboles más similares entre sí que las que no están asociadas. En primer lugar, se obtiene un alineamiento múltiple de secuencia, reducido al conjunto de especies comunes en las que aparezcan ambas proteínas del par. De cada alineamiento múltiple de secuencia, asociado a cada una de las proteínas, se construye la correspondiente matriz de distancia entre secuencias. Estas matrices se utilizan generalmente para construir el árbol filogenético de la proteína. Pero en este método se utilizan directamente las matrices (en representación de los árboles) para calcular la correlación lineal entre las proteínas del par. Así, una correlación elevada entre las matrices se interpreta como indicativo de una alta similitud entre los árboles, y por tanto se considera una asociación funcional.

  Este método presenta el inconveniente de que necesita un alineamiento múltiple de secuencia de calidad, con secuencias de las dos proteínas del par en las mismas especies.

  Un esquema de este método aparece en la Figura B.1(d), y una descripción detallada en [Pazos and Valencia, 2001] y en su servidor web [Ochoa and Pazos, 2010].

- **Mutaciones correlacionadas (I2H, *In Silico Two-Hybrid*)**
  En este caso también se emplea la co-evolución, pero cuantificando el grado de covariación entre los pares de aminoácidos de las proteínas (lo cual se denomina mutaciones correlacionadas). Estas posiciones de la secuencia que cambian tanto en una como en la otra proteína del par que interacciona o se asocia funcionalmente, pueden ser debidas a modificaciones (mutaciones) compensatorias, para estabilizar el cambio producido en una proteína con el de la otra. Así, se contabilizan por separado las mutaciones, mediante coeficientes de correlación entre cada par de residuos.

Posteriormente se calcula la probabilidad de interacción, según la comparación de la distribución de los valores de correlación de las mutaciones intraproteína (de cada proteína del par), con la correlación de las mutaciones interproteína.

Este método presenta el mismo inconveniente del método MT, necesitando un alineamiento múltiple de secuencia de calidad.

Un esquema de este método aparece en la Figura B.1(e), y una descripción detallada en [Pazos and Valencia, 2002].

Todos estos métodos se basan en información genómica, y en las secuencias de las proteínas. Otros métodos computacionales de predicción de interacción o asociación funcional entre proteínas se basan en coevolución de dominios [Sprinzak and Margalit, 2001], coevolución de niveles de expresión génica [Fraser et al., 2004] u homología (interólogos) [Yu et al., 2004].

B.5. Métodos de Determinación de Rutas Biológicas

A partir de la combinación de interacciones o asociaciones funcionales por pares, determinadas según los métodos descritos en la sección anterior, se pueden derivar rutas biológicas (definidas en la sección 2.4.3).

La comparación de rutas biológicas reconstruidas para diferentes especies revela flexibilidad, con muchas variaciones específicas de la especie. Estas desviaciones de la ruta canónica se pueden usar para identificar dianas de fármacos cuando ciertas enzimas alternativas son específicas de un patógeno [Gabaldón and Huynen, 2004]. Por lo tanto, resulta interesante la construcción de redes biológicas a partir del genoma.

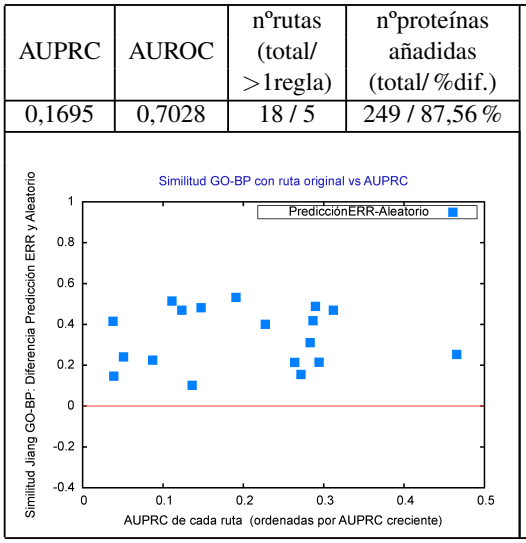
Tradicionalmente se ha potenciado la producción natural de sistemas biológicos, para posteriormente reconstruirlos. Existen tres enfoques para el diseño de rutas: la combinación de fragmentos de rutas existentes, la modificación o extensión de una ruta previa, o la creación de cada paso de la ruta independientemente. Centrándose en la última opción, existen diversos métodos y herramientas de definición de nuevas rutas desde cero (o *de novo*) [Karp et al., 2002; Adriaens et al., 2008; Prather and Martin, 2008], basados en diferentes enfoques que combinan el uso de homología para determinar la función molecular de las proteínas, con datos de experimentación a gran escala y técnicas basadas en contexto para identificar sus patrones funcionales [Gabaldón and Huynen, 2004]. Se han desarrollado diversas herramientas para la reconstrucción y validación de rutas biológicas, como PathoLogic [Karp et al., 2002], PathMiner [McShan et al., 2003] o PathFinder [Goesmann et al., 2002], aunque muchas se limitan a la visualización y manipulación de la ruta.

Como consecuencia de las distintas definiciones de ruta biológica [Bader et al., 2006], su implementación no es la misma en bases de datos diferentes, como por ejemplo Reactome [Matthews et al., 2009], KEGG [Kanehisa and Goto, 2000] o MetaCyc [Caspi et al., 2010]. Los esfuerzos para representarlos de forma común están en marcha (como *Pathway Commons* [Cerami et al., 2011]), pero todavía no se han resuelto las diferencias.

Apéndice C

Resumen de Resultados Extensión Rutas Metabólicas para Comparación. Sistema ERR-PRyC y ERR-PDR

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1695	0,7028	18 / 5	249 / 87,56 %



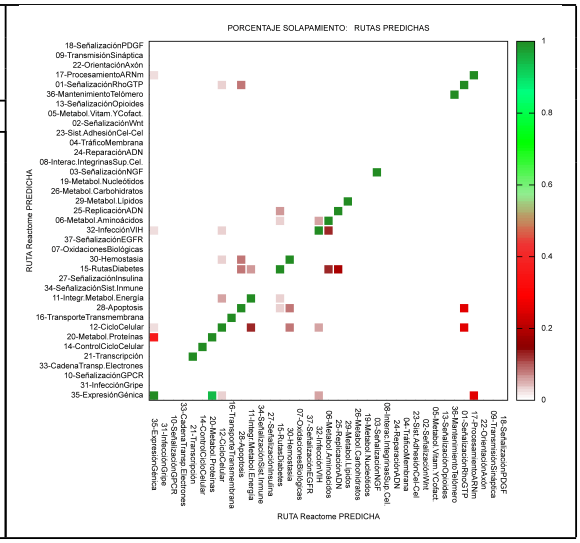
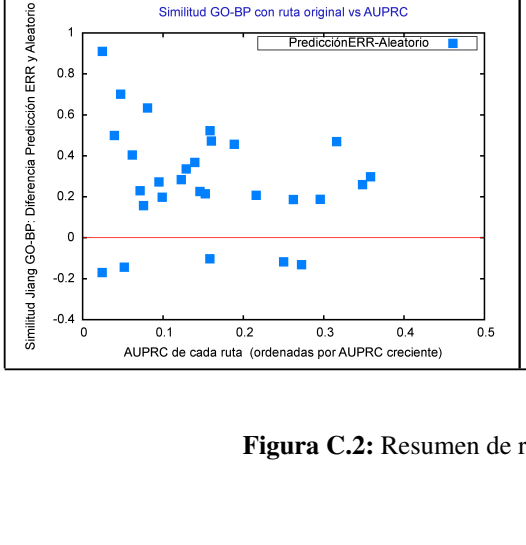


Figura C.1: Resumen de resultados sistema ERR-PRyC.

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1337	0,6914	28 / 15	383 / 85,90 %



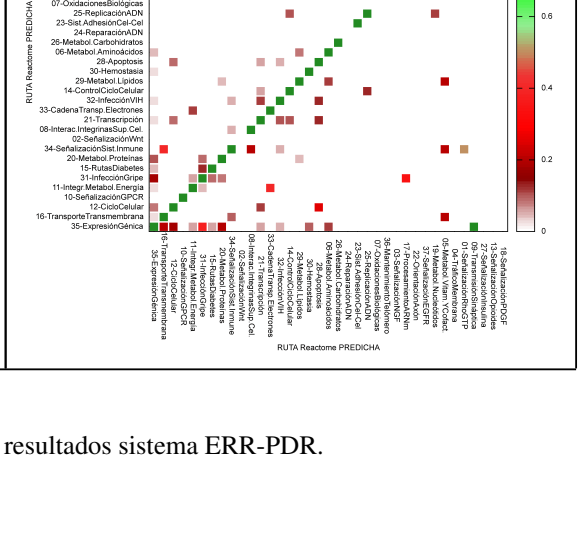


Figura C.2: Resumen de resultados sistema ERR-PDR.

Apéndice D

Resultados Detallados de Extensión por Ruta/Clase. Sistema ERR-PRyC y ERR-PDR

Tabla D.1: Resultados de la extensión por ruta individual, ordenadas por AUPRC creciente. Sistema ERR-PRyC.

Id. clase	Id. ruta Reactome	Nombre ruta Reactome	AUPRC	AUROC	Tamaño de ruta no redundante	Nº proteínas predichas por ERR	Nº prot. predichas por Glaab et al.	Nº prot. predichas en común	Nº reglas extensión	Media diversidad reglas	Media precisión en test de reglas	Similitud GO-BP predicciones
18	REACT_16888	Señalización de PDGF	0,0055	0,5000	12	0	4	0	0	0	0	0
9	REACT_13685	Transmisión Sináptica	0,0236	0,5705	26	0	1	0	0	0	0	0
22	REACT_18266	Orientación por Axón	0,0265	0,5875	33	0	6	0	0	0	0	0
17	REACT_1675	Procesamiento de ARNm	0,0377	0,5530	24	4	3	0	1	1	0	0,42
1	REACT_11044	Señalización de Rho GTP	0,0388	0,6366	33	4	2	0	1	1	0	0,33
36	REACT_7970	Mantenimiento del Telómero	0,0508	0,6835	26	4	0	0	1	1	0,20	0,24
		ALEATORIO (MEDIO)	**0,0524**	**0,4913**								
13	REACT_15295	Señalización de Opioides	0,0527	0,6474	13	0	2	0	0	0	0	0
5	REACT_11193	Metabol.Vitaminas y Cofactore	0,0532	0,8606	35	0	0	0	0	0	0	0
2	REACT_11045	Señalización de Wnt	0,0572	0,6890	27	0	2	0	0	0	0	0
23	REACT_19331	Sist.Adhesión Célula-Célula	0,0676	0,6240	12	0	0	0	0	0	0	0
4	REACT_11123	Tráfico de Membrana	0,0688	0,6377	33	0	0	0	0	0	0	0
24	REACT_216	Reparación del ADN	0,0721	0,6533	84	0	4	0	0	0	0	0
8	REACT_13552	Interac.Integrinas Sup.Celular	0,0747	0,6287	23	0	1	0	0	0	0	0
3	REACT_11061	Señalización de NGF	0,0871	0,6887	71	9	7	0	1	1	0	0,62
19	REACT_1698	Metabolismo de Nucleótidos	0,0974	0,6666	62	0	1	0	0	0	0	0
26	REACT_474	Metabolismo de Carbohidratos	0,1029	0,6924	53	0	0	0	0	0	0	0
29	REACT_602	Metabolismo de Lípidos	0,1111	0,6967	126	1	1	0	1	1	0	0,52
25	REACT_383	Replicación del ADN	0,1233	0,6654	61	11	2	0	1	1	0,50	0,70
6	REACT_13	Metabolismo de Aminoácidos	0,1362	0,6716	97	8	0	0	1	1	0,25	0,51
32	REACT_6185	Infección VIH	0,1474	0,7020	120	19	7	1	2	1	0,13	0,76
37	REACT_9417	Señalización de EGFR	0,1554	0,5380	16	0	4	0	0	0	0	0
		MEDIA	**0,1695**	**0,7028**								
7	REACT_13433	Oxidaciones Biológicas	0,1728	0,8084	39	0	1	0	0	0	0	0
30	REACT_604	Hemostasia	0,1907	0,7410	85	13	12	1	3	0,72	0,20	0,80
15	REACT_15380	Rutas de la Diabetes	0,2272	0,6856	186	32	3	0	2	1	0,25	0,65
27	REACT_498	Señalización Insulina	0,2421	0,6674	18	0	3	0	0	0	0	0
34	REACT_6900	Señalización Sistema Inmune	0,2428	0,7379	110	0	21	0	0	0	0	0
11	REACT_1505	Integración Metabol. Energía	0,2638	0,7583	144	16	2	0	1	1	0	0,65
28	REACT_578	Apoptosis	0,2717	0,7958	66	13	13	5	1	1	0,75	0,64
16	REACT_15518	Transporte Transmembrana de	0,2829	0,9268	36	4	0	0	1	1	0	0,41
12	REACT_152	Ciclo Celular, fase mitótica	0,2864	0,7730	197	37	12	0	2	0,87	0,75	0,73
20	REACT_17015	Metabolismo de Proteínas	0,2895	0,7290	148	16	5	0	1	1	0,63	0,72
14	REACT_1538	Controles del Ciclo Celular	0,2941	0,8338	59	10	2	0	2	1	0,67	0,47
21	REACT_1788	Transcripción	0,3119	0,7102	101	6	6	2	1	1	0,89	0,82
33	REACT_6305	Cadena Transp. Electrones	0,3528	0,9269	76	0	0	0	0	0	0	0
10	REACT_14797	Señalización de GPCR	0,3750	0,7706	117	0	2	0	0	0	0	0
31	REACT_6167	Infección Gripe	0,4126	0,8072	119	0	2	0	0	0	0	0
35	REACT_71	Expresión Génica	0,4653	0,7378	274	42	19	2	1	1	0,79	0,61

