UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

“Evaluación funcional del interactoma en la red metabólica de *Saccharomyces cerevisiae*”

TESIS QUE PARA OBTENER EL GRADO DE:

LICENCIADO EN INVESTIGACIÓN BIOMÉDICA BÁSICA

PRESENTA:

RAÚL ANTONIO ORTIZ MERINO

ASESOR DE TESIS:

DR. GABRIEL DEL RÍO GUERRA

CIUDAD UNIVERSITARIA Septiembre, 2010

AGRADECIMIENTOS

Esta tesis es el punto en donde culmina mi formación académica a nivel superior. Debido a que fue un proceso que abarcó aspectos tanto académicos como familiares y sociales procuraré agradecerle a quienes participaron en cualquiera de ellos sin que la precedencia en orden signifique precedencia en importancia.

Gracias a mi familia, específicamente a mi Nane que es el motor que nos ha impulsado a tantos. A mi mamá, a quien le agradezco por obvias razones, a mi hermano Carlos por ser el mejor y el peor de mis ejemplos, y a mi papá, el a veces desconocido que mejor me conoce y que siempre me ha apoyado a pesar de todo. Igual le agradezco tanto a mi tío Alejandro Toledo como a sus retoños Gaby y Pancho que me han inspirado por su dedicación y desempeño sobresaliente, espero algún día poderles retribuir el favor. Mi tío Felipe tampoco se puede quedar afuera, le agradezco por alentarme cada que hay oportunidad y sea esta una oportunidad para rellenar el cofre del tesoro en memoria de mi tío Alejandro. No menciono a todos los demás hermanos tíos, primos y sobrinos por la extensión de la lista y no me gustaría hacer más omisiones de las necesarias.

Agradezco a la Universidad Nacional Autónoma de México que en su esfuerzo y dedicación ha logrado formar a muchísimos estudiantes que, como yo, tienen en sus manos el progreso tanto de la misma universidad como de la nación.

Tampoco puedo dejar de agradecer a todos aquellos que estuvieron ahí para permitirme aprender con su ejemplo y que se han tomado la molestia de enseñarme. Entre ellos incluyo al Dr. Juan Luis Rendón Gómez, a la Dra. Rocío Salceda Sacanelles y al Dr. Gabriel del Río así como a los alumnos adscritos a sus respectivos laboratorios a quienes no hubiera tenido el gusto de conocer (y entender) sin la Licenciatura en Investigación Biomédica Básica y su cuerpo tanto de alumnos como de profesores y personal administrativo. A mis compañeros de la licenciatura los menciono aparte para hacer énfasis en el hecho de que gracias a ustedes prefiero caminar entre gigantes antes de sobresalir entre mediocres.

Al Adrián por el hospedaje, los viajes y por ser el vínculo con el ambiente genómico y sus perspectivas.

Igualmente no hubiera llegado a este punto sin El Mostro. Toda mi vida he querido pertenecer a un equipo como este, gracias por dejarme ser su amigo, su compañero y su hermano. Gracias por darme chance de desafanarme de la academia y, aún fuera de ella, forzarme a ser profesional.

Otra lección fundamental se la debo al buen José Luis Becherell Robles, gracias por hacer que me diera cuenta de que no estamos aquí para siempre y que hay que ser terco para aferrarse hasta las últimas, espero hacer que valga la pena.

Gracias a lo conocido por que nos distingue como humanos y a lo desconocido por ponernos a prueba.

"…gracias a la vida que me ha dado tanto…"

Mercedes Sosa
cantora argentina
(1935-2009)

ABREVIATURAS

ABC – **A**rea **B**ajo la **C**urva
ADN – **A**cido **d**esoxi**r**ribonucleico
ARN – **A**cido **r**ibo**n**ucleico
ARNm – **ARN m**ensajero
BioGRID – **Bio**logical **G**eneral **R**epository for **I**nteraction **D**atasets
CG – **C**omponente **G**igante
Clos – Cercanía (**Clos**eness)
Cluster – Coeficiente de empacamiento (**Cluster**ing coefficient)
Deg – Grado (**Deg**ree)
DIP – **D**atabase of **I**nteracting **P**roteins
EC – Clasificación enzimática (**E**nzyme **c**lassification)
ES – **E**rror e**s**tándar
EMA – Análisis de modo elemental (**E**lementary **M**ode **A**nalysis)
Excen – **Excen**tricidad
FBA – Análisis de balance de flujos (**F**lux **B**alance **A**nalysis)
FN – **f**alsos **n**egativos
FP – **F**alsos **p**ositivos
IC – **I**ntervalo de **C**onfianza
IntNets – Redes de **IM** (**Int**eractome **Net**work**s**)
Int.Met – Redes metabólicas y de **IM** (**Int**eractome and **met**abolic **net**works)
IM – **I**nteracciones **m**oleculares
IPP – **I**nteracciones **p**roteína-**p**roteína
KEGG – **K**yoto **E**nciclopedia of **G**enes and **G**enomes
mDist – Distancia promedio (**m**ean **dist**ance)
meanClosenessCent – cercanía promedio
meanClusteringCoef – coeficiente de empacamiento promedio
meanDegree – grado promedio
meanDistance – promedio de **mDist**
meanExcentricity – excentricidad promedio
meanSphereDegree – grado esférico promedio
MetNets – Redes metabólicas (**Met**abolic **net**works)
minE – **Mín**imo **e**rror
MOMA – Minimización de ajuste metabólico (**M**inimization **O**f **M**etabolic **A**djustment)
OCCI – Indice general de la centralidad de cercanía (**O**verall **C**loseness **C**entrality **I**ndex)
REMG – **R**ich On **E**ssential **M**etabolic **G**enes
ROC – **R**eceiver **O**perating **C**haracteristic
S – Intermediación (betweenes**s**)
SGDP – **S**accharomyces **G**enome **D**eletion **P**roject
sphereDeg – Grado esférico (**sphere deg**ree)
TN – verdaderos negativos (**t**rue **n**egatives)
TP – Verdaderos positivos (**t**rue **p**ositives)
trav – Traversity
travA – TraversityA
YIPD – **Y**east **I**nteracting **P**roteins **D**atabase

ÍNDICE

RESUMEN

La información obtenida a partir de distintas aproximaciones “ómicas” (genómica, proteómica, transcriptómica, metabolómica, entre otras) requiere del desarrollo de estrategias experimentales que permitan integrarlos, además de herramientas para asimilar la información en forma de modelos.

Existen distintas aproximaciones para representar los mecanismos moleculares de éstos sistemas pero, en cualquiera que sea el caso, se debe mostrar su capacidad para reproducir características del sistema al que modelan. La reconstrucción y el análisis de distintas representaciones de la red metabólica de la levadura *Saccharomyces cerevisiae* ha permitido predecir el conjunto de genes metabólicos experimentalmente descritos como esenciales (o críticos) para la supervivencia de este organismo. El proceso consiste en modelar las relaciones genéticas del organismo en forma de red y utilizar características relativas a cada uno de sus elementos como un parámetro para diferenciar los genes esenciales de los que no los son. Tal análisis ya ha sido utilizado sobre distintas redes metabólicas mostrando una eficiencia limitada en la predicción de genes críticos en el metabolismo.

Con este trabajo se pretende evaluar distintas medidas de centralidad como índices de la importancia relativa de un gen y su capacidad para clasificar genes esenciales en diferentes modelos del metabolismo. Esto al analizar la suposición de que las relaciones de una red metabólica pueden ser enriquecidas mediante la incorporación de diversos tipos de IM reportados en distintas bases de datos disponibles públicamente. Esta propuesta implica que los genes involucrados en el metabolismo celular mediante distintos tipos de IM deberían formar una red cuyos elementos centrales incluyen genes esenciales. Como resultado, se generaron varios modelos capaces de reproducir los genes esenciales para el metabolismo con mayor probabilidad que los anteriores.

I. INTRODUCCIÓN

I.1 Biología de sistemas

I.1.a La era genómica de la biología

El desarrollo de la secuenciación de genomas como un procedimiento automático y de alto rendimiento ha sido seguido por innovaciones tecnológicas que proveen mediciones a gran escala de las distintas especies moleculares existentes dentro de la célula. Las descripciones para determinado tiempo y/o condición celular que pueden arrojar estos tipos de datos, referidos como "ómicos", pueden ser clasificadas en componentes, interacciones y estados funcionales.

En la categoría de componentes se busca información acerca del contenido molecular específico de la célula e incluye a la genómica, transcriptómica, proteómica y metabolómica describiendo genes, elementos transcripcionales[*], proteínas y metabolitos respectivamente, con la localizómica tratando de identificar su disposición subcelular. La información sobre interacciones pretende especificar las conexiones entre las distintas especies moleculares definiendo la interactómica que puede tener distintos enfoques dependiendo del tipo de molécula cuya interacción es reportada, siendo las de tipo proteína-DNA y proteína-proteína las más comunes. Finalmente, al examinar los estados funcionales se espera obtener información sobre el comportamiento del sistema representado por fenotipos, por ello ha sido llamado fenómica, e incluye estrategias de alto rendimiento para determinar la adecuación celular en respuesta a perturbaciones genéticas y/o ambientales entre las que se encuentran análisis sistemáticos de mutación, tamizajes farmacológicos y de interferencia mediada por RNA, así como estudios de flujos metabólicos (Joyce y Palsson 2006).

[*] Factores de transcripción y ARNm.

La secuenciación de genomas fue uno de los primeros esfuerzos en la biología por practicar un enfoque científico basado en descubrimientos con el objetivo de definir todos los elementos de un sistema y crear bases de datos que contengan dicha información. Esta nueva aproximación científica, que ha catalizado el desarrollo de las distintas disciplinas genómicas, está fortaleciendo la visión de que la biología es una ciencia informacional proveyendo nuevas herramientas de alto rendimiento para monitorear y perturbar los sistemas biológicos, estimulando la creación de nuevos métodos que permitan hacer esto de forma sistemática (Ideker 2001) y permitiendo colectar conjuntos de datos en forma comprensible para ganar información sobre los componentes de los sistemas biológicos (Kitano 2002). De tal forma, actualmente se sabe que el almacenaje y procesamiento de la información necesaria para los distintos programas celulares reside en varios niveles de organización: genoma, transcriptoma, proteoma y metaboloma, entre los que existe un flujo bidireccional (Oltvai y Barabási 2002).

I.1.b Información biológica.

La información biológica tiene varias características importantes:

- Opera con niveles de información múltiples y jerárquicos: ADN→ ARNm→ proteínas→ interacciones[*]→ vías→ redes[†]→ células→ tejidos o comunidades celulares→ organismos→ poblaciones→ ecologías.

- Es procesada en redes complejas. Las aproximaciones genómicas han impulsado la visión de que los sistemas biológicos están compuestos fundamentalmente de dos tipos de información: genes, que codifican la maquinaria molecular responsable de la ejecución de las funciones vitales, y

[*] entre distintas especies moleculares como macromoléculas y metabolitos.
[†] ver sección V.1.a para una definición

redes de interacciones regulatorias, que especifiquen cómo se expresan tales genes.

- Las redes que la integran suelen ser robustas. Esto significa que varias perturbaciones sencillas no tendrían grandes efectos sobre ellas pero contienen elementos clave sobre los que alguna perturbación pudiera tener efectos profundos; éstos ofrecen blancos poderosos para la comprensión y manipulación del sistema.

Las bases de datos biológicos han surgido como depósitos centrales para la gran cantidad de información experimental sirviendo para cubrir las demandas impuestas por la genómica funcional y otras aproximaciones a nivel de sistemas. Aunque las bases de datos para secuencias, ya sea de ácidos nucleicos o de proteínas, siguen considerándose como las más grandes, más utilizadas y mejor mantenidas, ha habido una explosión en el interés por bases de datos que guarden otros tipos de información (Ideker 2001). Inclusive, cada año se publica un número especial en la revista Nuclear Acids Research describiendo las principales bases de datos nuevas o actualizadas evidenciando la creciente cantidad de éstas.

Existen varios ejemplos de bases de datos con distintos tipos de IM, como el repositorio biológico general para conjuntos de datos de interacciones BioGRID (Stark 2006), la base de datos DIP de proteínas que interactúan (Salwinski 2004), la base de datos y catálogo de herramientas IntAct para el almacenaje, presentación y análisis de interacciones protéicas del European Bioinformatics Institute (Hermjakob 2004; S. Kerrien 2007), la base de datos MPact de IPP en levadura del Munich Information Center for Protein Sequences (Güldener 2006) y resultados de análisis sistemáticos a gran escala mediante dobles híbridos YIPD (Ito 2000; Ito 2001) y YPLND (Uetz 2000), el modelo probabilístico de red genética funcional YeastNet (Lee 2007) que representa interacciones genéticas diversas y KEGG que presenta información de vías metabólicas (Ogata 1999).

Una vez acumulados, dichos datos necesitan mantenimiento sistemático, anotación consistente, verificación y actualización. Por ello requieren del desarrollo de estrategias que permitan integrarlos para poder analizarlos esperando hacer descubrimientos, de sistemas computacionales, para guardar, catalogar y condensar los datos, además de herramientas para asimilar la información en forma de modelos buscando predecir comportamientos que puedan ser demostrados experimentalmente (Ideker 2001).

I.1.c La biología de sistemas

La visión de sistemas ha sido propuesta y probada desde hace ya varios años y se podría decir que se originó con el trabajo de Norbert Wiener (1965) quien propuso el concepto de cibernética desarrollando fórmulas matemáticas para describir sistemas fisiológicos. Un poco después el filósofo Ludwig Von Bertalanffy (1976) intentó establecer la teoría general de sistemas, pero es hasta ahora cuando se ha abierto la posibilidad de realizar estudios a nivel de sistemas usando moléculas biológicas. En este sentido existen tres tendencias tecnológicas principales: las técnicas de manipulación genética han aumentado su rendimiento al automatizarse y estandarizarse en varios órdenes de magnitud, la disponibilidad de secuencias genómicas completas ha estimulado el desarrollo de varios proyectos sistemáticos y, las tecnologías de disrupción genética en *trans* han permitido la aplicación de perturbaciones genéticas a un amplio rango de organismos (Ideker 2001).