AUPRC < clasificadores Aleatorio o por Defecto (individuales)
Sin extensión <= 20% tamaño de la ruta

Tabla D.2: Resultados de la extensión por ruta individual, ordenadas por AUPRC creciente. Sistema ERR-PDR.

Id. clase	Id. ruta Reactome	Nombre ruta Reactome	AUPRC	AUROC	Tamaño de ruta no redundante	Nº proteínas predichas por ERR	Nº prot. predichas por Glaab et al.	Nº prot. predichas en común	Nº reglas extensión	Media diversidad reglas	Media precisión en test de reglas	Similitud GO-BP predicciones
18	REACT_16888	Señalización de PDGF	0,0188	0,6611	12	0	4	0	0	0	0	0
13	REACT_15295	Señalización de Opioides	0,0203	0,6313	13	0	2	0	0	0	0	0
27	REACT_498	Señalización Insulina	0,0243	0,6242	18	1	3	0	1	1	0	0,83
9	REACT_13685	Transmisión Sináptica	0,0244	0,5765	26	1	1	0	1	1	0	0,91
1	REACT_11044	Señalización de Rho GTP	0,0395	0,6156	33	2	2	0	1	1	0	0,50
4	REACT_11123	Tráfico de Membrana	0,0418	0,5927	33	0	0	0	0	0	0	0
5	REACT_11193	Metabol.Vitaminas y Cofactore	0,0472	0,6823	35	5	0	0	1	1	0	0,83
19	REACT_1698	Metabolismo de Nucleótidos	0,0517	0,6048	62	9	1	0	2	0,84	0	0,38
		ALEATORIO (MEDIO)	**0,0524**	**0,4913**								
37	REACT_9417	Señalización de EGFR	0,0569	0,6967	16	0	4	0	0	0	0	0
22	REACT_18266	Orientación por Axón	0,0572	0,7164	33	0	6	0	0	0	0	0
17	REACT_1675	Procesamiento de ARNm	0,0616	0,4899	24	3	3	0	1	1	0,50	0,57
3	REACT_11061	Señalización de NGF	0,0626	0,6354	71	0	7	0	0	0	0	0
36	REACT_7970	Mantenimiento del Telómero	0,0713	0,7685	26	5	0	0	1	1	0,33	0,41
7	REACT_13433	Oxidaciones Biológicas	0,0729	0,6060	39	0	1	0	0	0	0	0
25	REACT_383	Replicación del ADN	0,0757	0,7429	61	8	2	0	3	0,90	0	0,46
23	REACT_19331	Sist.Adhesión Célula-Célula	0,0806	0,5697	12	1	0	0	1	1	0	0,91
24	REACT_216	Reparación del ADN	0,0896	0,7188	84	0	4	0	0	0	0	0
26	REACT_474	Metabolismo de Carbohidratos	0,0948	0,6276	53	8	0	0	2	0,83	0,13	0,53
6	REACT_13	Metabolismo de Aminoácidos	0,0990	0,6268	97	18	0	0	2	0,87	0	0,55
28	REACT_578	Apoptosis	0,1227	0,6898	66	8	13	0	1	1	0,33	0,60
30	REACT_604	Hemostasia	0,1287	0,7624	85	11	12	0	2	1	0	0,66
		MEDIA	**0,1337**	**0,6914**								
29	REACT_602	Metabolismo de Lípidos	0,1394	0,7254	126	25	1	0	3	0,68	0	0,71
14	REACT_1538	Controles del Ciclo Celular	0,1460	0,8437	59	10	2	0	1	1	0,25	0,51
32	REACT_6185	Infección VIH	0,1525	0,6190	120	20	7	0	2	0,57	0,50	0,64
33	REACT_6305	Cadena Transp. Electrones	0,1581	0,7750	76	5	0	0	1	1	0,50	0,11
21	REACT_1788	Transcripción	0,1585	0,6669	101	18	6	3	2	1	0,50	0,80
8	REACT_13552	Interac.Integrinas Sup.Celular	0,1602	0,7557	23	5	1	0	1	1	1	0,68
2	REACT_11045	Señalización de Wnt	0,1839	0,7986	27	0	2	0	0	0	0	0
34	REACT_6900	Señalización Sistema Inmune	0,1885	0,7201	110	22	21	0	3	0,53	0	0,75
20	REACT_17015	Metabolismo de Proteínas	0,2158	0,7024	148	26	5	0	3	0,61	0,33	0,56
15	REACT_15380	Rutas de la Diabetes	0,2499	0,6832	186	36	3	0	2	1	0,50	0,32
31	REACT_6167	Infección Gripe	0,2620	0,7726	119	24	2	0	4	0,43	0,42	0,51
11	REACT_1505	Integración Metabol. Energía	0,2723	0,7157	144	18	2	0	2	0,87	0,25	0,26
10	REACT_14797	Señalización de GPCR	0,2955	0,8272	117	14	2	0	2	1	0,30	0,68
12	REACT_152	Ciclo Celular, fase mitótica	0,3159	0,6983	197	23	12	0	1	1	0,63	0,72
16	REACT_15518	Transporte Transmembrana de	0,3483	0,9488	36	5	0	0	1	1	0	0,56
35	REACT_71	Expresión Génica	0,3581	0,6889	274	52	19	2	13	0,29	0,34	0,63

AUPRC < clasificadores Aleatorio o por Defecto (individuales)
Sin extensión <= 20% tamaño de la ruta

Apéndice E

Resultados Cuantitativos Homólogos Anotados y Predichos por ERR-PDR

Tabla E.1: Homólogas de proteínas predichas por ERR-PDR y anotadas en Reactome de conjuntos de entrenamiento y test.

Id. clase	Nombre ruta Reactome	**Cjto.A:** Anotadas Reactome entrenamiento y test	**Cjto.B:** Predichas por ERR en cjto.A	**Homólogas de cjto.A en no anotadas** (Aplicación(8187) / Resto No Anotadas(10607))	**Homólogas de cjto.B en no anotadas** (Aplicación(8187) / Resto No Anotadas(10607))	**Homólogas de cjto.B en no anotadas Y predichas por ERR** (Aplicación / Resto No Anotadas)	**Cjto.C:** Anotadas Reactome no incluidas en entrenamiento ni test	**Homólogas de cjto.A en cjto.C**	**Homólogas de cjto.B en cjto.C**	**Homólogas de cjto.B en cjto.C Y predichas por ERR**
1	Señalización de Rho GTP	33	2	0 / 34	0 / 27	0 / 0	91	42	19	0
2	Señalización de Wnt	27	0	0 / 7	-	-	25	11	11	-
3	Señalización de NGF	71	0	0 / 141	-	-	137	96	96	-
4	Tráfico de Membrana	33	0	0 / 44	-	-	17	11	11	-
5	Metabol.Vitaminas y Cofactores	35	3	0 / 17	0 / 0	0 / 0	11	13	0	0
6	Metabolismo de Aminoácidos	97	4	0 / 37	0 / 1	0 / 0	64	55	3	0
7	Oxidaciones Biológicas	39	0	0 / 48	-	-	73	68	68	-
8	Interac.Integrinas Sup.Celular	23	3	0 / 67	0 / 0	0 / 0	57	42	7	1
9	Transmisión Sináptica	26	2	0 / 40	0 / 3	0 / 0	43	24	3	0
10	Señalización de GPCR	117	11	0 / 144	0 / 15	0 / 0	574	500	38	7
11	Integración Metabol. Energía	144	5	0 / 59	0 / 2	0 / 0	68	57	0	0
12	Ciclo Celular, fase mitótica	197	14	0 / 53	0 / 5	0 / 0	95	58	1	0
13	Señalización de Opioides	13	0	0 / 17	-	-	50	39	39	-
14	Controles del Ciclo Celular	59	7	0 / 7	0 / 0	0 / 0	52	18	1	0
15	Rutas de la Diabetes	186	5	0 / 158	0 / 0	0 / 0	81	64	0	0
16	Transporte Transmembrana de	36	2	0 / 21	0 / 0	0 / 0	87	79	2	0
17	Procesamiento de ARNm	24	3	0 / 1	0 / 0	0 / 0	8	1	0	0
18	Señalización de PDGF	12	0	0 / 14	-	-	52	22	22	-
19	Metabolismo de Nucleótidos	62	0	0 / 22	-	-	25	23	0	-
20	Metabolismo de Proteínas	148	23	0 / 44	0 / 5	0 / 0	52	27	1	0
21	Transcripción	101	16	0 / 21	0 / 1	0 / 0	30	4	0	0
22	Orientación por Axón	33	0	0 / 53	-	-	106	46	46	-
23	Sist.Adhesión Célula-Célula	12	2	0 / 34	0 / 0	0 / 0	35	25	3	0
24	Reparación del ADN	84	0	0 / 41	-	-	20	7	7	-
25	Replicación del ADN	61	6	0 / 3	0 / 1	0 / 0	33	17	3	0
26	Metabolismo de Carbohidratos	53	0	0 / 40	-	-	43	45	6	-
27	Señalización Insulina	18	3	0 / 11	0 / 0	0 / 0	21	12	1	0
28	Apoptosis	66	4	0 / 66	0 / 0	0 / 0	57	26	0	0
29	Metabolismo de Lípidos	126	6	0 / 101	0 / 5	0 / 0	92	57	1	0
30	Hemostasia	85	7	0 / 252	0 / 87	0 / 1	143	71	12	0
31	Infección Gripe	119	29	0 / 60	0 / 4	0 / 0	44	19	1	0
32	Infección VIH	120	16	0 / 79	0 / 0	0 / 0	60	28	0	0
33	Cadena Transp. Electrones	76	3	0 / 5	0 / 0	0 / 0	1	1	0	0
34	Señalización Sistema Inmune	110	12	0 / 147	0 / 4	0 / 0	150	68	2	0
35	Expresión Génica	274	79	0 / 79	0 / 22	0 / 3	105	52	7	0
36	Mantenimiento del Telómero	26	3	0 / 1	0 / 0	0 / 0	19	15	0	0
37	Señalización de EGFR	16	0	0 / 23	-	-	31	11	11	-

AUPRC < clasificadores Aleatorio o por Defecto (individuales)
Sin extensión <= 20% tamaño de la ruta

Tabla E.2: Homólogas de proteínas predichas por ERR-PDR y anotadas en Reactome redundantes a entrenamiento y test.