Esta aparición-reaparición de la biología de sistemas ha provocado que su definición siga abierta pero puede considerarse como un dominio de la ciencia biológica que estudia la dinámica resultante de las interacciones de los componentes constitutivos de los sistemas biológicos bajo un enfoque de sistema (en oposición a la visión reduccionista del fenómeno biológico).

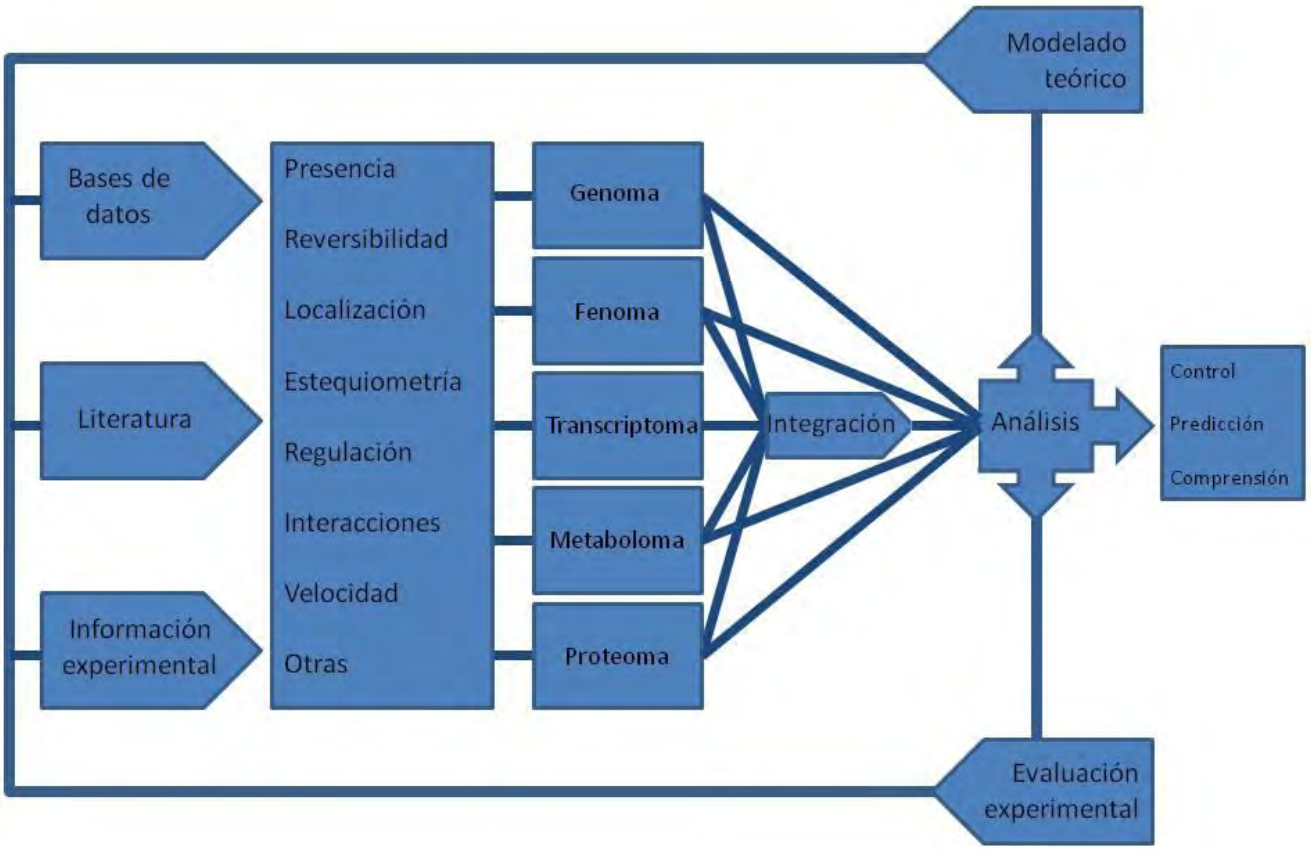
La biología de sistemas asume como tareas centrales a) reunir toda la información de cada uno de los distintos niveles de información de sistemas biológicos individuales en forma comprensible y, b) integrar esos datos, desplegarlos gráficamente y generar modelos matemáticos o computacionales que puedan ser útiles para predecir el comportamiento del sistema (Ideker 2001).

Esta disciplina no explora genes individuales o proteínas una a la vez como se ha hecho de forma exitosa durante los últimos años, en lugar de eso, investiga el comportamiento y las relaciones de los elementos de un sistema biológico particular. Por otra parte, al identificar todas las moléculas de un organismo se provee un catálogo de componentes individuales que no es suficiente por sí mismo para entender la complejidad de sistema de estudio (Kitano 2002; Palsson 2000) pero existen observaciones que apoyan el hecho de que las interacciones entre los componentes del sistema pueden proveer la mejor y más completa información sobre el comportamiento celular (Reed 2003).

Existen dos perspectivas complementarias en esta disciplina: una llamada “arriba-abajo” y otra “abajo-arriba”. La primera se enfrenta con grandes conjuntos de datos y pretende obtener una representación del comportamiento del sistema lo suficientemente amplia como para permitir el descubrimiento de patrones genéricos, buscando obtener poder predictivo sobre los mecanismos biológicos. En la segunda perspectiva se intenta deducir las propiedades emergentes de módulos[*] (o subsistemas) bien caracterizados y mecanísticamente detallados. Este último acercamiento empieza modelando el comportamiento de procesos individuales en partes manejables del sistema con el fin de crear modelos que puedan ser combinados hasta obtener un modelo que represente al sistema completo (Bruggeman y Westerhoff 2006).

[*] Un módulo es, por definición, una entidad discreta cuya función es separable de la de otros módulos. Esta visión implica que una fracción de los componentes celulares llevan a cabo una función casi autónoma. Los módulos pueden estar compuestos por distintos tipos de moléculas y presentan funciones discretas que resultan de las interacciones entre sus componentes (Hartwell 1999).

El factor común en ambas aproximaciones consiste en que usan un ciclo iterativo que inicia con datos experimentales y sigue con el análisis e integración de éstos terminando con la formulación y prueba de nuevas hipótesis de co- e inter- regulación. Tal ciclo se ilustra en el siguiente diagrama.



Ciclo iterativo de la biología de sistemas (Diagrama). Comienza con distintas fuentes de información que contienen diversos parámetros característicos de algún sistema. Tales datos representan los procesos celulares desde distintas perspectivas "ómicas" de manera que pueden ser analizados (y algunas veces integrados) buscando obtener control o poder predictivo y generando nuevos datos teóricos o experimentales. Esta nueva información puede ser incorporada al ciclo hasta alcanzar la comprensión del sistema.

I.2 Modelado teórico

I.2.a Redes: enfoques y alternativas

Las recientes tecnologías experimentales de alto rendimiento han permitido el desarrollo de nuevas perspectivas sobre las propiedades funcionales de la célula ayudando a revelar a detalle la composición molecular y complejidad de las celular (Xenarios y Eisenberg 2001). Esta explosión, tanto en la variedad como en la cantidad de información presenta varias complicaciones que limitan la utilidad de cada tipo de información por separado (Ideker 2001; De Keersmaecker 2006). Por ejemplo, las tecnologías de alto rendimiento producen datos en distintos formatos por lo que la habilidad de analizarlos en forma integral, lógica y simultánea permanece como un reto clave para el desarrollo de la biología de sistemas. Los modelos *in silico* representan una forma de lidiar con ello, entre otras formas, mediante la integración de diversos conjuntos de datos "omicos" (Palsson 2002). Dicha integración permite observar cómo se distribuyen las funciones celulares entre grupos de componentes heterogéneos que interactúan entre ellos dentro de grandes redes[*] (Hartwell 1999).

Desde este punto de vista, la información sobre componentes celulares incluye secuencias genómicas y macromoléculas con distintos tipos de interacciones que pueden servir para establecer las conexiones de las redes bioquímicas dentro de la célula. Los ejemplos más comunes de tales interacciones podrían ser las de tipo proteína-proteína (**IPP**) y proteína-ADN, útiles para modelar redes de interacción de proteínas y de regulación transcripcional. Otro tipo de interacciones son las químico-genéticas, éstas describen la relación entre los genes codificantes para las proteínas que utilizan un mismo metabolito como sustrato o como producto y por ello se utilizan para construir redes metabólicas. Entre las reconstrucciones de redes a escala celular se puede considerar a las metabólicas entre las mejor documentadas, mientras que las redes de regulación

[*] compuestas, por definición, de dos conjuntos: uno de nodos o vértices (elementos) y uno de aristas (relaciones); ver sección V.1.a

o de procesos de transcripción y/o traducción siguen en desarrollo, dejando otros tipos de redes, como las de señalización y de ARN pequeños, aún como prospectos.

El proceso de reconstrucción de redes comprende cuatro pasos fundamentales: 1)reconstrucción automática basada en información genómica, 2)curado manual por expertos, 3)conversión de la red en un modelo computacional mediante su representación matemática y, por último, 4)la integración de datos de alto rendimiento. Este procedimiento tiene potencial para evaluar los distintos conjuntos de datos a escala genómica en forma integral al colocarlos en un contexto funcional y estructurado (Feist 2009). De tal modo se espera que se puedan responder preguntas clave para la expansión del conocimiento biológico actual como la razón del surgimiento de funciones que no son predecibles a partir de interacciones o componentes individuales (redes de regulación o enfermedades multifactoriales entre otras).

Lograr una conexión entre los componentes e interacciones por un lado y los estados celulares (representados mediante estudios de expresión genética, niveles de metabolitos, flujo metabólico o caracterización fenotípica de cepas de deleción) en el otro permanece como un reto importante para la biología de sistemas (Herrgård 2006). Se trata de un desafío porque hasta el momento no existe evidencia de que el conocimiento biológico actual, aún en desarrollo, contenga listas completas[*], además de que los métodos de análisis e integración son diversos y producen resultados heterogéneos. Esto ha fortalecido la idea de que un sistema biológico no es simplemente un ensamble de genes y proteínas por lo que sus propiedades no pueden ser comprendidas totalmente con sólo hacer diagramas de sus interconexiones (Kitano 2002).

[*] de componentes, interacciones o estados celulares; si bien los genomas idealmente representan la totalidad de genes de algún organismo, existen casos en los que los elementos anotados no corresponden a algún componente celular determinado o representan moléculas sin función anotada.

I.2.b Análisis y evaluación funcional

El proceso de construcción de modelos matemáticos para procesos biológicos complejos debe presentar un ciclo de retroalimentación con dos componentes: experimental e *in silico* (ver diagrama en sección I.1.c). En este contexto, los modelos obtenidos mediante cualquier tipo de aproximación tienen que ser capaces de reproducir alguna característica observada en el sistema al que pertenecen[*]. Idealmente, se espera desarrollar un proceso iterativo en el que el modelo pueda ayudar a identificar objetivos que, al ser verificados, permitan realizar mejoras y actualizaciones. (Palsson 2000). Así, tras el establecimiento de un modelo *in silico* a escala celular para un organismo particular, se espera caracterizar las variaciones en el comportamiento de la red para probar el modelo y proponer experimentos adicionales (Palsson 2002). Una vez que se obtiene poder predictivo se puede llegar a principios generales que permitan comprender los sistemas biológicos. Entre los primeros ejemplos de tales principios se encuentra el control metabólico y jerárquico (Fell 1997) descubierto al utilizar aproximaciones como la teoría de sistemas bioquímicos (Savageau 1976).

Varios sistemas biológicos además de otros de naturaleza diversa ya han sido modelados y analizados en forma de redes (Oltvai y Barabási 2002), en donde la elección de un método para analizarlas depende del conocimiento disponible. Un análisis dinámico permite hacerse una idea de cómo los cambios en el comportamiento de algunos componentes (modificaciones a los parámetros mediante algún estímulo) pueden afectar otras partes del sistema y se han conducido algunos de estos estudios en distintas especies microbianas presentando distintas características. Un ejemplo consiste en el uso de modelos conducidos por datos y basados en restricciones permite definir las posibles funciones de una red al confinarlas en un espacio solución (Palsson 2000). Ese espectro en el que es posible que se encuentre la solución se puede reducir al incorporar perfiles de expresión, permitiendo identificar y analizar funciones

[*] habilidad que puede surgir, por ejemplo, al asumir la relación entre el fenotipo y la estructura molecular de la célula y ha sido referido como el paradigma estructura-función (del Río 2009).

celulares. En tal caso, cuando se toman en cuenta la abundancia o concentración de los componentes celulares se puede obtener cierta resolución temporal al definir las constantes de velocidad del sistema como un todo en lugar de modelar eventos bioquímicos individuales.

En contraste, un estudio de estado estacionario puede ser realizado usando solamente una estructura de red sin tener que incluir parámetros tan específicos (Kitano 2002). Además, analizando la estructura de las redes se pueden integrar, hacer comparaciones, buscar modularidad (complejos y motivos), tratar de predecir funciones y hacer estadística topológica (Zhang 2007). Por ejemplo, es posible estudiar las propiedades globales de las redes de IM mediante la teoría de gráficos o redes (Sawinski y Eisenberg 2003) y definir varias medidas de centralidad para cuantificar la importancia relativa de cada elemento dentro de la red, ya sea en forma local[*] o global[†] (Thibert 2005). Algunas de ellas ya han sido usadas para evaluar distintos tipos de redes en varios sistemas (Acencio y Lemke 2009; Albert 2005; Friedel y Zimmer 2006; Jeong 2001) y su utilidad ha sido revisada en varias ocasiones (Barabási y Oltvai 2004; Albert 2005).

La identificación los genes esenciales (o críticos) que, cuando están ausentes, confieren un fenotipo letal ha sido descrita como “la tarea más importante de la validación de blancos genómicos” (Chalker y Lunsford 2002) ya que éstos pueden ser utilizados como punto de referencia para evaluar las funciones representadas en distintos modelos. Sin embargo, su identificación experimental es difícil y lenta incluso en los organismos más simples y más estudiados, además de que se trata de fenotipos relacionados a condiciones ambientales como el medio de crecimiento. Por ello, existen aproximaciones que buscan identificar estos genes y van más allá de los métodos experimentales directos mediante el uso de modelos teóricos. Entre ellos se encuentran la genómica comparativa (Arigoni 1998), el uso de estadística Bayesiana

[*] en el sentido de que sus valores dependen solamente del nodo en cuestión y de aquellos con los que mantiene aristas

[†] cuyos valores dependen de la estructura general de la red

(Lamichhane 2003), de inteligencia artificial (Seringhaus 2006) y de análisis de centralidad (del Río 2009).

Ya se ha estudiado si existe una relación entre la estructura de las redes y propiedades funcionales de las proteínas como su esencialidad. Una de las primeras y más importantes conexiones en este sentido fue realizada al ordenar todas las proteínas de la base de datos DIP según el número de aristas incidentes sobre un vértice buscando correlación con el efecto fenotípico de su remoción individual. De tal forma, las proteínas altamente conectadas dentro de una red de IPP resultaron tener una probabilidad de ser esenciales hasta tres veces mayor que las proteínas con pocas interacciones (Jeong 2001).

Dicha correlación fue confirmada por varios estudios posteriores aunque poco se sabía sobre su causa. Jeong y colaboradores sugirieron que la sobre-representación de proteínas esenciales entre los nodos con alto grado se debía a su papel manteniendo la conectividad[*] de la red. Sin embargo, surgieron explicaciones alternativas como el hecho de que al aumentar las conexiones de una proteína aumenta la probabilidad de que estén implicadas en más de una interacción hasta llegar el punto en que se vuelven esenciales (He y Zhang 2006). En el 2008 se realizó un intento sistemático por comprobar el determinante topológico de la esencialidad y se encontró que la esencialidad de un gen tiene que ver con sus conexiones locales más que con el mantenimiento de la conectividad de la red. Además, también se rechazó la propuesta de He y colaboradores proponiendo una alternativa en la que los complejos protéicos están compuestos casi homogéneamente por proteínas esenciales o por no esenciales (Zotenko 2008).

[*] Ver definición en la sección V.1.a.

I.3 Modelos experimentales en biología de sistemas

I.3.a Organismos modelo

La verificación o falsificación de hipótesis requiere de conjuntos de datos lo más completos posibles. El hecho de que los organismos multicelulares contengan distintos tipos de células con gran variedad y número de moléculas provoca que la biología de sistemas deba empezar a probarse en cultivos celulares bien descritos (Bruggeman y Westerhoff 2006). Antes de que las secuencias genómicas anotadas se volvieran disponibles, las principales fuentes de información para reconstruir procesos biológicos eran provistas por la literatura existente sobre la caracterización bioquímica de un número selecto de organismos. Por eso, era de esperarse que el desarrollo de tales modelos tenga lugar en sistemas u organismos comúnmente utilizados como modelos para el análisis, la interpretación y la predicción de relaciones entre genotipos y fenotipos. Acorde con lo anterior, algunas de las primeras reconstrucciones metabólicas se hicieron para *Bacillus subtilis*, *Escherichia coli*, *Saccharomyces cerevisiae, Caenorhabditis elegans* y *Drosophila melanogaster* (Palsson 2000; Feist 2009).

I.3.b *Saccharomyces cerevisiae*

S. cerevisiae es un hongo unicelular ampliamente utilizado en la producción de pan, cerveza y vino que ha sido empleado en la biología molecular y celular desde sus inicios. Por ello, su mapa genético completo fue descrito desde hace varias décadas y su secuencia genómica completa fue la primera en ser descrita para un organismo eucarionte (Goffeau 1997). Esto ha permitido su revisión amplia y constante convirtiéndola en un recurso indispensable para el análisis detallado de la función genética y la arquitectura genómica (Clayton 1997).

Algunos de los avances tecnológicos más recientes y aún en desarrollo están aumentando las formas en que es posible manipular el material genómico

de este organismo. Uno de estos adelantos relativamente recientes es la elaboración de una colección de deleciones de la gran mayoría de los genes de la levadura (Shoemaker 1996; Winzeler 1999) permitiendo la caracterización sistemática de su genoma para fenotipos que pueden ensayarse a gran escala. Adicionalmente, esta recopilación de mutantes identifica individualmente cada genotipo mutante con un código de barras de 20 pares de bases que puede ser usado para analizar varios genotipos en paralelo y poder distinguir cada uno dentro de una población de cepas mutantes. De tal forma, se puede evaluar la adecuación de cada cepa de deleción, para cada condición dada, en un experimento sencillo. Con la disponibilidad de todas estas cepas con deleciones de un solo gen, la investigación en marcha y a futuro intenta caracterizar fenotipos relacionados a los distintos genes (Ideker 2001).

Incluso se han analizado los genes que no son esenciales al caracterizar una parte considerable del total de interacciones genéticas posibles en *S. cerevisiae* (Costanzo 2010). Tales interacciones fueron cuantificadas según su adecuación ante distintas condiciones de crecimiento distinguiendo desde el caso más severo en que la falta de la interacción es letal (letalidad sintética) hasta los casos en que se disminuye o aumenta incluso de manera sutil (dependiendo de distintos niveles de confianza).

I.3.c El metabolismo de *S. cerevisiae* como objeto de estudio

Para lograr un conocimiento profundo de las capacidades metabólicas celulares se pueden utilizar métodos como el modelado teórico. Estos modelos deben permitir el análisis detallado de las interacciones entre las vías individuales que funcionan dentro de la célula, donde el primer paso natural es reconstruir la red metabólica subyacente. Uno de los procedimientos para diseñar modelos metabólicos de toda una célula se basa en la reconstrucción de redes metabólicas usando secuencias genómicas, bases de datos de vías metabólicas y literatura bioquímica (Förster 2003; Förster 2003b; Duarte 2004). Tales redes constituyen un

primer modelo que ya ha sido evaluado en su capacidad para reproducir fenómenos celulares relacionados al metabolismo (como el crecimiento de cepas de deleción) obteniendo resultados consistentes con las mediciones experimentales.

La red metabólica de *S. cerevisiae* ha sido reconstruida usando información genómica, bioquímica y fisiológica disponible actualmente, donde las reacciones metabólicas fueron compartamentalizadas dentro del citosol o la mitocondria e incluyendo los pasos de transporte entre tales compartimientos y su ambiente (Förster 2003; Förster 2003b). Dichos modelos metabólicos ya han sido utilizados para ser reconciliados y analizados con datos de crecimiento en cepas mutantes como se menciona arriba. Esto ha mostrado que un modelo *in silico* a escala genómica puede ser utilizado como marco estructurado para llenar sistemáticamente los espacios vacíos en nuestro conocimiento sobre un organismo y formar repositorios centrales que incorporen datos genómicos, transcriptómicos, proteómicos y metabolómicos.

Las redes metabólicas han sido modeladas extensivamente usando estudios de estado estacionario. Entre las aproximaciones utilizadas se encuentran los métodos como **FBA** (Förster 2003), **MOMA** y **EMA** (Rocha 2010). Estos métodos utilizan la topología de la red, la estequiometria de las reacciones químicas y de transporte y, en algunos casos, las velocidades máximas de algunas reacciones. Todos ellos han sido útiles para entender la estructura y dinámica del flujo metabólico pero arrojan predicciones experimentales distintas que aún no pueden separar el papel de la topología del de los demás parámetros funcionales de la red.

Existen estudios para *S. cerevisiae* en los que se han utilizado distintos conjuntos públicos de datos y se han analizado de diferentes formas. Tal es el caso de los genes esenciales listados en el YDP que se han evaluado según sus características genómicas (Seringhaus 2006). También, al estudiar la topología de

redes metabólicas, se ha demostrado que esta es suficiente para predecir la viabilidad de cepas mutantes, por lo menos tan bien como con el FBA (Wunderlich y Mirny 2006). Se ha utilizado la filogenia de los genes de la base de datos DIP para predecir su esencialidad usando inteligencia artificial (Saha y Heber 2006) y estas dos últimas aproximaciones han sido comparadas sobre el mismo conjunto de datos (Hwang 2009). También se han usado características topológicas así como de función y localización celular para predecir genes esenciales en una red construida a partir de interacciones físicas, metabólicas y regulación transcripcional listadas en la base de datos BioGRID obteniendo resultados aceptables (Acencio y Lemke 2009).

II. OBJETIVOS

II.1 Objetivo general

Reconstruir redes genéticas mediante relaciones químico-genéticas (metabólicas) e IM diversas. Utilizar su topología para predecir genes esenciales que participan en el metabolismo.

II.2 Objetivos particulares

- Utilizar redes químico-genéticas disponibles públicamente (KEGG e iND750) para añadirles las relaciones genéticas representadas mediante distintos tipos de IM anotadas y disponibles en bases de datos públicas (IPP obtenidas mediante aproximaciones de alto rendimiento, co-citación en la literatura y curado manual por expertos, además de interacciones genéticas probabilísiticas).

- Hacer uso de distintas medidas de centralidad local y global para evaluar si la predicción de genes esenciales mejora al incluir IM diversas en la red del metabolismo de la levadura.

- Determinar los genes esenciales identificados al utilizar relaciones químico-genéticas y compararlos con aquellos obtenidos al añadir IM. Probar si la existencia de genes que se predicen como esenciales solamente al agregar IM podría revelar la importancia de la IM sobre la relación químico-genética para la esencialidad de dichos genes.

III. HIPÓTESIS

La esencialidad de los genes que participan en el metabolismo no radica solamente en la función enzimática de las proteínas que codifican.

IV. JUSTIFICACIÓN

Dos de los tres tipos principales de redes a escala celular (de IPP y de regulación transcripcional) son reconstruidos a partir de datos a gran escala que están sujetos a cierta imprecisión. En contraste, las redes metabólicas son derivadas de vías metabólicas inferidas mediante experimentos bioquímicos ampliamente refinados reduciendo la posibilidad de errores. Por tanto, una buena reconstrucción metabólica puede reflejar, por lo menos, la mayoría de las capacidades metabólicas de una célula que no tienen por qué ser específicas para algún estado fisiológico particular y debe ser considerada como un paso inicial hacia la comprensión de la relación entre el genotipo y el fenotipo de un organismo en específico.

Al utilizar este tipo de redes se ha demostrado que los genes esenciales ocupan un lugar importante en su estructura por lo que pueden identificarse usando medidas topológicas buscando la representación de fenotipos particulares asociados a las relaciones químico-genéticas del metabolismo de *S. cerevisiae* (del Río 2009). Sin embargo, ninguno de los diversos trabajos que buscan predecir los genes críticos es capaz de predecir correctamente todos los genes que son esenciales para el metabolismo de la levadura. Tales inconsistencias entre la información y las simulaciones han señalado deficiencias que promueven la visión de que el estado incompleto de las redes puede limitar su capacidad para identificar genes esenciales. Por tanto, se puede buscar su optimización al integrar datos de distintas fuentes planteando nuevas hipótesis.

Existen trabajos en los que se comparan distintas características de los genes esenciales utilizando un mismo conjunto de datos en el que se pueden representar uno o más tipos de interacciones (ya sea metabólicas, IPP o proteína-ADN). Hasta donde se sabe, aún no se ha intentado probar distintas combinaciones de representaciones metabólicas, interactomas y medidas de

centralidad para predecir genes esenciales que participan en el metabolismo como se intenta en este trabajo.

Alternativamente, es posible explicar las limitaciones en la predicción de genes metabólicos esenciales considerando que la función de éstos es consecuencia de un proceso dinámico y no estructural como lo simulan las medidas de centralidad. Por ejemplo, es posible suponer que la esencialidad de genes metabólicos depende de la velocidad de reacción de las enzimas que codifican y de su localización en la red. De tal suerte, una medida que evalúa solo la topología de la red no podría tomar en cuenta dichas consideraciones y, como consecuencia, no sería capaz de predecir correctamente los genes críticos.

De esta forma, probar distintas reconstrucciones metabólicas para determinar si la esencialidad de sus genes depende de la estructura de la red que conforman, de su dinámica o bien de ambas, representa una pregunta abierta en el estudio del metabolismo y la biología de sistemas. Este trabajo busca dar respuesta a esta pregunta a partir un análisis bioinformático sobre la información acumulada en distintas bases de datos disponibles públicamente.

V. MÉTODOS

V.1 Redes

V.1.a Conceptos y definiciones

En donde quiera que existan elementos pertenecientes a grupos se pueden definir conjuntos. Un conjunto es una colección de objetos llamados miembros o elementos formalmente listados entre { } cuya definición más común consiste en ser únicos y no tener un orden específico pudiendo existir un conjunto vacío denotado por el símbolo ∅. Existen varias operaciones básicas y comunes aplicables a los conjuntos como la unión[*] que contiene a todos los elementos encontrados y la intersección[†] que incluye los elementos comunes. Ambas operaciones implican a todos los conjuntos participantes y son conmutativas de manera que no importa el orden en que los elementos sean añadidos o listados (Orwand 1999).