Id. clase	Nombre ruta Reactome	**Cjto.A:** Anotadas Reactome no incluidas ni en entrenamiento ni en test	**Cjto.B:** Predichas por ERR en cjto.A	**Homólogas de cjto.A en no anotadas** (Aplicación(8187) / Resto No Anotadas(10607))	**Homólogas de cjto.B en no anotadas** (Aplicación(8187) / Resto No Anotadas(10607))	**Homólogas de cjto.B en no anotadas Y predichas por ERR** (Aplicación / Resto No Anotadas)
1	Señalización de Rho GTP	91	0	3 / 66	0 / 0	0 / 0
2	Señalización de Wnt	25	0	0 / 52	-	-
3	Señalización de NGF	137	0	6 / 232	-	-
4	Tráfico de Membrana	17	0	2 / 15	-	-
5	Metabol.Vitaminas y Cofactores	11	0	0 / 8	0 / 0	0 / 0
6	Metabolismo de Aminoácidos	64	0	2 / 65	0 / 0	0 / 0
7	Oxidaciones Biológicas	73	0	1 / 36	-	-
8	Interac.Integrinas Sup.Celular	57	3	3 / 102	0 / 0	0 / 0
9	Transmisión Sináptica	43	0	3 / 95	0 / 0	0 / 0
10	Señalización de GPCR	574	18	10 / 252	2 / 14	0 / 0
11	Integración Metabol. Energía	68	0	1 / 79	0 / 0	0 / 0
12	Ciclo Celular, fase mitótica	95	2	3 / 220	0 / 2	0 / 0
13	Señalización de Opioides	50	0	0 / 103	-	-
14	Controles del Ciclo Celular	52	0	1 / 111	0 / 0	0 / 0
15	Rutas de la Diabetes	81	2	5 / 262	0 / 3	0 / 0
16	Transporte Transmembrana de	87	0	0 / 23	0 / 0	0 / 0
17	Procesamiento de ARNm	8	0	0 / 25	0 / 0	0 / 0
18	Señalización de PDGF	52	0	3 / 165	-	-
19	Metabolismo de Nucleótidos	25	0	0 / 12	-	-
20	Metabolismo de Proteínas	52	0	2 / 75	0 / 0	0 / 0
21	Transcripción	30	2	3 / 99	0 / 2	0 / 0
22	Orientación por Axón	106	0	6 / 193	-	-
23	Sist.Adhesión Célula-Célula	35	0	5 / 49	0 / 0	0 / 0
24	Reparación del ADN	20	0	0 / 28	-	-
25	Replicación del ADN	33	0	0 / 82	0 / 0	0 / 0
26	Metabolismo de Carbohidratos	43	1	2 / 53	-	-
27	Señalización Insulina	21	1	2 / 117	0 / 0	0 / 0
28	Apoptosis	57	2	0 / 141	0 / 4	0 / 0
29	Metabolismo de Lípidos	92	0	6 / 145	0 / 0	0 / 0
30	Hemostasia	143	2	8 / 303	0 / 8	0 / 0
31	Infección Gripe	44	1	1 / 25	0 / 0	0 / 0
32	Infección VIH	60	1	3 / 122	0 / 0	0 / 0
33	Cadena Transp. Electrones	1	0	0 / 0	0 / 0	0 / 0
34	Señalización Sistema Inmune	150	5	5 / 256	0 / 5	0 / 0
35	Expresión Génica	105	10	6 / 94	0 / 3	0 / 0
36	Mantenimiento del Telómero	19	0	0 / 17	0 / 0	0 / 0
37	Señalización de EGFR	31	0	4 / 113	-	-

AUPRC < clasificadores Aleatorio o por Defecto (individuales)
Sin extensión <= 20% tamaño de la ruta

Apéndice F

Mapas de Agrupación de Proteínas por Propiedades Simples

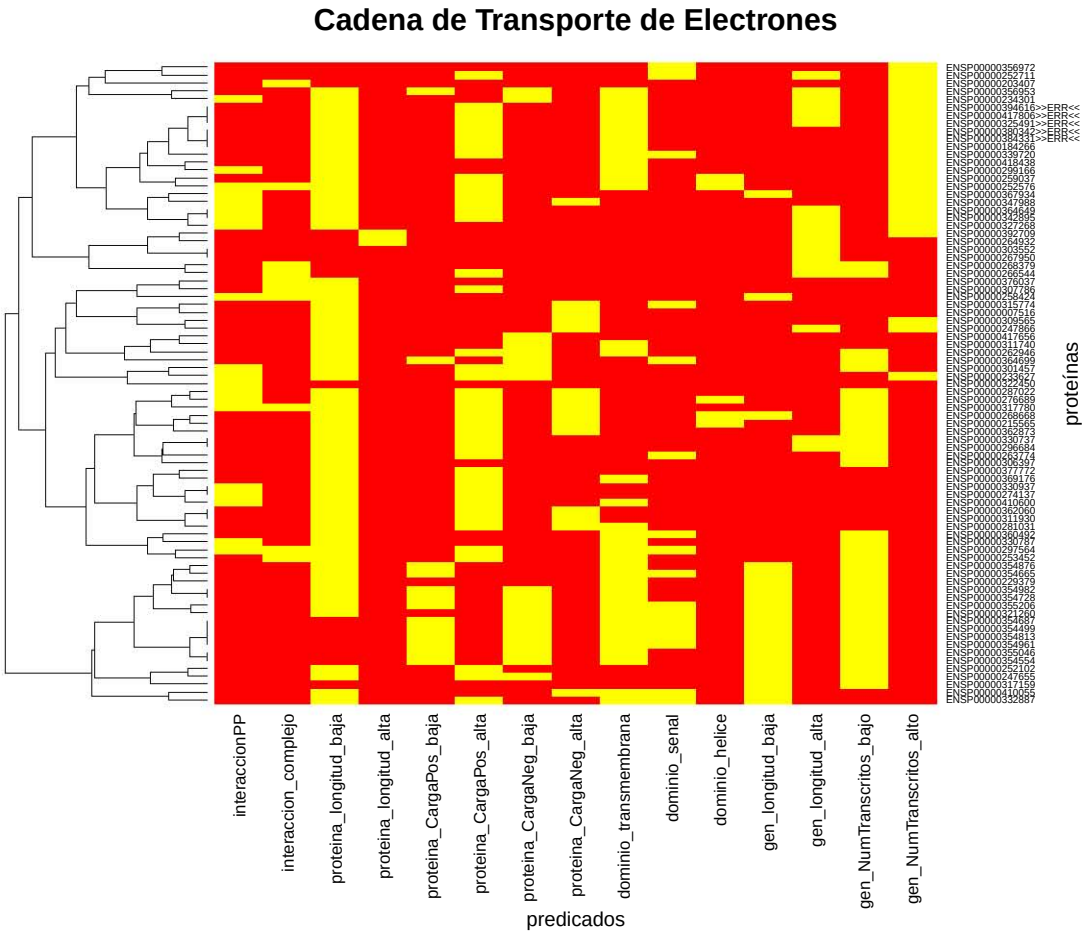


Figura F.1: Mapa de agrupación de proteínas por propiedades simples. Se incluyen proteínas de la ruta original y predichas por ERR (con el sufijo $>> ERR <<$ en las etiquetas de las filas). No hay predichas por Glaab et al. Cada propiedad simple se representa con un predicado lógico (cada columna). Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0). Se usan los identificadores de UniProt. Ruta *Cadena de transporte de electrones*.

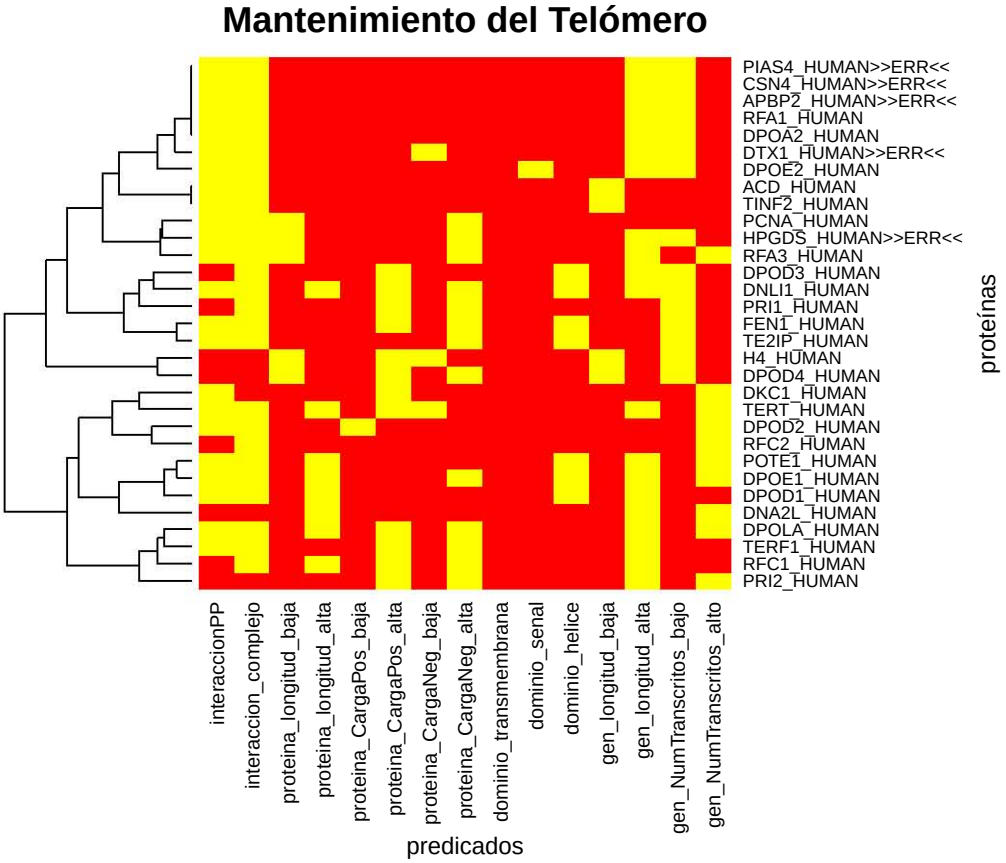


Figura F.2: Mapa de agrupación de proteínas por propiedades simples. Se incluyen proteínas de la ruta original y predichas por ERR (con el sufijo $>> ERR <<$ en las etiquetas de las filas). No hay predichas por Glaab et al. Cada propiedad simple se representa con un predicado lógico (cada columna). Para cada proteína, el amarillo representa que la propiedad es cierta (1) y el rojo que es falsa (0). Se usan los identificadores de UniProt. Ruta *Mantenimiento del telómero*.