La teoría de redes surge de la teoría de conjuntos para describir las propiedades matemáticas de las relaciones y, dentro de ella, una red es un conjunto de nodos o vértices ***V*** unidos por un conjunto de bordes o aristas ***A*** con dirección opcional (Huber 2007). La representación computacional de una red depende, además del propósito con el que son elaboradas, de la naturaleza y cantidad de datos que se espera que contengan incluyendo todas las relaciones representadas y variando la forma en que se presentan los vértices y/o las aristas. En una lista de adyacencia se representa cada **vecino**[‡] de cada nodo y en una matriz de adyacencia presenta cada nodo tanto en las columnas como en las filas representando su interacción mediante operadores binarios 0 si no existe, 1 cuando existe. Una lista parental sirve para redes con forma de árbol y es muy

[*] a veces llamada suma o máximo y denotada por el símbolo ∪ o el operador lógico || o "OR"; un elemento se encuentra en la unión si está en alguno de los conjuntos que participan en ella.
[†] también conocida como producto o mínimo y denotada por el símbolo ∩ o el operador lógico | o "AND".
[‡] sucesor o nodo que participa en una conexión mediante una arista.

compacta porque sólo necesita listarse el nodo del que surge el actual, excepto para el nodo raíz de donde surgen todos los demás (Orwand 1999).

En este trabajo se utilizaron listas de adyacencia principalmente por su conveniencia para la mayoría de los algoritmos utilizados. Una lista de adyacencia puede ser representada como un objeto de Perl llamado *hash*[*] en donde cada arista es representada por los vértices que la componen garantizando tanto la unicidad como el orden inespecífico que caracterizan a los conjuntos.

Con este contexto, y de acuerdo con el paradigma estructura-función antes mencionado, un fenotipo puede ser representado como una red **R** con **n** genes definida por un conjunto de genes **V** (vértices o nodos) y un conjunto de relaciones **A** (aristas) de manera que: $\mathbf{V}(\mathbf{R}) = \{v_1, v_2, \ldots, v_n\}$ y $\mathbf{A}(\mathbf{R}) = \{v_1v_2, v_2v_4, \ldots, v_xv_y\}$ que no pueden ser conjuntos vacíos y en donde n es el número total de genes de la red. Los genes críticos se pueden definir como el conjunto **c** de manera que: $\mathbf{c} \in \mathbf{V}(\mathbf{R})$ por lo que $\mathbf{V}(\mathbf{R})$ incluye los genes en **c** que son identificables al ser ordenados usando los valores arrojados por distintas operaciones matemáticas (medidas de centralidad, ver abajo) (del Río 2009). Estas definiciones hacen posible la elaboración de programas que implementen operaciones para construir redes y algoritmos para analizarlas. A continuación se mencionan algunas de las características medibles en redes que serán utilizadas este trabajo.

Todos los algoritmos empleados para ordenar los genes y distinguirlos como esenciales o no dependen del procesamiento de los vértices y/o de las aristas en algún orden. La secuencia de aristas necesarias para atravesar la red desde el nodo *n* hacia el nodo *m* puede definir una **ruta** o varias, con ella se puede definir la **conectividad** de tal forma que el nodo *n* y el nodo *m* están conectados si existe una ruta entre ellos. La sucesión más pequeña de aristas o **ruta más corta** define la **distancia geodésica** o, simplemente, **distancia**. Un **ciclo** es una secuencia de aristas que lleva a algún vértice visitado previamente y aquel

[*] lista de valores que pueden ser accesados mediante llaves únicas.

conjunto de vértices donde todos son alcanzables entre sí, es decir. Un **componente conectado** está compuesto por un ciclo o varios entrelazados y el más grande de ellos (en el caso de haber más de uno) es llamado **componente gigante** (**CG**) (Orwand 1999).

Las medidas de centralidad están definidas dentro de la teoría de redes como algoritmos que evalúan la importancia relativa de un nodo dentro de una red. Las medidas de centralidad utilizadas en este estudio son listadas a continuación:

- **clos**: El recíproco de la distancia promedio entre un nodo y todos los demás nodos de la red.

  $$C(x) = \frac{n-1}{\sum_{y \in A, y \neq x} d(x,y)} = \frac{1}{\bar{d}}$$

  donde *d*(*x*,*y*) es la distancia entres el nodo *x* y el nodo *y*, A es el conjunto de todos los nodos y $\bar{d}$ es la distancia promedio entre *x* y todos los demás nodos.

- **cluster**: La razón del número *e* de conexiones entre los vecinos del nodo *n* entre el número máximo de conexiones del mismo nodo *n.* Para nodos con 0 o 1 conexiones es igual a 0.

  $$CE(n) = \frac{e}{k(n)[k(n-1)]}$$

- **deg**: El número *k* de aristas que llegan a un nodo *n* y que salen de él. Si no se define como entrante o saliente entonces se trata del número de aristas en donde participa *n* sin importar su dirección.

- **excen**: La distancia geodésica entre el nodo *n* y el nodo más alejado de éste. Si no existe una ruta a partir del nodo *n* su valor es 0.

- **mDist**: El promedio de todas las distancias geodésicas entre el nodo *n* y todos los nodos a los que se pueda llegar desde él. Si el grado de *n* es igual a 0, se asume que la distancia promedio también es igual a 0.

- **S**: La cantidad de rutas más cortas de la red que atraviesan el nodo *n*.

- **sphereDeg**: El número de vértices alcanzables a una distancia de 2.

- **travA**: También puede ser llamada conectividad dinámica (*dk*) y se define al contar cuantas veces es atravesado un nodo al conectar cada par de nodos según su orden en la red. Se relaciona con la intermediación de la forma:

$$SNN = \frac{dk}{N(N-1)}$$

- **trav**: Igual que la anterior pero calculada según el orden de los vértices.

Clos, excen, mDist y S son medidas globales de centralidad mientras que las demás son medidas locales, en el sentido de que dependen solamente de las conexiones del gen.

Los nodos de una red pueden ordenarse de mayor a menor según una medida de centralidad. Si los nodos se ordenan con respecto al inverso aditivo[*] de la misma medida de centralidad el orden de los nodos se invierte, es decir, se ordenan de menor a mayor. Se calcularon los inversos aditivos para los valores obtenidos con las medidas de centralidad cluster y excen (**clusteInv** y **excenInv**, respectivamente). Solamente se usaron los dos anteriores debido a que son los únicos que obtuvieron resultados positivos en estudios anteriores (del Río 2009).

[*] El valor de un nodo según el inverso aditivo de una medida de centralidad se define como $C_{ia}(n) = -C(n)$, donde $C(n)$ y $C_{ia}(n)$ son el valor del nodo *n* según la medida de centralidad C y el inverso aditivo de la medida de centralidad C, respectivamente.

Además de los valores de centralidad y el número de vértices y/o aristas pueden ser calculadas otras propiedades tanto de las redes, como de sus CGs. Entre dichas propiedades se encuentran: el **diámetro** (distancia máxima entre dos nodos), el parámetro **OCCI**[*] (Ma y Zeng 2003), **meanClusteringCoef**, **meanClosenessCent**, **meanExcentricity**, **meanDistance**, **meanDegree** y **meanSphereDegree**.

V.1.b Redes metabólicas de *S. cerevisiae*

Se utilizaron distintas redes metabólicas generadas a partir de la base de datos KEGG (ftp://ftp.genome.jp/pub/kegg/pathway/organisms/sce/) y de una red construida a partir de datos de experimentales (Duarte 2004). En la tabla 1 se listan las características de cada una de dichas redes y se describen en los párrafos siguientes.

Red	Metabolitos más conectados (eliminados)	Vías	ECno	Genes Hipotéticos	Compartimientos
KEGG					
KEGGtype			Si		
KEGGpath					
KEGGtypepath					
KEGG2	H_2O, ATP, ADP, NAD+, NADH, NADP, NADPH			ND	ND
KEGG2type					
KEGG2path					
KEGG2typepath					
Palsson_HIPOT_0	H_2O, H+	Si			
Palsson_HIPOT_1	H_2O, H+, Pi				
Palsson_HIPOT_2	H_2O, H+, Pi, ATP		ND	Si	
Palsson_HIPOT_3	H_2O, H+, Pi, ATP, Glu-L				
Palsson_HIPOT_4	H_2O, H+, Pi, ATP, Glu-L, ADP				
Palsson_nonHIPOT_0	H_2O, H+				Si
Palsson_nonHIPOT_1	H_2O, H+, Pi				
Palsson_nonHIPOT_2	H_2O, H+, Pi, ATP			No	
Palsson_nonHIPOT_3	H_2O, H+, Pi, ATP, Glu-L				
Palsson_nonHIPOT_4	H_2O, H+, Pi, ATP, Glu-L, ADP				

Tabla 1 Propiedades de las redes metabólicas utilizadas.
ECno: Enzyme classification number; **ND**: no determinado.

Todas las redes KEGG fueron construidas a partir de la información presente en la base de datos de reacciones metabólicas KEGG PATHWAY en las

[*] que representa la distribución de los valores de la cercanía "clos"

cuáles se conectaron los genes anotados como enzimas* a través de sus metabolitos sin contar aquellos mayormente conectados: agua, ATP, ADP, NAD^+, NADH, $NADP^+$, NADPH y oxígeno. La red KEGG2 fue construida con la información de KEGG PATHWAY como tal y a partir de ellas fueron derivadas las redes type, considerando el tipo de reacción anotado†, así como las redes path, en la que los genes anotados como pertenecientes a distintas vías metabólicas no deberían ser unidos a menos de que tal gen se presente en más de una vía y, por último, las redes typepath con las dos consideraciones anteriores (del Río 2009).

Las redes Palsson_ fueron derivadas a partir de la red iND750 que ha sido curada a mano tomando en cuenta la compartamentalización celular. A partir de ésta se construyeron dos grupos, en el primero, Palsson_HIPOT, fueron consideradas todas las reacciones contenidas en el modelo original y en el segundo, Palsson_NONHIPOT, se eliminaron las reacciones hipotéticas. Las redes _0, _1, _2, _3 y _4 se obtuvieron eliminando: agua y H^+, las anteriores y Pi, las anteriores y ATP, las anteriores y L-Glu, además de las anteriores y ADP, respectivamente (del Río 2009). La identificación y remoción de las aristas en las que participan los metabolitos más conectados responde a que, de tal forma, el cálculo de la ruta más corta entre dos metabolitos resulta más representativo. Lo anterior surge de la observación de que al utilizar tales metabolitos, el promedio de la distancia más corta resulto cercano a 3 en distintos organismos lo que significaría que cualquier metabolito podría ser convertido en otro en ~3 pasos. Dicha conclusión va en contra de lo observado, por ejemplo, en la glucólisis donde se ha demostrado bioquímicamente que se requiere de 9 pasos y al considerarse tanto el ATP como el ADP se reduce a 2 (Ma y Zeng 2003).

* con clasificación **EC**; aquellos genes con el mismo número EC y diferente sustrato fueron forzados a aceptar todos los sustratos para tal número EC.
† reversible o irreversible.

V.1.c Redes de interacciones moleculares

Los genes de *S. cerevisiae* que codifican para las proteínas cuya interacción es reportada en las bases de datos BioGRID, DIP, IntAct, MPact, Yeastnet, YIPD y YPLND fueron utilizados como vértices para construir redes donde la existencia de interacción da lugar a las aristas (Ver sección V.1.a Conceptos y definiciones). En la tabla 2 se comparan las características de cada uno de los conjuntos de datos y la dirección web de donde se obtuvieron. Adicionalmente se calcularon tanto el conjunto unión como el conjunto intersección con todas las interacciones diferentes presentes en las bases de datos anteriores; el conjunto intersección resultó vacío por lo que no fue considerado más adelante.

Red	IG	IP	AR	CL	CM	Sitio web
Biogrid	Si	Si	Si	Si	Si	http://www.thebiogrid.org/
DIP	No	Si	No	Si	No	http://dip.doe-mbi.ucla.edu/
Intact	No	Si	No	Si	Si	ftp://ftp.ebi.ac.uk/pub/databases/intact/current
Mpact	No	Si	No	Si	Si	ftp://ftpmips.gsf.de/yeast/PPI/
Yeastnet	Si	No	No	No	No	http://www.yeastnet.org/
YIPD	No	Si	Si	No	No	http://itolab.cb.k.u-tokyo.ac.jp/Y2H/
YPLND	No	Si	Si	No	No	http://depts.washington.edu/sfields/yp_interactions/index.html

Tabla 2 Tipos de información y fuente de las bases de datos de interacciones moleculares utilizadas. **IG**: Interacción Genética; **IP**: Interacción proteína-proteína; **AR**: Alto rendimiento; **CL**: co-citación en la literatura; **CM**: Curada manualmente.

V.1.d Integración

Se evaluó un número total de 170 redes. Esto se logró calculando el conjunto unión para cada una de las 18 diferentes redes metabólicas (KEGG, KEGGPATH, KEGGTYPE, KEGGTYPEPATH 1 y 2 además de Palsson_HIPOT y _NONHIPOT de la _0 a la _4) con las 8 redes de IM (Biogrid, Dip, Intact, Mpact, Yeastnet, YIPD, YPLND y Union) tomando en cuenta las redes originales. Las redes metabólicas, las de IM y las que son resultado de la unión entre redes metabólicas y redes de IM serán referidas como **MetNets**, **IntNets** e **Int.Met**, respectivamente. Debido a particularidades de algunos de los algoritmos mediante

los que se calculan los valores de centralidad, se retiraron los "self-loops"[*] de cada una de las redes antes de analizarlas.