Apéndice G

Figuras Interpretación Extensión Rutas con Sistema ERR-PRyC

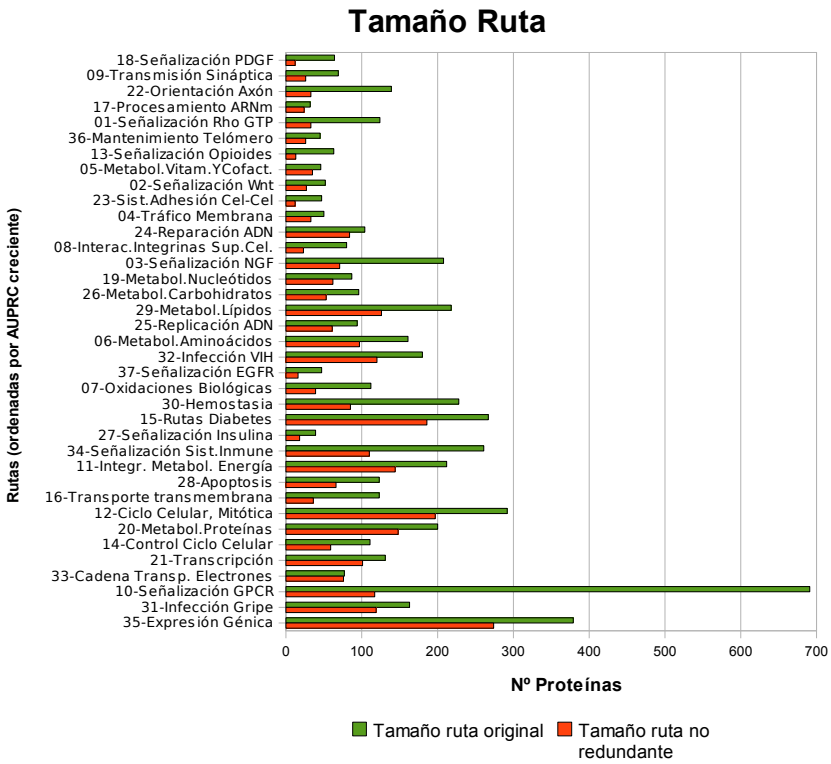


Figura G.1: Análisis de rendimiento frente a tamaño de ruta. Sistema ERR-PRyC. Rutas ordenadas de arriba a abajo, de menor a mayor AUPRC. Las barras verdes representan la cantidad de proteínas de la ruta original, y las barras naranjas la cantidad de proteínas tras eliminar las redundantes entre sí (las que se usan en el aprendizaje).

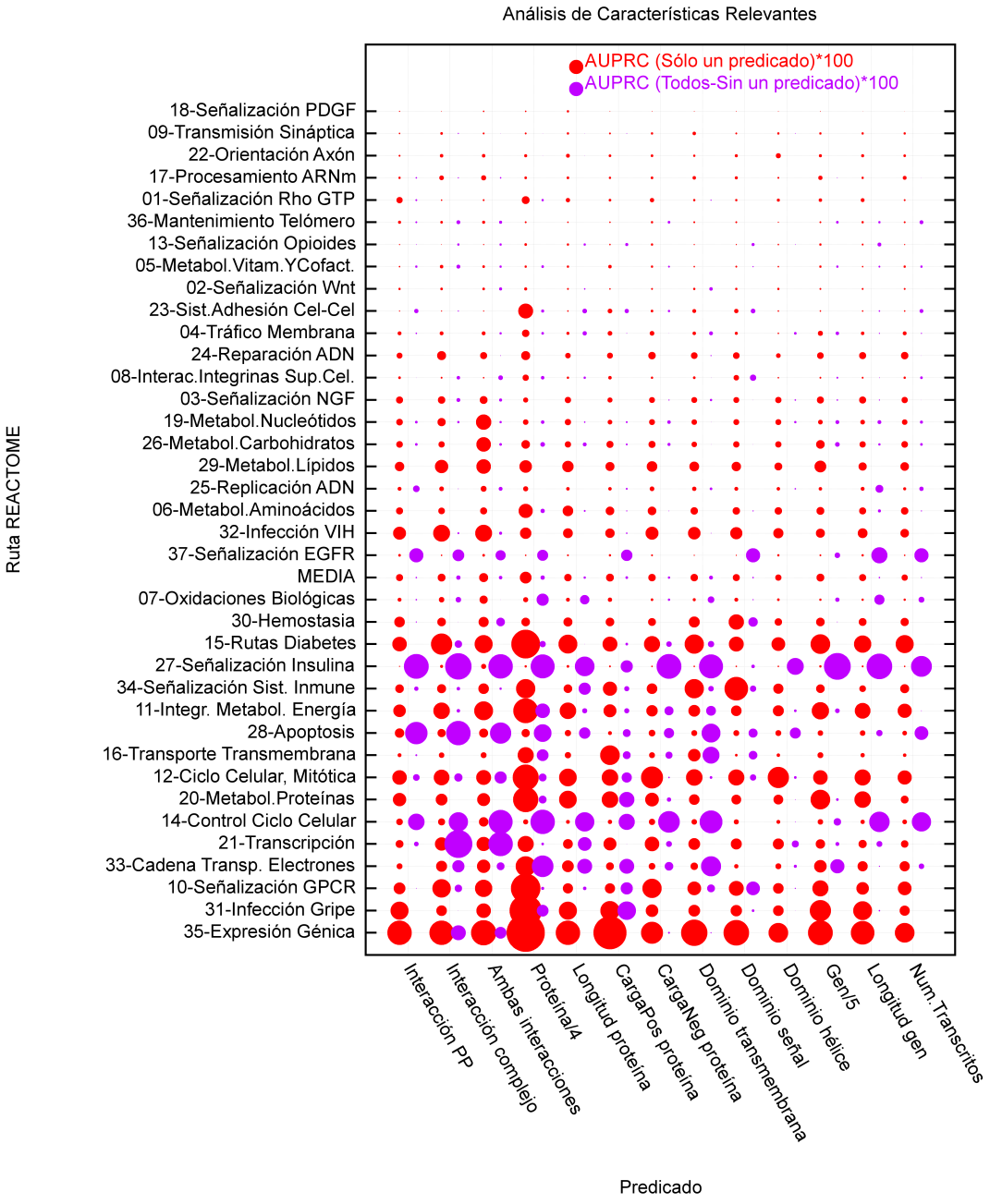
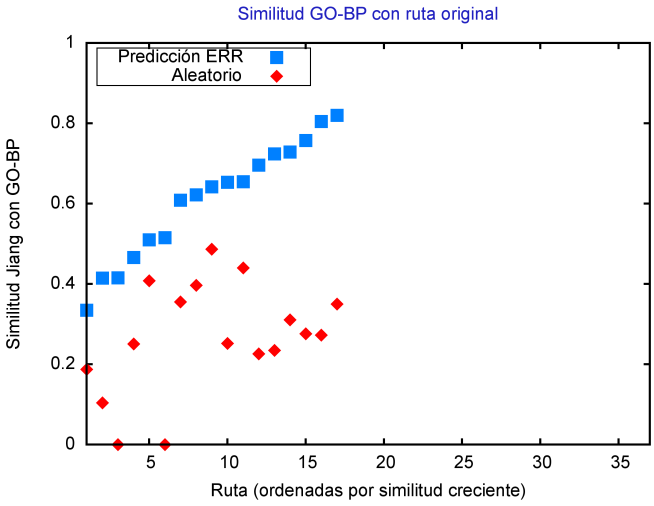
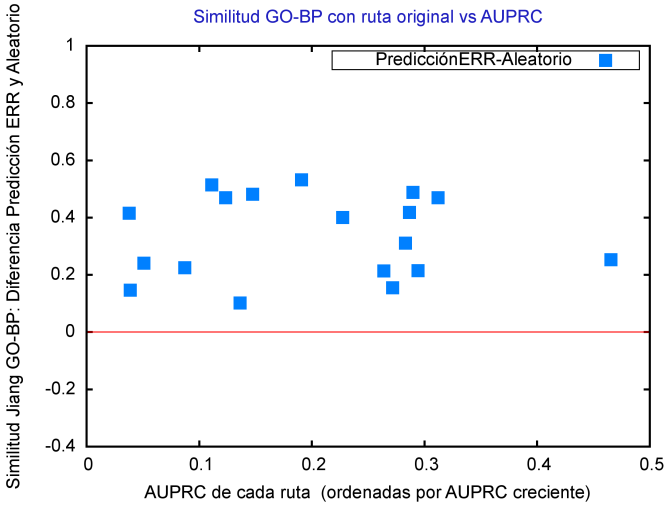


Figura G.2: Análisis de predicados relevantes en el aprendizaje. Los círculos rojos (izquierda) representan una propiedad relevante por sí misma. Los círculos morados (derecha) representan una propiedad relevante en combinación con otras. Las rutas (filas) están ordenadas, de abajo a arriba, de mejor a peor AUPRC, según el sistema. Sistema ERR-PRyC.



(a) Orden por similitud creciente.



(b) Orden por AUPRC creciente.

Figura G.3: Similitud de anotación funcional entre proteínas de la ruta original y proteínas añadidas (por predicción y aleatoriamente). Sistema ERR-PRyC. Las rutas sin extensión no se representan. (a) Rutas ordenadas por similitud creciente en el grupo de predicciones. Cada punto representa la similitud absoluta de las proteínas añadidas a la ruta original. (b) Rutas ordenadas por AUPRC creciente en el grupo de predicciones. Cada punto representa la diferencia de similitud a la ruta original entre las proteínas predichas y las proteínas aleatorias (*Sim.Predichas–Sim.Aleatorias*) para esa ruta. Así, la línea roja representa la inexistencia de mejora de las predicciones frente a la aleatoriedad, en términos de similitud.

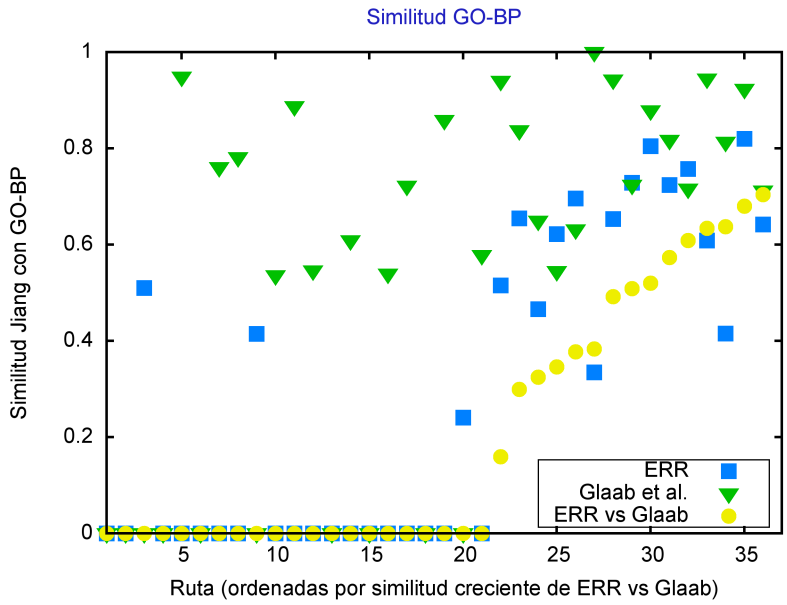


Figura G.4: Similitud de anotación funcional entre proteínas de la ruta original y las proteínas añadidas (ERR-PRyC y Glaab et al.) y entre ambos sistemas de extensión.