V.2 Evaluación funcional

V.2.a Genes esenciales

Todas las redes y sus CGs fueron evaluadas en su capacidad de predecir los genes reportados como esenciales en la base de datos del **SGDP** (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html). Los genes de una red que aparecen en dicho inventario se consideraron como esenciales dejando a los demás como no esenciales. Así se obtuvieron una lista de genes esenciales y otra de genes no esenciales para cada una de las redes. En el caso de las redes Int.MetNets sólo se consideraron los genes esenciales de la parte metabólica de la red buscando una mejor representación enfocada en metabolismo. Lo anterior debido que se trata de la información que parece más confiable (ver sección I.2.a) y a la importancia de las redes metabólicas en tareas como la ingeniería metabólica, el estudio de trastornos metabólicos, la predicción de potencial patogénico, la identificación de blancos farmacéuticos y el análisis de flujo de carbono.

V.2.b Las medidas de centralidad como predictores de genes esenciales

Se puede proponer una operación matemática sobre **A**(**R**) para clasificar **V**(**R**) y obtener **c** (ver sección V.1.a). En este caso, tal operación matemática es la medida de centralidad de una red. Si dos fenotipos p y q no comparten el mismo conjunto de genes críticos, entonces $\mathbf{c}_p \neq \mathbf{c}_q$ de modo que no es posible obtener ambos conjuntos usando la misma operación matemática sobre la misma red. La

[*] Tomando en cuenta que las aristas de una red se definen en base al par de nodos *u, v* que la integran, si *u*=*v* se trata de un "self-loop", si no, se define como una arista propia.

red con el mayor número de genes esenciales identificado será referida como red **REMG** (del Río 2009).

Nótese que la meta de este estudio es reconstruir redes a partir de otras que han probado una capacidad limitada para predecir genes esenciales buscando superar las limitaciones observadas anteriormente y explicar la esencialidad de los genes metabólicos. En este sentido, se realizó el mismo tipo de evaluación para distintos tipos de redes (o partes de ellas) para comparar distintas alternativas. Todas las redes y sus respectivos CGs fueron evaluados para predecir los genes reportados como esenciales en la base de datos del SGDP mediante 11 medidas de centralidad: clos, cluster, clusteInv, deg, excen, excentricitInv, meanDist, S, sphereDeg travA y trav (ver la sección V.1.a Conceptos y definiciones).

V.2.c Evaluación de índices topológicos para clasificar genes como esenciales

Los genes de cada red se ordenaron de los de mayor valor a los de menor valor según la medida de centralidad utilizada. Al hacer esto con todas las medidas se obtuvieron tablas de genes ordenados de mayor a menor valor de centralidad para cada red y se relacionaron con las listas de genes esenciales para evaluar el desempeño predictivo para cada centralidad y red.

Para evaluar el desempeño predictivo de una centralidad se miden parámetros estadísticos que dependen del número de **TP**s, **FP**s, **TN**s y **FN**s definidos en términos de los fenotipos medidos experimentalmente. De tal modo se pueden definir la sensibilidad[*], la especificidad[†], la exactitud[‡] y la tasa de falsos

[*] TP/(TP+FP)
[†] TN/(TN+FN)
[‡] (TP+TN)/(TP+TN+FP+FN)

positivos*, donde los positivos son las mutaciones viables y los negativos representan a las mutaciones reportadas como letales (genes críticos).

Al graficar la sensibilidad contra la especificidad del método de predicción se obtienen las llamadas curvas **ROC** que han sido ampliamente usadas para juzgar la capacidad de discriminación de métodos estadísticos que combinan varios factores para propósitos predictivos. Se ha propuesto un gran número de medidas teóricas para reducir una curva ROC a un sólo índice cuantitativo basándose en el supuesto de que las distribuciones que subyacen a los grupos positivo y negativo son Gaussianas, siendo el **ABC** el más popular entre ellos. Tal índice varía desde 0.5 (sin exactitud aparente) hasta 1.0 (perfectamente exacto) para predicciones cada vez mejores.

El ABC mide la probabilidad de distinguir los genes esenciales de los no esenciales y está relacionada con pruebas estadísticas no paramétricas como la de Wilcoxon o la de Mann-Whitney. Este índice depende del cálculo integral y no asume supuestos sobre la distribución de los datos que cuando son extensivos tienden razonablemente bien hacia una curva ROC cuya forma implica dos distribuciones Normales.

Cuando los puntos de la curva ROC se someten a un programa iterativo de estimación de máxima similitud se obtienen diferencias entre medias y proporción de varianza como parámetros. El error estándar se utiliza para cuantificar que tan variable es tal área bajo la curva de muestras diferentes con tamaños similares requiriendo del cálculo de dos probabilidades intermedias Q_1 (probabilidad de que dos mediciones negativas tomadas al azar sean clasificadas mejor que una positiva escogida aleatoriamente) y Q_2 (probabilidad de que una medición negativa escogida al azar sea mejor clasificada que dos mediciones positivas escogidas aleatoriamente) útiles para calcular las probabilidades α y β de cometer errores tipo I o tipo II, respectivamente y construir intervalos de confianza. Esto permite

* FDR = FP/TP+FP

que el área derivada a partir de ambos parámetros sea acompañada por el error estándar para construir intervalos de confianza y realizar pruebas de significancia.

La equivalencia entre este análisis y la estadística de Wilcoxon o a la U de Mann Whitney normalizada por el número de pares posibles, implica que un modelo efectivo deberá generar un valor de ABC significativamente mayor que 0.50 (valor esperado al azar) (Hanley y McNeil 1982), por tanto, fueron calculados los **IC** de 0.1, 0.05 y 0.01 (90%, 95% y 99%) y el **ES** para el ABC. También fue calculado el **minE** de cada curva ROC para rastrear al punto de la curva más cercano a la predicción perfecta, además de la exactitud predictiva (accuracy) para tal error (Wunderlich y Mirny 2006).

V.2.d Análisis comparativo de los distintos tipos de redes y medidas de centralidad

La estadística, tanto descriptiva como inferencial para los datos obtenidos, fue realizada mediante funciones básicas del software estadístico R (Becker 1988) y de algunos paquetes adicionales indicados a continuación. Para representar los datos se elaboraron gráficas de tipo scatterplot y boxplot usando el paquete *scatterplot3d* y la función *boxplot* respectivamente. Los datos fueron descritos mediante funciones del paquete *psych*, se utilizaron las pruebas de normalidad Shapiro-Wilk y Anderson-Darling presentes en el paquete *nortest* así como la homogeneidad de las varianzas que fue evaluada con la prueba Fligner-Kileen y la prueba de significancia de Wilcoxon que se utilizó para evaluar las diferencias entre los grupos. El contenido de genes de la mejor red fue analizado mediante un diagrama de Venn Euler elaborado utilizando el paquete *venneuler*. Las tablas utilizadas como input fueron elaboradas mediante scripts de AWK y el output fue redirigido tanto a figuras en formato jpeg como a archivos de texto.

V.3 Manejo de archivos y programas de computadora

La información proveniente de las distintas bases de datos fue manejada utilizando un sistema tipo Linux y, una vez descargada, se utilizaron programas de AWK para construir listas de adyacencia tomando los identificadores de cada gen cuya interacción fue reportada. La transformación de las bases de datos en conjuntos y en redes, además de las operaciones (de conjuntos; ver sección V.1.a) entre éstas, fueron realizadas mediante programas escritos en PERL. La evaluación funcional fue realizada usando programas codificados en Java como implementaciones de los métodos reportados por Thibert (2005) y Cusack (2007), para calcular los valores de las medidas de centralidad, y del método desarrollado por Aurora Labastida (del Río 2009), basado en la utilización de curvas ROC, cuya ABC fue calculada usando un método empírico y utilizada como un estimado de qué tan bueno es el modelo para diferenciar los genes esenciales de los no esenciales. El análisis estadístico y las gráficas fueron realizados utilizando el software estadístico R (Becker 1988).

VI. RESULTADOS

VI.1 Evaluación funcional

VI.1.a Redes Metabólicas

Los valores obtenidos para cada gen de las 18 redes metabólicas, evaluadas según 11 criterios de centralidad, fueron utilizados para valorar la capacidad de cada medida para identificar los genes esenciales de cada red. Se considera que una medida es capaz de identificar a los genes esenciales si producía un ABC con IC de 99% mayor a 0.5.

Como se puede observar en las figuras 1 y 2, ninguna medida de centralidad fue capaz de identificar correctamente a todos los genes esenciales (ABC=1) y solamente en 27 de las 198 posibilidades se encontró un ABC significativamente mayor a 0.5 (IC de 99%; tabla 3). Los dos mejores casos se obtuvieron usando excentricitInv y clos en las redes Palsson_nonHIPOT_1 y KEGG2path, respectivamente, por lo que no se puede decir que hay una diferencia marcada con respecto al uso de redes metabólicas pero las centralidades globales si parecen ser mejores.

metnet	centralidad	ABC	99%ls	99%li
Palsson_nonHIPOT_1	excentricitInv	0.585	0.621	0.550
KEGG2path	clos	0.586	0.632	0.540
Palsson_nonHIPOT_0	excentricitInv	0.573	0.610	0.537
KEGGpath	cluster	0.583	0.630	0.536
KEGG	cluster	0.583	0.631	0.535
KEGGpath	clos	0.580	0.625	0.535
KEGG2typepath	clos	0.573	0.619	0.526
KEGG2	cluster	0.574	0.624	0.525
KEGGpath	excentricitInv	0.568	0.613	0.524
KEGG2path	cluster	0.572	0.620	0.524
KEGGtypepath	clos	0.569	0.614	0.523
Palsson_nonHIPOT_1	travAlea	0.556	0.589	0.522
Palsson_nonHIPOT_1	clos	0.559	0.597	0.522
Palsson_nonHIPOT_4	meanDist	0.556	0.591	0.521
Palsson_nonHIPOT_1	travGab	0.553	0.586	0.520
Palsson_nonHIPOT_3	cluster	0.548	0.586	0.511
Palsson_nonHIPOT_0	travAlea	0.543	0.576	0.509
KEGG2typepath	excentricitInv	0.554	0.601	0.508
Palsson_nonHIPOT_0	travGab	0.540	0.573	0.507
KEGG2path	excentricitInv	0.553	0.600	0.506
Palsson_nonHIPOT_2	cluster	0.543	0.580	0.505
Palsson_nonHIPOT_3	excentricitInv	0.538	0.573	0.504
Palsson_nonHIPOT_0	clos	0.540	0.578	0.503
Palsson_nonHIPOT_4	excen	0.539	0.576	0.503
Palsson_nonHIPOT_2	excentricitInv	0.537	0.572	0.502
Palsson_nonHIPOT_4	travAlea	0.534	0.567	0.501
Palsson_nonHIPOT_4	cluster	0.540	0.580	0.501

Tabla 3 Redes metabólicas y centralidades con las que se obtuvieron valores de área bajo la curva (ABC) significativamente mayores a 0.5

VI.1.b Redes de IM

En la figura 3 se puede notar que, al utilizar los mismos 11 índices de centralidad sobre las redes de IM, la mayoría (52 de 88; tabla 4) resultaron mejores que un clasificador aleatorio en contraste con lo que se observa con las redes metabólicas. Además, se puede notar que algunas centralidades locales obtienen valores más altos que las medidas globales.

intnet	centralidad	AUC	99%ls	99%li
Intact	deg	0.735	0.749	0.720
Yeastnet	deg	0.693	0.708	0.679
Yeastnet	sphereDeg	0.693	0.707	0.678
Yeastnet	SNN	0.693	0.707	0.678
Union	deg	0.690	0.704	0.676
Yeastnet	clos	0.688	0.702	0.673
Intact	clos	0.681	0.696	0.666
Dip	deg	0.670	0.686	0.654
Intact	SNN	0.668	0.683	0.653
Intact	sphereDeg	0.668	0.683	0.653
Intact	travGab	0.656	0.671	0.640
Intact	travAlea	0.655	0.671	0.640
Union	clos	0.643	0.657	0.629
Intact	cluster	0.643	0.658	0.629
Union	sphereDeg	0.639	0.654	0.625
Union	SNN	0.639	0.653	0.625
Dip	cluster	0.632	0.648	0.617
Biogrid	deg	0.629	0.641	0.616
Intact	excentricitInv	0.627	0.643	0.611
Dip	sphereDeg	0.623	0.639	0.607
Dip	SNN	0.623	0.639	0.607
Yeastnet	excentricitInv	0.622	0.637	0.607
Dip	clusteInv	0.618	0.634	0.603
Union	travGab	0.616	0.631	0.602
Dip	clos	0.618	0.634	0.601
Union	travAlea	0.615	0.630	0.601
Yeastnet	travAlea	0.616	0.631	0.601
Yeastnet	travGab	0.615	0.630	0.599
Dip	travGab	0.614	0.630	0.597
Dip	travAlea	0.613	0.630	0.597
Intact	clusteInv	0.609	0.624	0.595
Yeastnet	cluster	0.590	0.605	0.575
Biogrid	travAlea	0.583	0.597	0.570
Biogrid	travGab	0.583	0.596	0.570
Biogrid	clos	0.567	0.580	0.554
Biogrid	sphereDeg	0.566	0.579	0.553
Biogrid	SNN	0.566	0.579	0.553
Mpact	deg	0.567	0.584	0.549
Mpact	cluster	0.563	0.581	0.545
Mpact	meanDist	0.561	0.578	0.543
Mpact	excen	0.558	0.576	0.541
Mpact	clusteInv	0.559	0.576	0.541
Dip	excentricitInv	0.557	0.574	0.541
Mpact	travGab	0.558	0.575	0.540
Mpact	travAlea	0.558	0.575	0.540
Biogrid	cluster	0.546	0.560	0.533
Union	cluster	0.545	0.560	0.530
Biogrid	excentricitInv	0.534	0.548	0.520
Union	excentricitInv	0.531	0.546	0.515
YPLND	SNN	0.544	0.581	0.506
YPLND	sphereDeg	0.542	0.579	0.505
YPLND	deg	0.540	0.577	0.503

Tabla 4 Redes de IM y centralidades con las que se obtuvieron valores de área bajo la curva (ABC) significativamente mayores a 0.5

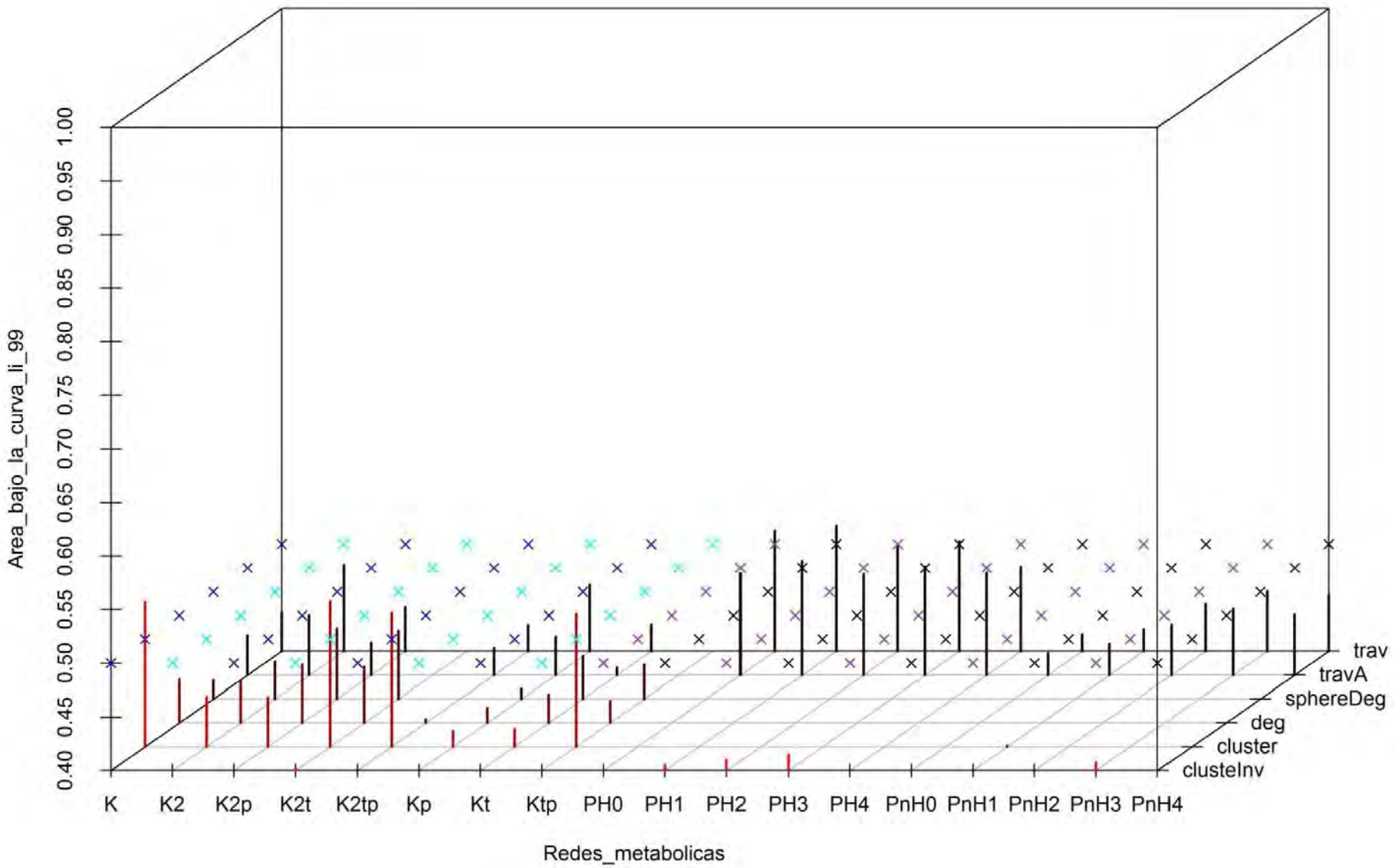


Figura 1 Se muestran los valores del límite inferior del intervalo de confianza de 99% para las centralidades locales de las redes metabólicas utilizadas. El eje de las ordenadas comienza en 0.4 y se marca con una x de color el punto por el que necesitaría pasar el valor de 0.5.

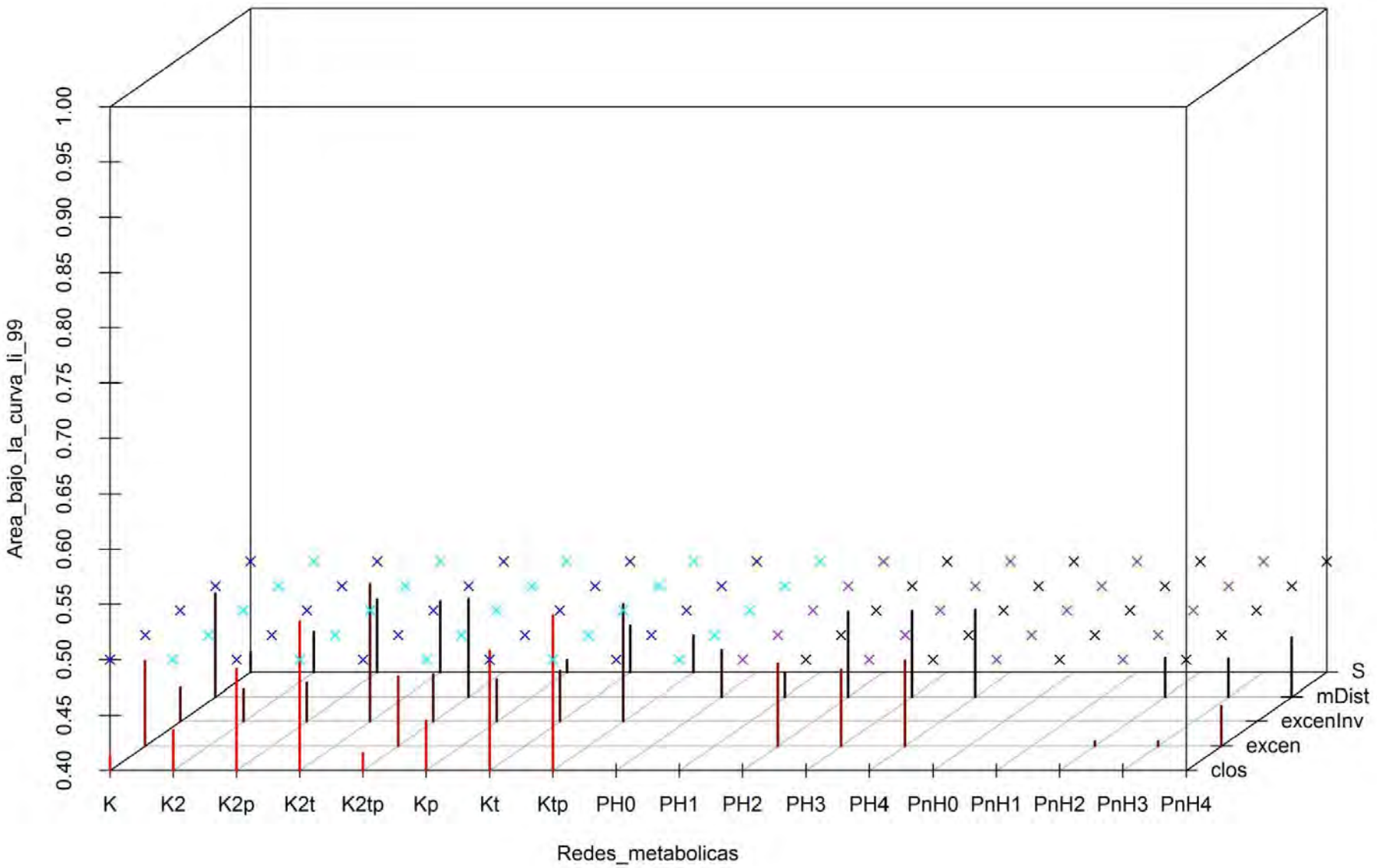


Figura 2 Se muestran los valores del límite inferior del intervalo de confianza de 99% para las centralidades globales de las redes metabólicas utilizadas. El eje de las ordenadas comienza en 0.4 y se marca con una x de color el punto por el que necesitaría pasar el valor de 0.5.

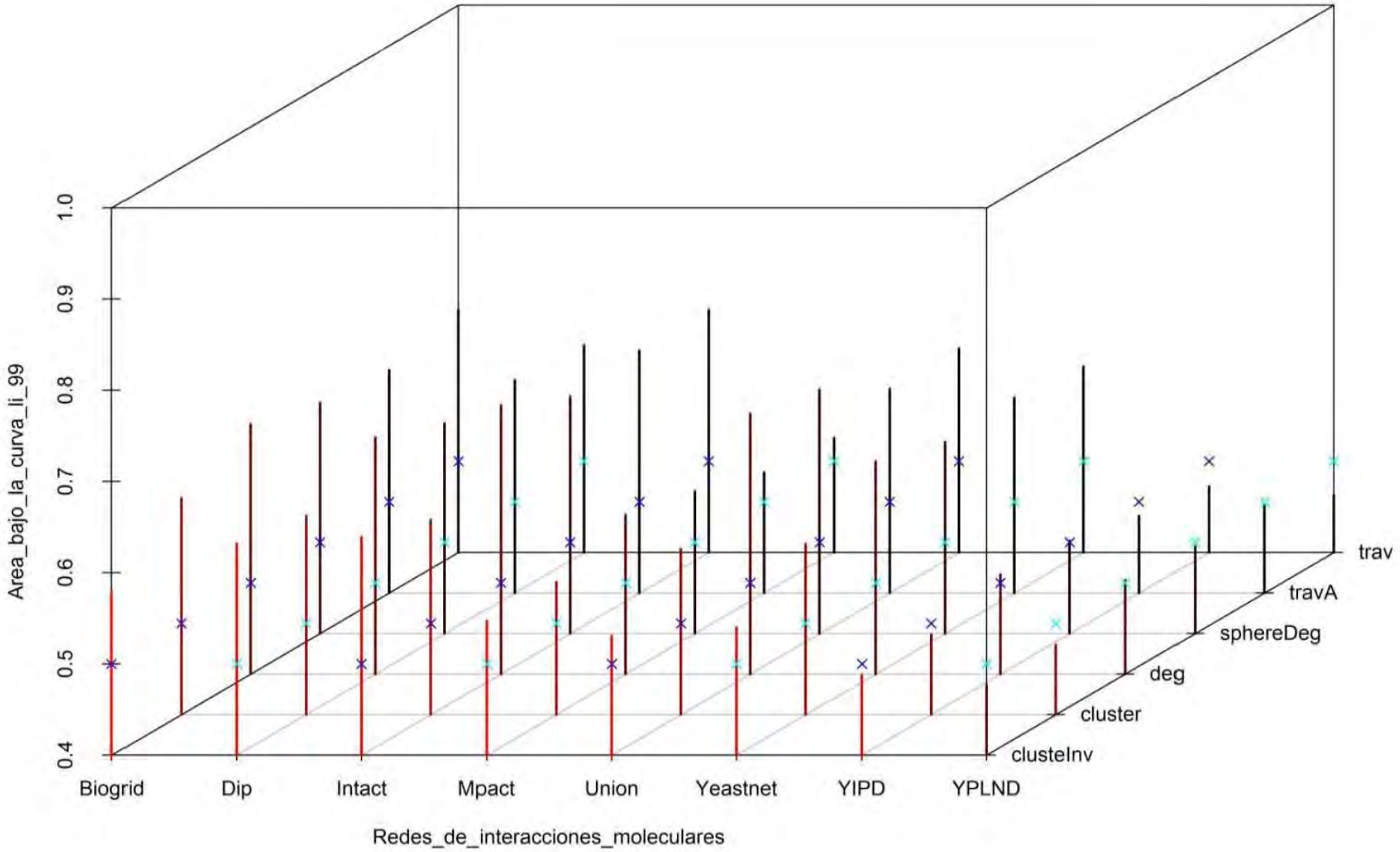


Figura 3 Se muestran los valores del límite inferior del intervalo de confianza de 99% para las centralidades locales (arriba) y globales (abajo) de las redes de IM. El eje de las ordenadas comienza en 0.4 y se marca con una x de color el punto por el que necesitaría pasar el valor de 0.5.