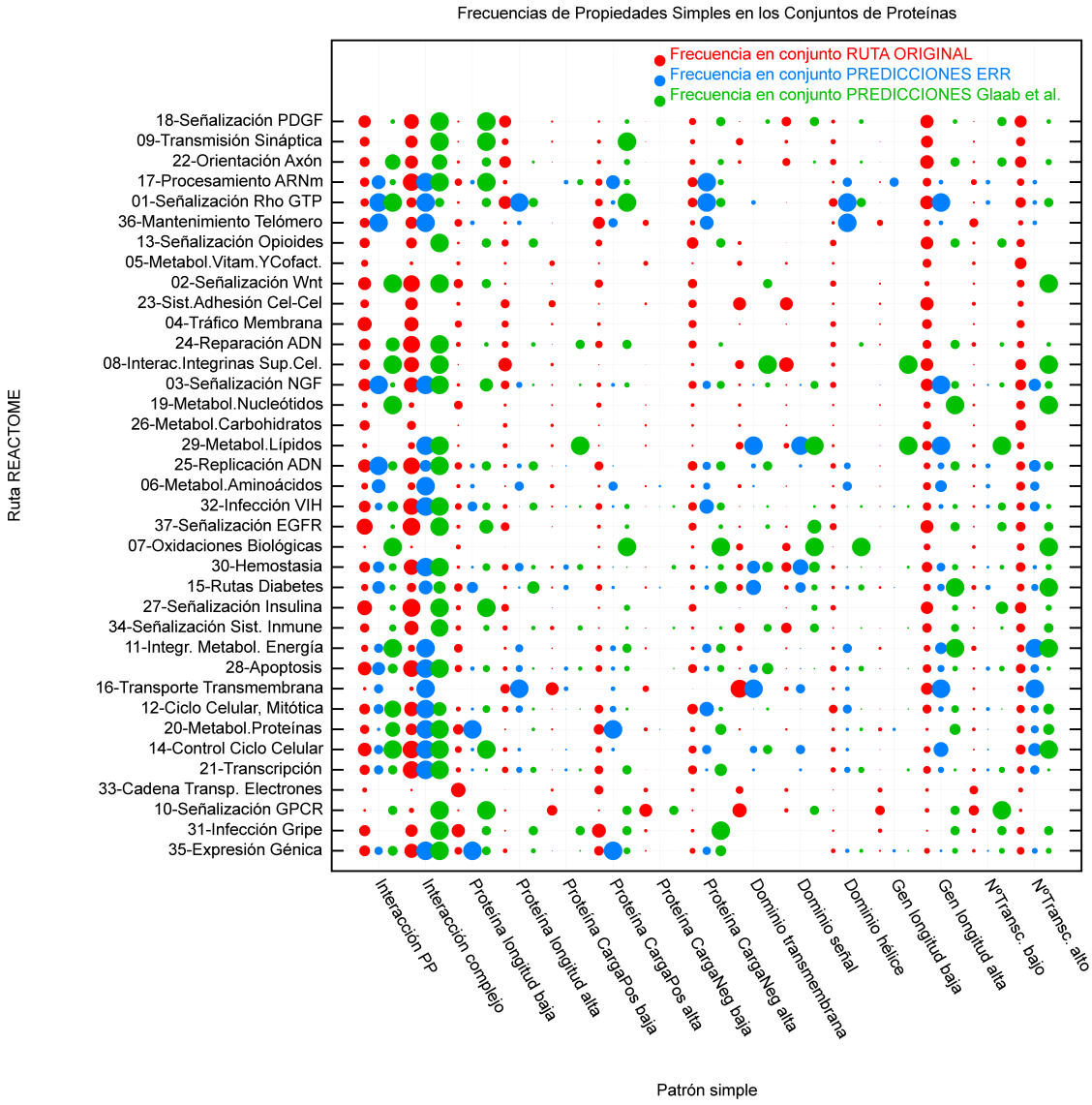
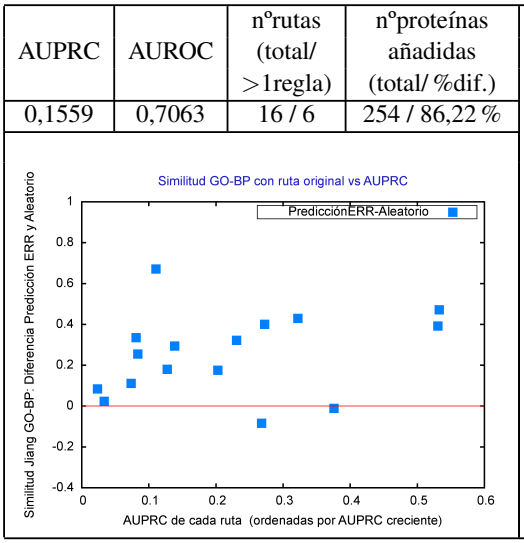


Figura G.5: Comparación de frecuencia de predicados simples por ruta. Sistema ERR-PRyC. Los círculos izquierdos/rojos representan la frecuencia en las proteínas de las rutas originales, los círculos centrales/azules la frecuencia en las proteínas predichas por el sistema ERR-PDR, y los círculos derechos/verdes la frecuencia en las proteínas expandidas por el método Glaab et al.

Apéndice H

Resumen de Resultados Extensión Rutas para Comparación. Varios Sistemas

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1559	0,7063	16 / 6	254 / 86,22 %



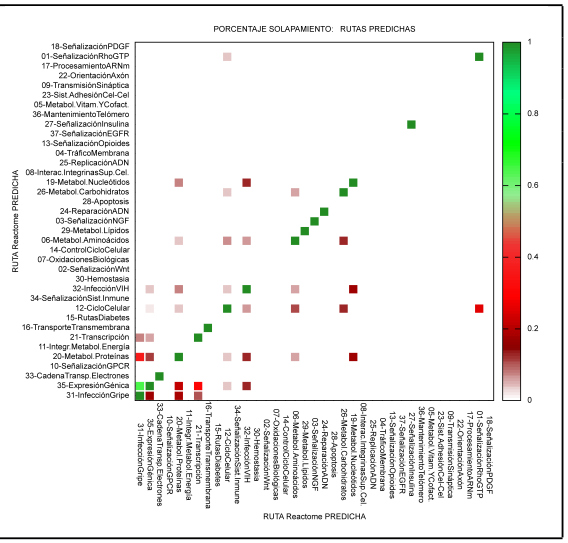
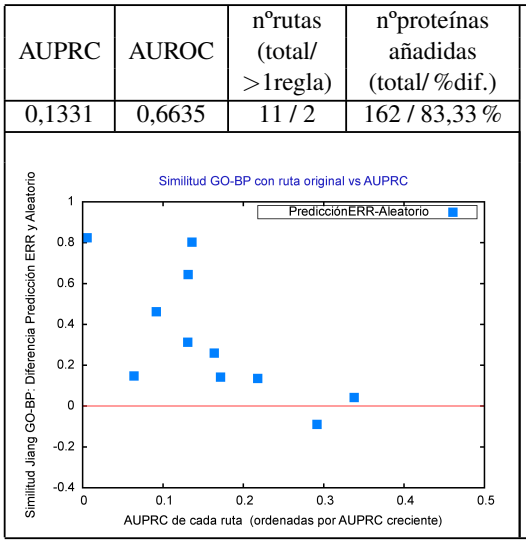


Figura H.1: Resumen de resultados sólo con interacciones PP (sin complejos).

AUPRC	AUROC	nºrutas (total/ >1regla)	nºproteínas añadidas (total/ %dif.)
0,1331	0,6635	11 / 2	162 / 83,33 %



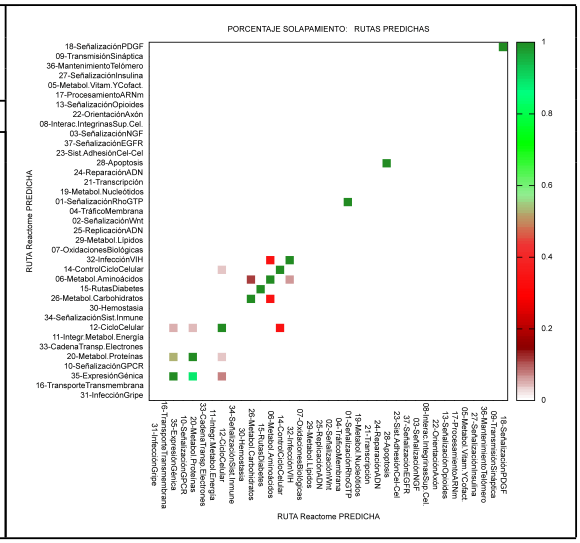


Figura H.2: Resumen de resultados sólo con complejos (sin interacciones PP).

Acrónimos

AA	Aprendizaje Automático
AAP	Aprendizaje Automático Proposicional
AAR	Aprendizaje Automático Relacional
ADN	Acido DesoxirriboNucleico
AFPP	Asociación/es Funcional/es entre Pares de Proteínas
AODE	Promedio de estimadores con una dependencia (del inglés, *Averaged One-Dependence Estimators*)
ARN	Acido RiboNucleico
ATP	Adenosín Trifosfato (del inglés, *Adenosine TriPhosphate*)
AUC	Area bajo la curva (del inglés, *Area Under Curve*)
AUPRC	Area bajo la curva PR (del inglés, *Area Under PR Curve*)
AUROC	Area bajo la curva ROC (del inglés, *Area Under ROC Curve*)
BLAST	Herramienta de búsqueda de alineamientos de secuencia de forma local (del inglés, *Basic Local Alignment Search Tool*)
CI	Clasificadores Individuales por clase
E/R	Entidad-Relación (modelo)
EGFR	Receptor del factor de crecimiento epidérmico (del inglés, *Epidermal Growth Factor Receptor*)
ERC	Constante aleatoria efímera (del inglés, *Ephemeral Random Constant*)
ERR	Extensión basada en Representación Relacional
ERR-PDR	Extensión basada en Representación Relacional que Prioriza la Diversidad de Reglas
ERR-PRyC	Extensión basada en Representación Relacional que Prioriza el Rendimiento y la Cobertura
FN	Falsos Positivos (del inglés, *False Negatives*)
FP	Falsos Positivos (del inglés, *False Positives*)
GC (método)	Conservación de genes adyacentes (del inglés, *Gene Context*)
GF (método)	Eventos de fusión de genes (del inglés, *Gene Fusion*)
GO	Ontología Génica (del inglés, *Gene Ontology*)
GO-BP	Ontología Génica de Proceso Biológico (del inglés, *Gene Ontology-Biological Process*)
GO-MF	Ontología Génica de Función Molecular (del inglés, *Gene Ontology-Molecular Function*)
GO-CC	Ontología Génica de Componente Celular (del inglés, *Gene Ontology-Cellular Component*)
GPCR	Receptor acoplado a proteínas G (del inglés, *G Protein-Coupled Receptor*)
GTP	Guanosín Trifosfato (del inglés, *Guanosine TriPhosphate*)

I2H (método)	Mutaciones correlacionadas (del inglés, *In Silico Two-Hybrid*)
ILP	Programación Lógica Inductiva (del inglés, *Inductive Logic Programming*)
IPP	Interacción/es Proteína-Proteína (también aparece como interacciones PP)
MC	Multi-Clasificador
MCC	Coeficiente de Correlación de Matthews (del inglés, *Matthews Correlation Coefficient*)
MT (método)	Similitud de árboles filogenéticos (del inglés, *MirrorTree*)
NGF	Factor de crecimiento nervioso (del inglés, *Nerve Growth Factor*)
ORC	Complejo de reconocimiento de los sitios de origen de la replicación (del inglés, *Origin Recognition Complex*)
PDGF	Factor de crecimiento derivado de plaquetas (del inglés, *Platelet-Derived Growth Factor*)
PG	Programación Genética
PP (método)	Perfiles filogenéticos (del inglés, *Phylogenetic Profiles*)
PR (curva)	Precisión-Sensibilidad (del inglés, *Precision-Recall*)
ROC (curva)	Característica operativa del receptor (del inglés, *Receiver Operating Characteristic*)
SNP	Polimorfismos de nucleótidos aislados (del inglés, *Single Nucleotide Polymorphisms*)
TERC	Componente ARN de la telomerasa (del inglés, *Telomerase RNA Component*)
TERT	Transcriptasa inversa de la telomerasa (del inglés, *Telomerase Reverse Transcriptase*)
TN	Verdaderos Negativos (del inglés, *True Negatives*)
TP	Verdaderos Positivos (del inglés, *True Positives*)
VIH	Virus de Inmunodeficiencia Humana

Bibliografía

Adriaens, M. E., Jaillard, M., Waagmeester, A., Coort, S. L. M., Pico, A. R., and Evelo, C. T. A. (2008). The public road to high-quality curated biological pathways. *Drug Discovery Today*, 13(19-20):856–862.

Aebersold, R. (2005). Molecular Systems Biology: a new journal for a new biology? *Molecular Systems Biology*, 1.

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 207–216. ACM.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). *Fast Discovery of Association Rules*, pages 307–328. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. Morgan Kaufmann.

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255.

Al-Shahib, A., Breitling, R., and Gilbert, D. (2005). Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics*, 4(3):195–203.

Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J., and Dopazo, J. (2006). BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research*, 34(suppl 2):W472–476.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa,

A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. F., and Hogue, C. W. V. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(suppl 1):418–424.

Allen, J. E., Pertea, M., and Salzberg, S. L. (2004). Computational gene prediction using multiple sources of evidence. *Genome Research*, 14(1):142–148.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S. F., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H.-C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., and Mori, H. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Research*, 16(5):686–691.

Armañanzas, R., Larrañaga, P., and Bielza, C. (2012). Ensemble transcript interaction networks: A case study on Alzheimer's disease. *Computer Methods and Programs in Biomedicine*, *in Press*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

Attwood, T. K. (2002). The PRINTS database: A resource for identification of protein families. *Briefings in Bioinformatics*, 3(3):252–263.

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(suppl 1):D504–D506.

Bader, S., Kühner, S., and Gavin, A.-C. (2008). Interaction networks for systems biology. *FEBS Letters*, 582(8):1220–1224.

Bailey, D. and O'Hare, P. (2005). Comparison of the SUMO1 and ubiquitin conjugation pathways during the inhibition of proteasome activity with evidence of SUMO1 recycling. *Biochemical Journal*, 392(2):271–281.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. Bradford Books, Cambridge, Massachusetts, U.S.A., Massachusetts.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Bandyopadhyay, S., Maulik, U., and Wang, J. T. L. (2007). *Analysis of biological data. A soft computing approach.*, volume 3. World Scientific, Singapore.

Bao, L. and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, 21(10):2185–2190.

Barabasi, A. L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5):60–69.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, 5(2):101–113.

Baudot, A., Martin, D., Mouren, P., Chevenet, F., Guenoche, A., Jacq, B., and Brun, C. (2006). PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks. *Bioinformatics*, 22(2):248–250.

Bendtsen, J. D., Jensen, L. J., Blom, N., von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering Design and Selection*, 17(4):349–356.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Biggs, N., Lloyd, E. K., and Wilson, R. J. (1986). *Graph Theory, 1736-1936*. Clarendon Press, New York, NY, USA.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press, New York.

Blockeel, H. (1998). *Top-Down Induction of First Order Logical Decision Trees*. PhD thesis, Informatics Section, Department of Computer Science, Faculty of Engineering.

Blockeel, H., Dehaspe, L., Demoen, B., Janssens, G., Ramon, J., and Vandecasteele, H. (2000). Executing query packs in ILP. In *Proceedings of the 10th International Conference on Inductive Logic Programming*, volume 1866 of *Lecture Notes in Artificial Intelligence*, pages 60–77. Springer.

Blockeel, H., Dehaspe, L., Ramon, J., Struyf, J., Assche, A. V., Vens, C., and Fierens, D. (2006a). The ACE data mining system. User's manual.

Blockeel, H. and Dzeroski, S. (1999). Experiments with TILDE in the river water quality domain. Technical report, Institut J. Stefan, Ljubljana.

Blockeel, H., Dzeroski, S., and Grbovic, J. (1999). Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. In *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, volume 1704 of *Lecture Notes in Artificial Intelligence*, pages 32–40. Springer.

Blockeel, H., Leander, S., Struyf, J., Dzeroski, S., and Clare, A. (2006b). Decision trees for hierarchical multilabel classification: A case study in functional genomics. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 4213, pages 18–29. Springer.

Blockeel, H., Page, D., and Srinivasan, A. (2005). Multi-instance tree learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 57–64.

Blockeel, H. and Raedt, L. D. (1998). Top-down induction of logical decision trees. *Artificial Intelligence*, 101 (1-2):285–297.

Blockeel, H., Raedt, L. D., and Ramon, J. (1998). Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine Learning*, pages 55–63. Morgan Kaufmann.

Borrajo, D., Gonzalez, J., and Isasi, P. (2006). *Aprendizaje Automatico*. Sanz Y Torres, Madrid.

Bouckaert, R. R. (2004). Bayesian Network Classifiers in Weka. Technical report, University of Waikato.

Bowers, P., Cokus, S., Eisenberg, D., and Yeates, T. (2004). Use of logic relationships to decipher protein network organization. *Science*, 306(5705):2246–2249.

Bratko, I. (2001). *Prolog Programming for Artificial Intelligence*. Addison Wesley, Harlow, England.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433(7025):531–537.

Cary, M. P., Bader, G. D., and Sander, C. (2005). Pathway information for systems biology. *FEBS Letters*, 579(8):1815–1820.

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P., and Karp, P. D. (2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 38(suppl 1):D473–D479.

Catanzaro, D., Pesenti, R., and Milinkovitch, M. C. (2007). An ant colony optimization algorithm for phylogenetic estimation under the minimum evolution principle. *BMC Evolutionary Biology*, 7:228.

Causier, B. (2004). Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrometry Reviews*, 23(5):350–367.

Cawley, S. L. and Pachter, L. (2003). HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, 19(suppl 2):ii36–ii41.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690.

Chagoyen, M. and Pazos, F. (2010). Quantifying the biological significance of gene ontology biological processes - implications for the analysis of systems-wide data. *Bioinformatics*, 26(3):378–384.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1):2.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(suppl 1):D572–574.

Che, D., Liu, Q., Rasheed, K., and Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Advances in Experimental Medicine and Biology*, 696:191–199.

Chen, J., Kelley, L. A., Muggleton, S., and Sternberg, M. J. E. (2008). Protein fold discovery using stochastic logic programs. In *Proceedings of the Probabilistic Inductive Logic Programming*, volume 4911 of *Lecture Notes in Artificial Intelligence*, pages 244–262.

Chen, J. J. (2007). Key aspects of analyzing microarray gene-expression data. *Pharmacogenomics*, 8(5):473–482.

Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, 26(1):73–79.

Cho, K. H., Choo, S. M., Jung, S. H., Kim, J. R., Choi, H. S., and Kim, J. (2007). Reverse engineering of gene regulatory networks. *IET Systems Biology*, 1(3):149–163.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630.

Clare, A. (2003). *Machine learning and data mining for yeast functional genomics*. PhD thesis, University of Wales Aberystwyth.

Clare, A., Karwath, A., Ougham, H., and King, R. D. (2006). Functional bioinformatics for Arabidopsis thaliana. *Bioinformatics*, 22(9):1130–1136.

Clare, A. and King, R. D. (2003). Data mining the yeast genome in a lazy functional language. In *Proceedings of the 5th International Symposium Practical Aspects of Declarative Languages*, volume 2562 of *Lecture Notes in Computer Science*, pages 19–36. Springer.

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.

Cleary, J. G. and Trigg, L. E. (1995). K*: an instance-based learner using an entropic distance measure. In *Proceedings of the 12th International Conference on Machine Learning*, pages 108–114. Morgan Kaufmann.

Consortium, T. U. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R., and Kornberg, R. D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, 288(5466):640–649.

Cussens, J. (2001). Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine learning*, pages 233–240. ACM.

de Miguel Castaño, A., Velthuis, M. G. P., and Martínez, E. M. (1999). *Diseño de bases de datos relacionales*. Ra-Ma, Madrid.

Dehaspe, L. and Raedt, L. D. (1997). Mining association rules in multiple relations. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, pages 125–132. Springer.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Reubenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka,

M., Novere, N. L., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1):1–38.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):P3.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71.

Do, C. B. and Katoh, K. (2008). Protein multiple sequence alignment. *Methods in Molecular Biology*, 484:379–413.

Domingos, P., Kok, S., Poon, H., Richardson, M., and Singla, P. (2006). Unifying logical and statistical AI. In *Proceedings of the 21th National Conference on Artificial Intelligence*, pages 2–7.

Drummond, C. and Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130.

Dzeroski, S. (2003). Multi-relational data mining: an introduction. *SIGKDD Explorer Newsletter*, 5(1):1–16.

Dzeroski, S. and Lavrac, N. (2001). *Relational Data Mining*. Springer.

Dzeroski, S., Raedt, L. D., and Driessens, K. (2001). Relational reinforcement learning. *Machine Learning*, 43(1/2):7–52.

Emanuelsson, O., Nielsen, H., and Heijne, G. V. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, 8(5):978–984.

Emde, W. and Wettschereck, D. (1996). Relational instance-based learning. In *Proceedings of the 13th International Conference on Machine Learning*, pages 122–130.

Enright, A., Iliopoulos, I., Kyrpides, N., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90.

Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical report, HP Laboratories.

Fields, S. and kyu Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–222.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. *Nucleic Acids Research*, 38(suppl 1):D557–562.

Fogel, G. B. (2008). Computational intelligence approaches for pattern discovery in biological systems. *Briefings in Bioinformatics*, 9(4):307–316.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the 15th International Conference on Machine Learning*, pages 144–151. Morgan Kaufmann Publishers Inc.

Fraser, H. B., Hirsh, A. E., Wall, D. P., and Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9033–9038.

Freund, Y. and Mason, L. (1999). The Alternating Decision Tree Learning Algorithm. In *Proceedings of the 16th International Conference on Machine Learning*, pages 124–133. Morgan Kaufmann Publishers Inc.

Friedberg, I. (2006). Automated protein function prediction - the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, 29(2-3):131–163.

Frohlich, H., Fellmann, M., Sültmann, H., Poustka, A., and Beissbarth, T. (2008). Predicting pathway membership via domain signatures. *Bioinformatics*, 24(19):2137–2142.

Gabaldón, T. and Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. *Cellular and Molecular Life Sciences*, 61(7):930–944.

Galperin, M. Y. and Fernández-Suárez, X. M. (2011). The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*.

García-Jiménez, B., Aler, R., Ledezma, A., and Sanchis, A. (2008a). Genetic Programming for predicting protein networks. In *Proceedings of the 11th Ibero-American Conference on Artificial Intelligence, IBERAMIA*, volume 5290 of *Lecture Notes in Artificial Intelligence*, pages 432–441. Springer.

García-Jiménez, B., Aler, R., Ledezma, A., and Sanchis, A. (2008b). Protein-protein functional association prediction using Genetic Programming. In *Proceedings of the International Conference on Genetic and Evolutionary Computation, GECCO*, pages 347–348. ACM.

García-Jiménez, B., Juan, D., Ezkurdia, I., Andrés-León, E., and Valencia, A. (2010a). Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS ONE*, 5(4):e9969.

García-Jiménez, B., Ledezma, A., and Sanchis, A. (2009). Modular Multi-Relational Framework for gene group function prediction. In *Proceedings of the 19th International Conference on Inductive Logic Programming. Poster*.

García-Jiménez, B., Ledezma, A., and Sanchis, A. (2010b). S.cerevisiae complex function prediction with Modular Multi-Relational Framework. In *Proceedings of the 23rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, IEA-AIE*, volume 6098 of *Lecture Notes in Artificial Intelligence*, pages 82–91.

García-Pedrajas, N. and de Haro García, A. (2008). *Output Coding Methods: Review and Experimental Comparison*, pages 327–344. Pattern Recognition Techniques, Technology and Applications. InTech, Austria.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., and Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*, pages 571–607. The Proteomics Protocols Handbook. Humana Press.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Gewehr, J. E., Szugat, M., and Zimmer, R. (2007). BioWeka extending the Weka framework for bioinformatics. *Bioinformatics*, 23(5):651–653.

Glaab, E., Baudot, A., Krasnogor, N., and Valencia, A. (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics*, 11(1):597.

Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J., and Giegerich, R. (2002). PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1):124–129.

Griep, S. and Hobohm, U. (2010). PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Research*, 38(suppl 1):D318–D319.

Groisman, R., Polanowska, J., Kuraoka, I., ichi Sawada, J., Saijo, M., Drapkin, R., Kisselev, A. F., Tanaka, K., and Nakatani, Y. (2003). The ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage. *Cell*, 113(3):357–367.

Gutmann, B. and Kersting, K. (2006). TildeCRF: Conditional Random Fields for Logical Sequences. In *Proceedings of the 15th European Conference on Machine Learning*, pages 174–185.