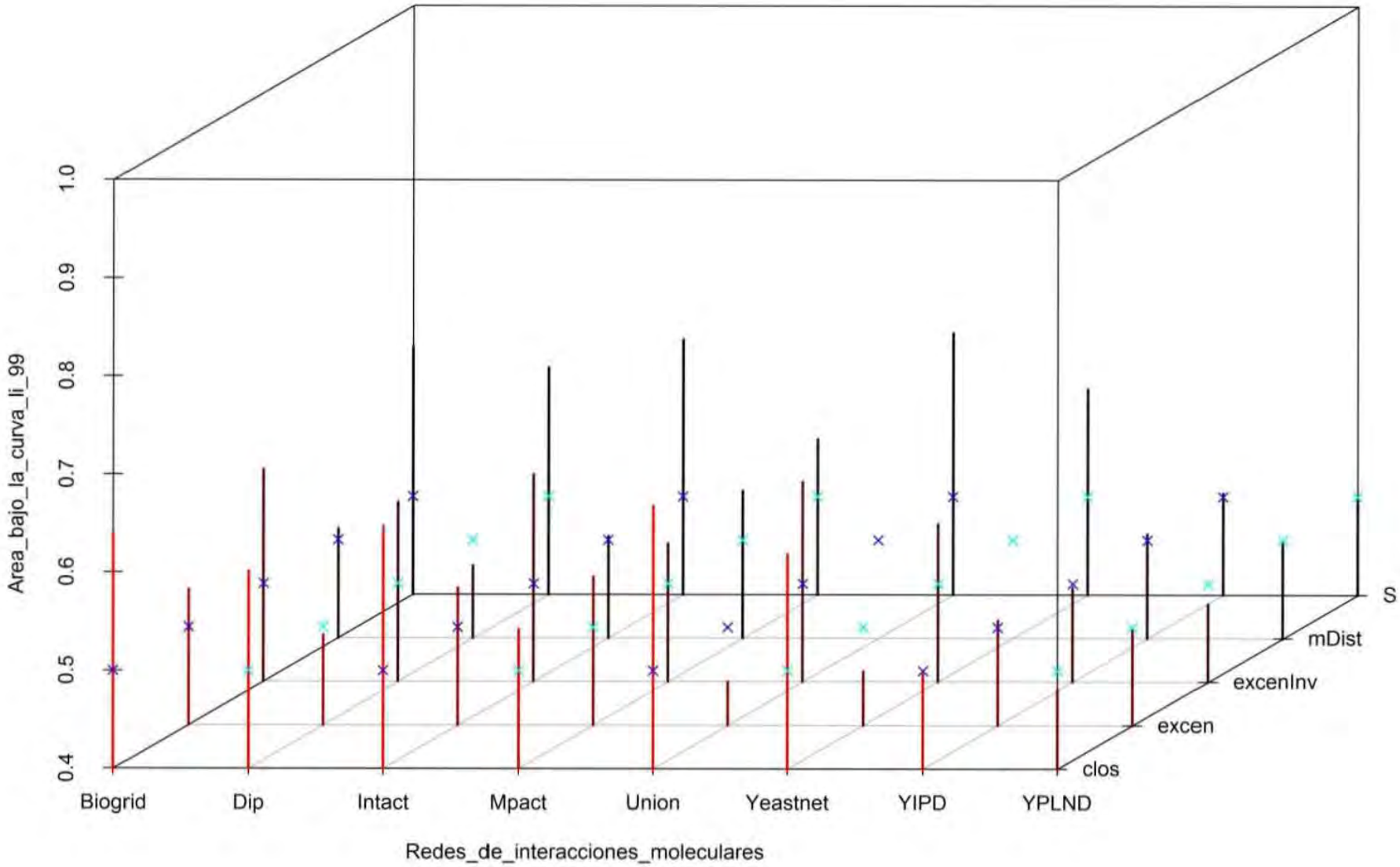


Figura 4 Se muestran los valores del límite inferior del intervalo de confianza de 99% para las centralidades locales (arriba) y globales (abajo) de las redes de IM. El eje de las ordenadas comienza en 0.4 y se marca con una x de color el punto por el que necesitaría pasar el valor de 0.5.

VI.1.c Optimización: Int.MetNets y CGs

Para explicar las limitaciones tanto de las redes metabólicas como de las de IM pueden proponerse dos alternativas que permitan obtener mejores resultados en la predicción de genes críticos y, a la vez, arrojen pistas para tratar de explicar su esencialidad. La primera tiene que ver con integrar las distintas fuentes de información y se basa en la evidencia de que las relaciones químico-genéticas cubren sólo una parte del amplio espectro de interacciones que puede tener un gen dentro de una célula. La segunda aborda la posibilidad de que la falta de conexiones entre genes limita la capacidad de distinguir los genes esenciales mediante medidas de centralidad por lo que se evaluó la subred conexa de mayor tamaño, el CG.

Para analizar las aproximaciones propuestas es necesario comparar los datos obtenidos inicialmente[*], con los valores conseguidos tras la integración de ambos tipos de redes en las llamadas Int.MetNets, además de la valoración de los CGs para los tres tipos de redes (evaluadas según sus genes esenciales para el metabolismo). Tal comparación se presenta en la figura 5 (ver valores en las tablas complementarias), en la que se puede corroborar que las redes de IM son mejores que las redes metabólicas para distinguir los genes esenciales, como se presentó en las figuras 1 a 4, y que la unión entre ellas resulta en valores aún más altos.

Adicionalmente, se puede notar que los CGs de las redes metabólicas resultan un poco mejores que las respectivas redes completas, a diferencia de lo que ocurre al evaluar los CGs de las IntNets, donde no se obtiene una mejoría notoria. Esto resalta la importancia de incluir las IM para completar las redes metabólicas buscando una representación que por un lado esté más completa (ver sección I.2.a y sección IV), y por el otro pueda resultar más útil (ver sección V.2.a).

[*] al utilizar las 11 diferentes medidas de centralidad para evaluar las 18 redes metabólicas y las 8 de IM

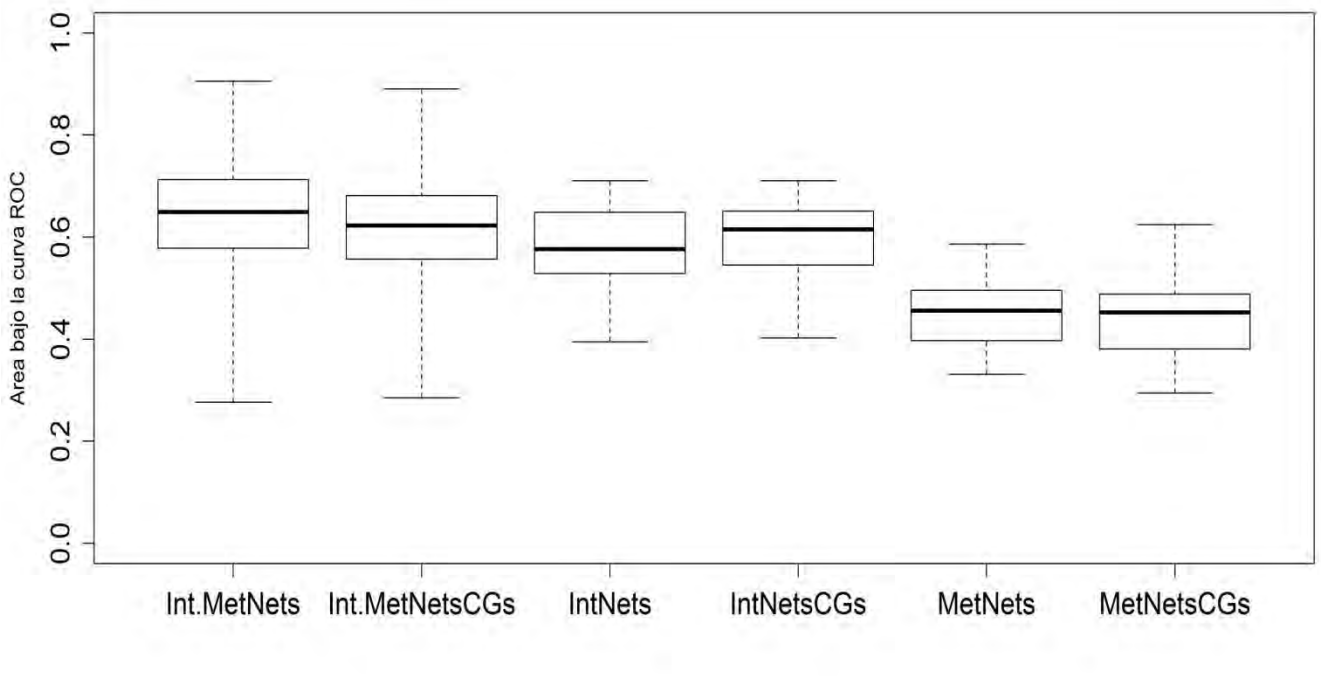


Figura 5 Comparación de los valores del área bajo la curva para todas las centralidades de las redes metabólicas (MetNets), de interacciones moleculares (Int.MetNets) y sus uniones, además de sus componentes gigantes (GCs). Los boxplots representan la mediana en el centro y los valores máximos y mínimos en los extremos

Se calcularon distintos parámetros tanto para las redes metabólicas como para las Int.MetNets buscando analizar más a detalle la distribución de los datos de ambos grupos y evaluar las diferencias observadas de forma adecuada. Así, en la tabla 3 se puede notar que las redes mixtas son mejores hasta un 30% obteniendo valores de ABC de hasta 0.91, además de que el sesgo negativo indica una tendencia hacia los valores más altos. Los valores de la media, mediana y curtosis para los distintos valores de ABC dan idea de distribuciones no normales, lo que fue corroborado mediante las pruebas Shapiro-Wilk ($p= 9.36x10^{-8}$ para Int.MetNets y $p= 9.126\ x10^{-4}$ paraMetNets) y Anderson-Darling ($p= 4.19\ x10^{-7}$ para Int.MetNets y $p= 7.289\ x10^{-3}$ paraMetNets). Las varianzas de ambos grupos no son homogéneas según la prueba Fligner-Killeen ($\chi^2=36.744$, $p= 1.347\ x10^{-9}$), lo que, sumado al hecho de no mostrar distribución normal, implica el

requerimiento de una prueba no paramétrica como la de Wilcoxon para evaluar las diferencias que resultaron significativas ($p<2.2x10^{-16}$).

Grupo	N	Media	Mediana	Desviación estándar	Mínimo	Máximo	Sesgo	Curtosis
MetNets	198	0.45	0.46	0.06	0.33	0.59	0.02	-0.88
IntNets	88	0.59	0.58	0.07	0.39	0.71	-0.23	-0.73
Int.MetNets	1584	0.65	0.65	0.11	0.28	0.91	-0.16	0.24

Tabla 5 Valores de distintos parámetros estadísticos para describir el ABC obtenido con las diferentes medidas de centralidad utilizadas sobre las redes metabólicas (MetNets) las redes de interacciones moleculares (IntNets) y sus uniones (Int.MetNets).

VI.2 Comparación

VI.2.a Medidas generales

Lo primero que se puede notar al comparar las redes metabólicas con las de IM es que las segundas son, en general, más grandes que las primeras contando con un mayor número tanto de vértices como de aristas (Tablas complementarias, gráficos 1 y 2). Al observar las redes Palsson solas, se puede advertir como aumentan el diámetro, la excentricidad y la distancia promedio al retirar los metabolitos más conectados, efecto que deja de notarse al unir estas últimas redes con las de IM. Así las Int.MetNets parecen reflejar principalmente las características de las redes de IM. De forma semejante, se observa que el diámetro de las redes analizadas depende en gran medida de su CG ya que dicho parámetro no varía entre la red y la parte completamente conectada, aún cuando los otros parámetros globales aumenten.

La cercanía es el recíproco de la distancia promedio por lo que ambos índices reflejan el mismo comportamiento. Así, el valor de la cercanía debería disminuir para todos los nodos de los CG en relación a los de la red completa mientras la distancia aumenta. Para comparar la distribución de los valores de

distancia y de cercanía se usan sus promedios, además, la distribución de la cercanía puede ser representada también por el índice OCCI. Los valores del promedio de la cercanía en los CGs son iguales o más bajos que en la red completa y esta diferencia es más notoria al observar los parámetros OCCI.

El hecho de que los CGs presenten valores más altos, con respecto a la red completa, para el OCCI como medida de las distancias dentro de la red, refleja la remoción de distancias muy cortas o cercanas a cero. Al observar otras medidas más locales como el coeficiente de empacamiento y el grado (simple o esférico) se puede notar cierto enriquecimiento que puede ser debido a que las distancias más cortas implican nodos con pocas conexiones que no están relacionados con el CG (Tablas complementarias).

VI.2.b Centralidades y redes

Al observar el panel superior de la figura 6 se puede distinguir que los valores de ABC más altos fueron obtenidos al utilizar centralidades locales como el coeficiente de empacamiento y el grado. En contraste, las medidas globales como la distancia promedio y la excentricidad arrojaron valores menores a 0.5, por debajo de lo que lo haría un clasificador aleatorio. El uso de los inversos aditivos de algunas medidas provoca estimados opuestos de la importancia de un gen. Sin embargo, eso no siempre resulta en una mejoría como ocurre con cluster y excen (Tablas complementarias).

Al comparar los valores clasificados según la red metabólica de donde provienen, como se muestra en la parte media de la figura 6, no se distingue alguna que resulte mejor en forma notoria. Aún así se puede observar cierta tendencia de las redes KEGG por valores más altos.