Gómez, M. J., Pazos, F., Guijarro, F. J., de Lorenzo, V., and Valencia, A. (2007). The environmental fate of organic pollutants through the global microbial metabolism. *Molecular Systems Biology*, 3:114.

Han, K., Park, B., Kim, H., Hong, J., and Park, J. (2004). HPID: The Human Protein Interaction Database. *Bioinformatics*, 20(15):2466–2470.

Hardison, R. C. (2003). Comparative genomics. *PLoS Biology*, 1(2):e58.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., and Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, 7(Suppl 1):S4.

Hawkins, T. and Kihara, D. (2007). Function prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology*, 5(1):1–30.

Herman, D., Ochoa, D., Juan, D., Lopez, D., Valencia, A., and Pazos, F. (2011). Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, 12(1):363.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004a). The HUPO PSI's molecular interaction format[mdash]a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004b). IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32(Database issue):452–5.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). Selection of representative protein data sets. *Protein Science*, 1(3):409–417.

Hochstrasser, M. (2009). Origin and function of ubiquitin-like proteins. *Nature*, 458(7237):422–429.

Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics*, 36(7):664.

Horton, P. and Nakai, K. (1997). Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 5, pages 147–152.

Huang, H., Hu, Z. Z., Arighi, C. N., and Wu, C. H. (2007). Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Frontiers in Bioscience: a Journal and Virtual Library*, 12:5071–5088.

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Research*, 37(suppl 1):690–697.

Huber, W., Carey, V. J., Long, L., Falcon, S., and Gentleman, R. (2007). Graphs in molecular biology. *BMC Bioinformatics*, 8 Suppl 6:S8.

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., Kovitz, B. L., Schilstra, M. J., Funahashi, A., Doyle, J. C., and Kitano, H. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Systems Biology*, 1(1):41–53.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(suppl 1):D211–215.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems Biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372.

Ihara, M., Yamamoto, H., and Kikuchi, A. (2005). SUMO-1 modification of PIASy, an E3 ligase, is necessary for PIASy-dependent activation of Tcf-4. *Molecular and Cellular Biology*, 25(9):3506–3518.

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., and Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. *Methods in Molecular Biology*, 593:25–48.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574.

Ivanovska, A., Vens, C., Colbach, N., Debeljak, M., and Dzeroski, S. (2008). The feasibility of co-existence between conventional and genetically modified crops: Using machine learning to analyse the output of simulation models. *Ecological Modelling*, 215(1-3):262–271.

Jansen, R. and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology*, 7(5):535–545.

Jassal, B. (2011). Pathway annotation and analysis with Reactome: the solute carrier class of membrane transporters. *Human Genomics*, 5(4):310–315.

Jensen, L. J. (2002). *Prediction of Protein Function from Sequence Derived Protein Features*. PhD thesis, Technical University of Denmark, Lyngby, Denmark.

Jensen, L. J. and Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics*, 27(24):3331–3332.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. (2002a). Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology*, 319(5):1257–1265.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. (2002b). Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology*, 319(5):1257–1265.

Jensen, L. J., Gupta, R., Staerfeldt, H. H., and Brunak, S. (2003a). Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, 19(5):635–642.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–416.

Jensen, L. J., Ussery, D. W., and Brunak, S. (2003b). Functionality of system components: Conservation of protein function in protein feature space. *Genome Research*, 13(11):2444–2449.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*.

John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann.

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, 105(3):934–939.

Juncker, A., Jensen, L. J., Pierleoni, A., Bernsel, A., Tress, M., Bork, P., von Heijne, G., Valencia, A., Ouzounis, C., Casadio, R., and Brunak, S. (2009). Sequence-based feature prediction and annotation of proteins. *Genome Biology*, 10(2):206.

Junker, B. H. and Schreiber, F. (2008). *Analysis of Biological Networks*. Wiley.

Kaelbling, L. P., Littman, M. L., and Moore, A. P. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):354–7.

Karp, P. D., Paley, S., and Romero, P. (2002). The Pathway Tools software. *Bioinformatics*, 18(suppl 1):S225–S232.

Kell, D. B. and Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1):99–105.

Kemper, B., Matsuzaki, T., Matsuoka, Y., Tsuruoka, Y., Kitano, H., Ananiadou, S., and Tsujii, J. (2010). PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381.

Keogh, E. J. (1999). Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches. In *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, pages 225–230.

Kersting, K. and Dick, U. (2004). Balios - the engine for bayesian logic programs. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 3202 of *Lecture Notes in Computer Science*, pages 549–551.

Kersting, K., Raedt, L. D., and Raiko, T. (2006). Logical Hidden Markov Models. *Journal of Artificial Intelligence Research*, 25:425–456.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., and Karp, P. D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33(Database issue):334–7.

Kim, J. H., Park, S. M., Kang, M. R., Oh, S. Y., Lee, T. H., Muller, M. T., and Chung, I. K. (2005). Ubiquitin ligase MKRN1 modulates telomere length homeostasis through a proteolysis of hTERT. *Genes & Development*, 19(7):776–781.

Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L., Coviello, G. M., Wright, W. E., Weinrich, S. L., and Shay, J. W. (1994). Specific association of human telomerase activity with immortal cells and cancer. *Science*, 266(5193):2011–2015.

King, R. D., Karwath, A., Clare, A., and Dehaspe, L. (2000a). Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining. *Yeast*, 1(4):283–293.

King, R. D., Karwath, A., Clare, A., and Dehaspe, L. (2001). The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, 17(5):445–454.

King, R. D., Karwath, A., Clare, A., and Dephaspe, L. (2000b). Genome scale prediction of protein functional class from sequence using data mining. In *Proceedings of the 6th ACM SIG International conference on Knowledge Discovery and Data Mining*, pages 384–389. ACM.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., and Clare, A. (2009). The automation of science. *Science*, 324(5923):85–89.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., and Oliver, S. G. (2004a). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252.

King, R. D., Wise, P. H., and Clare, A. (2004b). Confirmation of data mining based predictions of protein function. *Bioinformatics*, 20(7):1110–1118.

Kitano, H. (2000). Perspectives on systems biology. *New Generation Computing*, 18(3):199–216.

Kitano, H. (2001). *Foundations of systems biology*. MIT Press.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

Klingstrom, T. and Plewczynski, D. (2011). Protein-protein interaction and pathway databases, a graphical review. *Briefings in Bioinformatics*, 12(6):702–713.

Korcsmaros, T., Szalay, M., Rovo, P., Palotai, R., Fazekas, D., Lenti, K., Farkas, I., Csermely, P., and Vellai, T. (2011). Signalogs: Orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE*, 6(5):e19240.

Koza, J. (1992). *Genetic Programming*. MIT Press.

Koza, J. (1994). *Genetic Programming II*. MIT Press.

Krallinger, M., Leitner, F., and Valencia, A. (2010). Analysis of biological processes and diseases using text mining approaches. *Methods in Molecular Biology*, 593:341–382.

Kramer, S. (1996). Structural regression trees. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 812–819.

Kramer, S. and Pfahringer, B. (2005). *Proceedings of the 15th International Conference on Inductive Logic Programming, ILP 2005, Bonn, Germany, August 10-13*, volume 3625. Springer.

Kramer, S., Pfahringer, B., and Helma, C. (1997). Stochastic propositionalization of non-determinate background knowledge. In *Proceedings of the 8th International Conference on Inductive Logic Programming*, volume 1446 of *Lecture Notes in Artificial Intelligence*, pages 80–94. Springer.

Landwehr, N., Kersting, K., and Raedt, L. D. (2005). nFOIL: Integrating Naive Bayes and FOIL. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 795–800.

Lankester, E. R. (1870). On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Annuals Magazine of Natural History*, 6:43.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., and Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112.

Lavrac, N., Dzeroski, S., and Grobelnik, M. (1991). Learning nonrecursive definitions of relations with LINUS. In *Proceedings of the 5th European Working Session on Learning*, volume 482 of *Lecture Notes in Artificial Intelligence*, pages 265–281. Springer.

Lee, B., Shin, M., Oh, Y., Oh, H., and Ryu, K. (2009). Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome Science*, 7(1):27.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews. Molecular Cell Biology*, 8(12):995–1005.

Leon, E. A., Ezkurdia, I., García-Jiménez, B., Valencia, A., and Juan, D. (2009). EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Research*, 37(suppl 1):D629–D635.

Leskovec, J. (2008). *Dynamics of large networks*. PhD thesis, Machine Learning Department, Carnegie Mellon University.

Li, N. (2008). Platelet-lymphocyte cross-talk. *Journal of Leukocyte Biology*, 83(5):1069–1078.

Li, W. and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Likic, V. A., McConville, M. J., Lithgow, T., and Bacic, A. (2010). Systems biology: The next frontier for bioinformatics. *Advances in Bioinformatics*, 2010:ID268925.

Linghu, B., Snitkin, E., Hu, Z., Xia, Y., and DeLisi, C. (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology*, 10(9):R91.

Lloyd, J. W. (1987). *Foundations of logic programming*. Springer, New York, NY, USA.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.

Lu, L. J., Sboner, A., Huang, Y. J., Lu, H. X., Gianoulis, T. A., Yip, K. Y., Kim, P. M., Montelione, G. T., and Gerstein, M. B. (2007). Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends in Biochemical Sciences*, 32(7):320–331.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953.

Luciano, J. and Stevens, R. (2007). e-Science and biological pathway semantics. *BMC Bioinformatics*, 8:S3.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Engineering*, 10(11):1241–1248.

López-Bigas, N. and Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, 32(10):3108–3114.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

Mahler, S., Robilliard, D., and Fonlupt, C. (2005). Tarpeian bloat control and generalization accuracy. In *Proceedings of the 8th European Conference on Genetic Programming*, pages 203–214.

Marc, V. (2005). Interactome modeling. *FEBS Letters*, 579(8):1834–1838.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics*, 8 Suppl 6:S5.

Mateos, A., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M., and Stolovitzky, G. (2002). Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Research*, 12(11):1703–1715.

Mathe, C., Sagot, M.-F., Schiex, T., and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19):4103–4117.

Matsuda, H. (1995). Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. In *Proceedings of the Genome Informatics Workshop*, volume 6, pages 19–28. Universal Academy Press.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl 1):D619–622.

McShan, D. C., Rao, S., and Shah, I. (2003). PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, 19(13):1692–1698.

Middendorf, M., Kundaje, A., Wiggins, C., Freund, Y., and Leslie, C. (2004). Predicting genetic regulatory response using classification. *Bioinformatics*, 20(suppl 1):i232–i240.

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

Morett, E., Korbel, J., Rajan, E., SaabRincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B., and Bork, P. (2003). Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, 21(7):790–795.

Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4):295–318.

Muggleton, S. (1995). Inverse entailment and progol. *New Generation Computing. Special issue on Inductive Logic Programming*, 13(3-4):245–286.

Muggleton, S., Chen, J., Watanabe, H., Dunbar, S., Baxter, C., Currie, R., Salazar, J. D., Taubert, J., and Sternberg, M. (2010). Variation of background knowledge in an industrial application of ILP. In *Proceedings of the 20th International Conference on Inductive Logic Programming,*, volume 6489 of *Lecture Notes in Artificial Intelligence*, pages 158–170. Springer.

Nagai, S., Davoodi, N., and Gasser, S. M. (2011). Nuclear organization in genome stability: SUMO connections. *Cell Research*, 21(3):474–485.

Nassif, H., Al-Ali, H., Khuri, S., Keirouz, W., and Page, D. (2009). An Inductive Logic Programming approach to model and classify hexose binding sites. In *Proceedings of the 19th International Conference on Inductive Logic Programming*, volume 5989 of *Lecture Notes in Artificial Intelligence*, pages 149–165.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Noble, D. (2006). *The music of life : biology beyond the genome*. Oxford University Press.

Novere, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villeger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741.

Ochoa, D. and Pazos, F. (2010). Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, 26(10):1370–1371.

Ooi, H. S., Schneider, G., Lim, T.-T., Chan, Y.-L., Eisenhaber, B., and Eisenhaber, F. (2010). *Biomolecular Pathway Databases*, volume 609 of *Data Mining Techniques for the Life Sciences*, pages 129–144. Humana Press.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., Rivas, J. D. L., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25(8):894–898.

Ou, G. and Murphey, Y. L. (2007). Multi-class pattern classification using neural networks. *Pattern Recognition*, 40(1):4–18.

Ouali, M. and King, R. D. (2000). Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, 9(6):1162–1176.

Ouzounis, C. A., Coulson, R. M., Enright, A. J., Kunin, V., and Pereira-Leal, J. B. (2003). Classification schemes for protein structure and function. *Nature Reviews. Genetics*, 4(7):508–519.

Ouzounis, C. A. and Valencia, A. (2003). Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics*, 19(17):2176–2190.

Page, D. and Craven, M. (2003). Biological applications of multi-relational data mining. *SIGKDD Explorations*, 5(1):69–79.

Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of Computational Biology: a Journal of Computational Molecular Cell Biology*, 9(2):401–411.

Pazos, F., Ranea, J. A. G., Juan, D., and Sternberg, M. J. E. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of Molecular Biology*, 352(4):1002–1015.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609–614.

Pazos, F. and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–227.

Pazos, F. and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, 27(20):2648–2655.

Pazos, F., Valencia, A., and Lorenzo, V. D. (2003). The organization of the microbial biodegradation network from a systems-biology perspective. *EMBO Reports*, 4(10):994–999.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443.

Peters, J. M., Franke, W. W., and Kleinschmidt, J. A. (1994). Distinct 19 S and 20 S subcomplexes of the 26 S proteasome and their distribution in the nucleus and the cytoplasm. *Journal of Biological Chemistry*, 269(10):7709–7718.

Peña-Castillo, L., Tasan, M., Myers, C., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W., Krumpelman, C., Tian, W., Obozinski, G., Qi, Y., Mostafavi, S., Lin, G., Berriz, G., Gibbons, F., Lanckriet, G., Qiu, J., Grant, C., Barutcuoglu, Z., Hill, D., Warde-Farley, D., Grouios, C., Ray, D., Blake, J., Deng, M., Jordan, M., Noble, W., Morris, Q., Klein-Seetharaman, J., Bar-Joseph, Z., Chen, T., Sun, F., Troyanskaya, O., Marcotte, E., Xu, D., Hughes, T., and Roth, F. (2008). A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, 9:S2.

Philippi, C., Loretz, B., Schaefer, U. F., and Lehr, C. M. (2010). Telomerase as an emerging target to fight cancer - opportunities and challenges for nanomedicine. *Journal of Controlled Release*, 146(2):228–240.

Plotkin, G. (1970). A note on inductive generalization. *Machine Intelligence*, 5:153–163.

Poli, R. (2001). General schema theory for Genetic Programming with subtree-swapping crossover. In *Proceedings of the 4th European Conference on Genetic Programming*, pages 143–159.

Poli, R. (2003). A simple but theoretically-motivated method to control bloat in Genetic Programming. In *Proceedings of the 6th European Conference on Genetic Programming*, pages 43–76.

Poli, R., Langdon, W., and Dignum, S. (2007). On the limiting distribution of program sizes in tree-based Genetic Programming. In *10th European Conference on Genetic Programming*, pages 193–204.

Prather, K. L. J. and Martin, C. H. (2008). De novo biosynthetic pathways: rational design of microbial chemical factories. *Current Opinion in Biotechnology*, 19(5):468–474.

Qi, Y., Bar-Joseph, Z., and F., J. J. K.-S. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, 63(3):490–500.

Quiles, M. C. (2005). *Integration of biological data systems, infrastructures and programmable tools*. PhD thesis, Universidad Autónoma de Madrid.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Quinlan, J. R. and Mostow, J. (1990). Learning logical definitions from relations. *Machine Learning*, pages 239–266.

Raedt, L. D. (1997). Logical settings for concept-learning. *Artificial Intelligence*, 95(1):187–201.

Raedt, L. D. (2008). *Logical and Relational Learning*. Springer.

Raedt, L. D. and Kersting, K. (2003). Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48.

Raedt, L. D., Kimmig, A., and Toivonen, H. (2007). ProbLog: A Probabilistic Prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2462–2467.

Raedt, L. D. and Laer, W. V. (1995). Inductive constraint logic. In *Proceedings of the 5th Workshop on Algorithmic Learning Theory*, volume 997 of *Lecture Notes in Computer Science*, pages 80–94.

Raval, A., Ghahramani, Z., and Wild, D. L. (2002). A bayesian network model for protein fold and remote homologue recognition. *Bioinformatics*, 18(6):788–801.

Re, M. and Valentini, G. (2009). Prediction of gene function using ensembles of SVMs and heterogeneous data sources. In *Applications of Supervised and Unsupervised Ensemble Methods*, volume 245 of *Studies in Computational Intelligence*, pages 79–91. Springer.

Rodríguez, J. M., Maietta, P., Ezkurdia, I., López, G., Wesselink, J.-J., Pietrelli, A., Valencia, A., and Tress, M. (2012). APPRIS: A system for annotating alternative splice isoforms. In *Proceedings of the 11th Spanish Symposium on Bioinformatics. Poster*.

Rojas, A., Juan, D., and Valencia, A. (2006). *Molecular Interactions: Learning form Protein Complexes*, volume 6 of *In Silico Technologies in Drug Target Identification and Validation*, chapter 8, pages 225–244.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, 60(12):2637–2650.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Genetics*, 19(1):55–72.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., and Mewes, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539–5545.

Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing*, volume 1. MIT Press, Cambridge, MA.

Sahami, M. (1996). Learning Limited Dependence Bayesian Classifiers. In *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining*, pages 335–338. AAAI Press.

Sakharkar, M. K., Sakharkar, K. R., and Pervaiz, S. (2007). Druggability of human disease genes. *The International Journal of Biochemistry and Cell Biology*, 39(6):1156–1164.

Saladrigas, M. V. (2006). Vocabulario inglés-español de bioquímica y biología molecular. *Panace@*, 7(24):265–275.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):449–51.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68.

Sato, T. and Kameya, Y. (2001). Parameter learning of logic programs for symbolic-statistical modeling. *Journal of Artificial Intelligence Research*, 15:391–454.

Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Sebban, M., Mokrousov, I., Rastogi, N., and Sola, C. (2002). A data-mining approach to spacer oligonucleotide typing of mycobacterium tuberculosis. *Bioinformatics*, 18(2):235–243.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, 3:88.

Sheng, Q., Moreau, Y., Smet, F. D., Marchal, K., and Moor, B. D. (2005). *Advances in Cluster Analysis of Microarray Data*, chapter 10, pages 153–173. Data Analysis and Visualization in Genomics and Proteomics. John Wiley & Sons, Ltd.

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38(suppl 1):D161–D166.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). BioMart - biological queries made easy. *BMC Genomics*, 10(1):22.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Smogorzewska, A. and de Lange, T. (2004). Regulation of telomerase by telomeric proteins. *Annual Review of Biochemistry*, 73(1):177–208.

Solé, R. (2009). *Redes complejas. Del genoma a Internet*. Tusquets, España.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692.

Srinivasan, A. (2007). Aleph: A Learning Engine for Proposing Hypotheses. The Aleph Manual. Technical report, Computing Laboratory, Oxford University.

Srinivasan, A., King, R. D., and Bristol, D. W. (1999). An assessment of submissions made to the predictive toxicology evaluation challenge. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 270–275.

Srinivasan, A., Muggleton, S. H., Sternberg, M. J. E., and King, R. D. (1996). Theories for mutagenicity: a study in first-order and feature-based induction. *Artificial Intelligence*, 85(1-2):277–299.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–539.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews. Genetics*, 2(7):493–503.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the Perceptron algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10(9):2997–3011.

Struyf, J., Dzeroski, S., Blockeel, H., and Clare, A. (2005). Hierarchical multi-classification with predictive clustering trees in functional genomics. In *Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence*, pages 272–283.

Takeyama, K., Aguiar, R. C. T., Gu, L., He, C., Freeman, G. J., Kutok, J. L., Aster, J. C., and Shipp, M. A. (2003). The BAL-binding protein BBAP and related deltex family members exhibit ubiquitin-protein isopeptide ligase activity. *Journal of Biological Chemistry*, 278(24):21930–21937.

Tarca, A. L., Carey, V. J., wen Chen, X., Romero, R., and Draghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, 3(6):e116.

Tetko, I. V., Rodchenkov, I. V., Walter, M. C., Rattei, T., and Mewes, H.-W. (2008). Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information. *Bioinformatics*, 24(5):621–628.

Thomsen, R. (2007). *Protein-Ligand Docking with Evolutionary Algorithms*, pages 167–195. Computational Intelligence in Bioinformatics. John Wiley & Sons, Inc.

Todorovski, L., Blockeel, H., and Dzeroski, S. (2002). Ranking with predictive clustering trees. In *M*.

Trajkovski, I., Zelezny, F., Lavrac, N., and Tolar, J. (2008). Learning relational descriptions of differentially expressed gene groups. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(1):16–25.

Tran, T., Satou, K., and Ho, T. (2005). Using Inductive Logic Programming for predicting protein-protein interactions from multiple genomic data. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 3721 of *Lecture Notes in Artificial Intelligence*, pages 321–330. Springer.

Tress, M. L., Wesselink, J.-J., Frankish, A., López, G., Goldman, N., Löytynoja, A., Massingham, T., Pardi, F., Whelan, S., Harrow, J., and Valencia, A. (2008). Determination and validation of principal gene products. *Bioinformatics*, 24(1):11–17.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., Zhang, H., and Consortium, T. F. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, 37(suppl 1):D555–559.

Valafar, F. (2002). Pattern recognition techniques in microarray data analysis: a survey. *Annals of the New York Academy of Sciences*, 980:41–64.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3):368–373.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience, New York.

Vens, C. (2007). *Complex Aggregates in Relational Learning*. PhD thesis, Department of Computer Science, K.U.Leuven, Leuven, Belgium.

Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., and Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(suppl 1):D433–D437.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

Wang, K., Ussery, D. W., and Brunak, S. (2009). Analysis and prediction of gene splice sites in four Aspergillus genomes. *Fungal Genetics and Biology*, 46(1, Supplement 1):S14–S18.

Webb, G. I., Boughton, J. R., and Wang, Z. (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*, 58(1):5–24.

Wilkinson, M. D. and Links, M. (2002). BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics*, 3(4):331–341.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition.

Woznica, A. (2006). Relational Weka. Technical report, University of Geneva.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530.

Xu, R., Anagnostopoulos, G. C., and Wunsch, D. C. (2007a). *Hybrid of Neural Classifier and Swarm Intelligence in Multiclass Cancer Diagnosis with Gene Expression Signatures*, pages 1–20. Computational Intelligence in Bioinformatics. John Wiley & Sons, Inc.

Xu, R., Wunsch, D. C., and Frank, R. L. (2007b). Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):681–692.

Xu, Y. (2011). Chemistry in human telomere biology: structure, function and targeting of telomere DNA/RNA. *Chemical Society Reviews*, 40(5):2719–2740.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1):69–90.

Ye, J., McGinnis, S., and Madden, T. L. (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Research*, 34(suppl 2):W6–9.

Yi, T.-M. and Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*, 232(4):1117–1129.

Yoon, B. J. (2009). Hidden Markov Models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the Yeast interactome network. *Science*, 322(5898):104–110.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. *Genome Research*, 14(6):1107–1118.

Zheng, Z. and Webb, G. I. (2000). Lazy learning of bayesian rules. *Machine Learning*, 41(1):53–84.

Zongker, D. and Punch, B. (1998). *lil-gp*: Genetic Programming System. Technical report, Michigan State University.