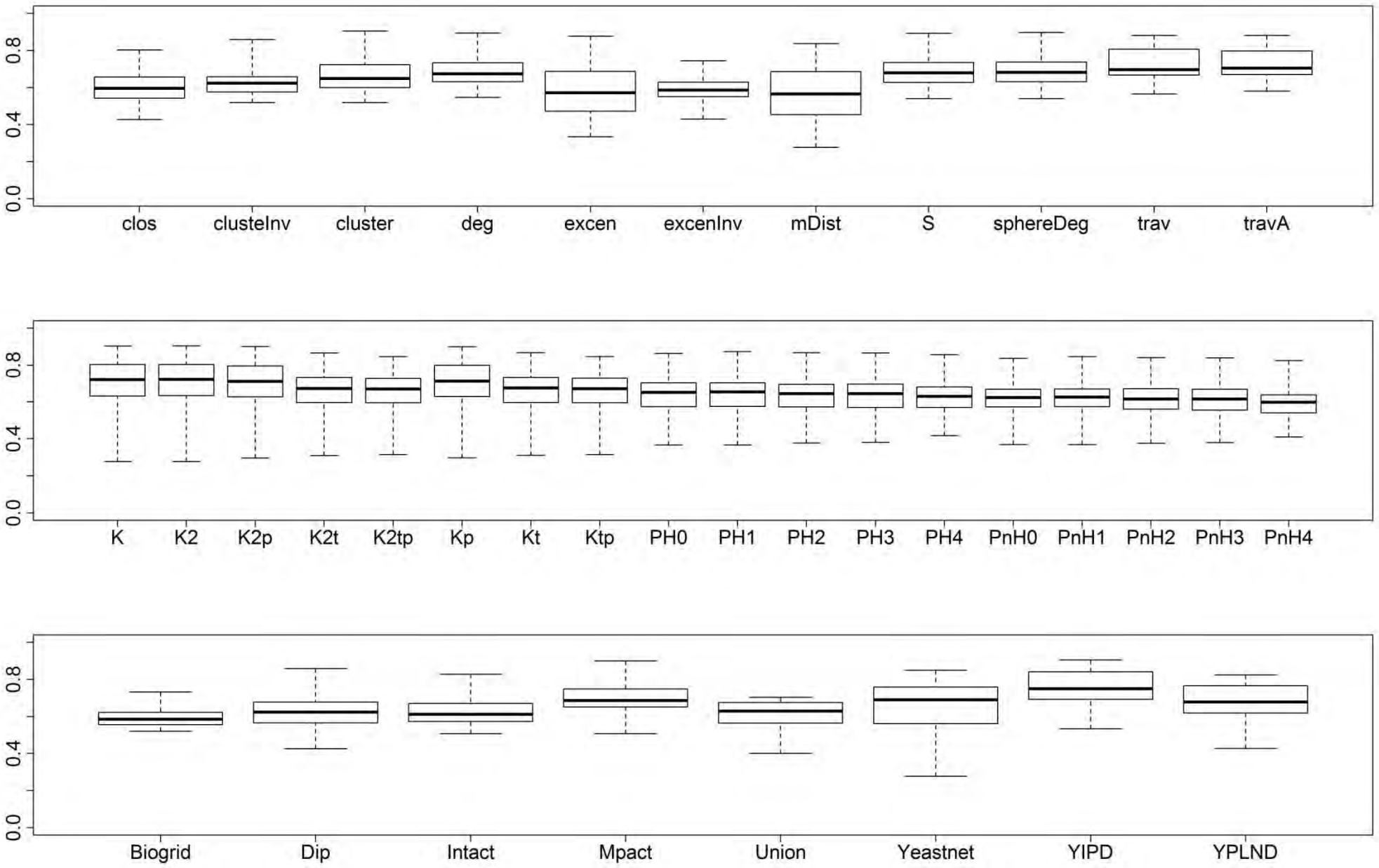


Figura 6 Valores del área bajo la curva para todas las redes Int.MetNets categorizados según las centralidades (arriba), las redes metabólicas (al centro) o las redes de interacciones moleculares (abajo) utilizadas. Los boxplots representan la mediana en el centro y los valores máximos y mínimos en los extremos.

En el panel inferior de la figura 6, donde se compara según la red de IM, las redes obtenidas a partir de las bases de datos YIPD, YPLND y Mpact alcanzan valores notablemente más altos que las demás. Estas últimas bases de datos coinciden en el hecho de que contienen información de IPP ya sea a partir de estudios de alto rendimiento, obtenidas por minería de textos o curadas manualmente por expertos en el área.

VI.3 Genes esenciales. Red REMG

La unión de las interacciones contenidas en las redes KEGG2 y YIPD puede ser considerada como una red REMG ya que, al analizarse según un criterio local de centralidad, como el coeficiente de empacamiento, es posible localizar genes anteriormente descritos como esenciales en forma más eficiente (YIPD.KEGG2cluster: ABC=0.905±0.023; sensibilidad=0.918; especificidad=0.866; exactitud=0.962; mínimo error=0.157; Tablas complementarias) que lo que se puede lograr usando la misma centralidad sobre la red metabólica (KEGG2cluster: ABC=0.574±0.05; sensibilidad=0.530; especificidad=0.684; exactitud=0.78; mínimo error=0.566; Tablas complementarias) o la red de IM subyacentes (YIPDcluster: ABC=0.517±0.021; sensibilidad=0.041; especificidad=0.980; exactitud=0.813; mínimo error=0.959).

A continuación, se analizaron los genes esenciales representados en estas redes para después compararlos con los predichos. Como se puede observar en la figura 7, entre los 609 genes de la red KEGG2, que representan ~10% de los genes de la levadura, se encuentran ~10% del total de genes descritos como esenciales por el YDP (134 de 1156) alrededor de la mitad de los cuales (64 de 134) están contenidos en la red construida a partir de la base de datos YIPD.

Para analizar los genes esenciales predichos, se comparó la forma de la curva ROC construida con la medida de centralidad con la que se obtuvieron los valores más altos con la forma de las curvas de aquellas redes a partir de las

cuales fue construida usando la misma centralidad y el mismo conjunto de genes esenciales.

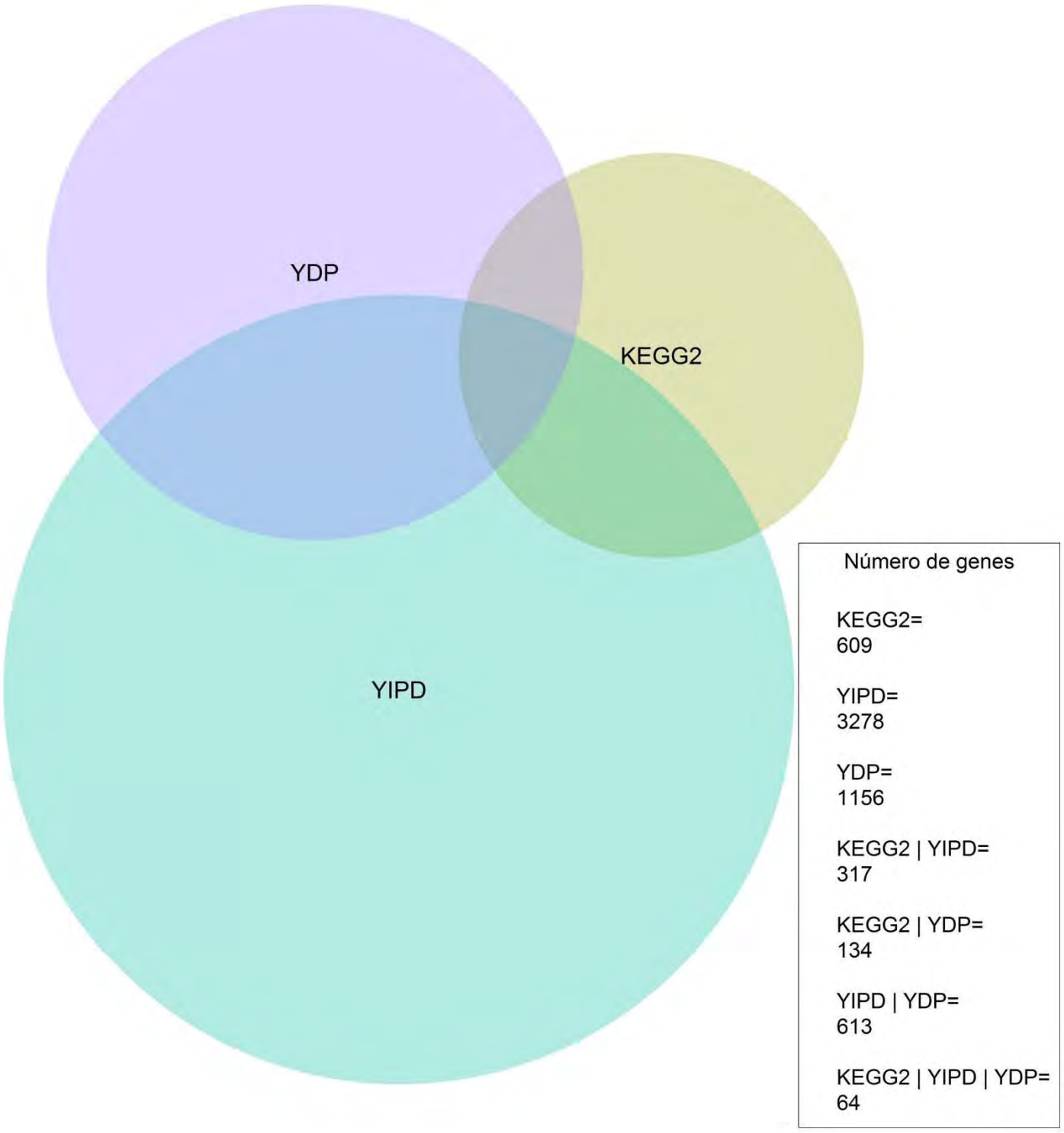


Figura 7 Diagrama de Venn-Euler representando los conjuntos de genes, tanto de las redes KEGG2 y YIPD, como de los genes esenciales descritos por el YDP. En la leyenda se muestra el número total de genes de cada conjunto y de algunas intersecciones entre ellos.

En la figura 8 se puede notar que la curva construida usando la red YIPD.KEGG2 se acerca más a la predicción perfecta que cualquiera de las curvas que resultan de las redes KEGG2 o YIPD partir de las cuales fue construida.

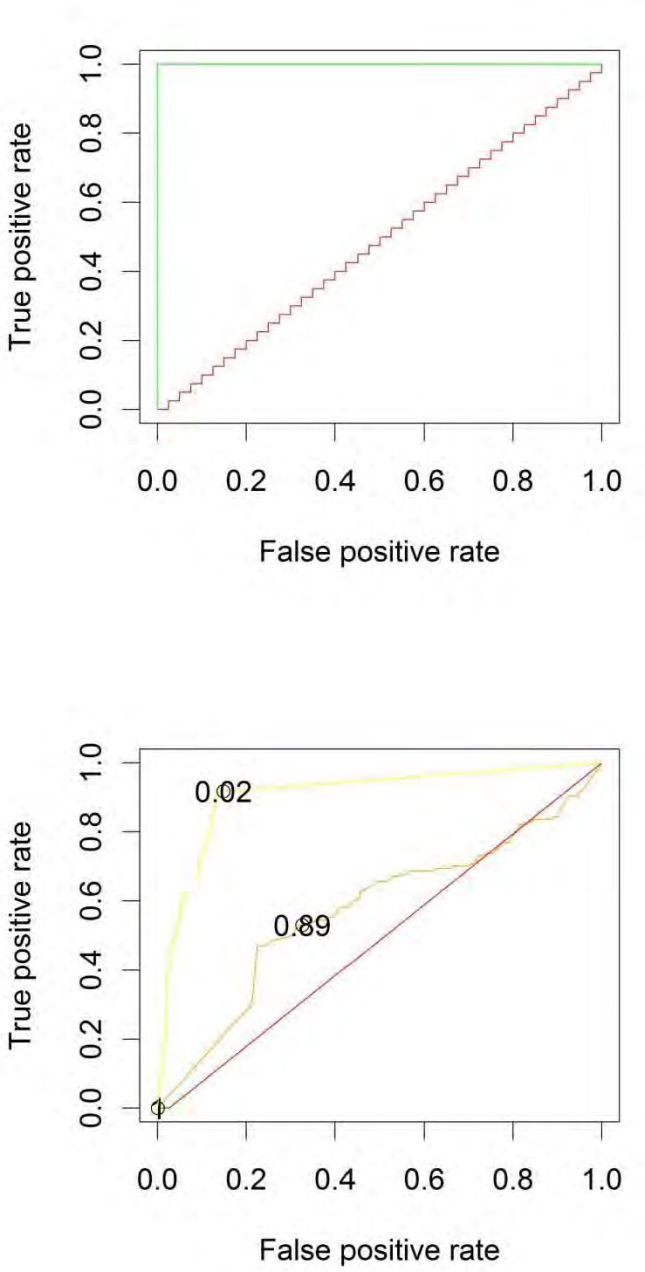


Figura 8 Curvas ROC. En el recuadro de arriba se muestra en verde una curva para un predictor perfecto y en rojo aquella para un predictor aleatorio. En el recuadro de abajo se muestran las curvas obtenidas usando el coeficiente de empacamiento como índice de centralidad para predecir genes esenciales metabólicos en las redes YIPD (rojo), KEGG2 (naranja) y YIPD.KEGG2 (amarillo) mostrando el punto de mínimo error para cada curva (1, 0.8888 y $0.02E^{-2}$, respectivamente).

Para observar un poco más a detalle la construcción de las curvas ROC, en las figuras 9 a 11 se ilustran los valores de sensibilidad, especificidad y error para cada valor de coeficiente de empacamiento utilizado como punto de corte para comparar la fracción de genes esenciales y no esenciales predichos en la red YIPD.KEGG2 y en aquellas de donde proviene. De ese modo se puede notar que los genes esenciales para el metabolismo representados en la red YIPD (figura 9) se pueden caracterizar por tener valores de cluster iguales a 0, o sea tienen solo una conexión. No obstante, al unir esta red con la red KEGG2 (figura 10) y evaluarla según los mismos genes metabólicos, se obtuvieron resultados más satisfactorios que con cualquiera de las dos redes por separado (figura 11).

Al analizar detenidamente los 134 genes esenciales a predecir se puede notar que en el punto de mínimo error obtenido con la red YIPD.KEGG2 es posible catalogar hasta 123 genes esenciales dejando solamente 11 sin ser clasificados (figura 11). En contraste, con la red metabólica KEGG2 se obtienen solamente 71 verdaderos positivos de 223 predichos como esenciales (figura 10). 52 de los genes esenciales predichos correctamente con la red YIPD.KEGG2 no se catalogan como tal con la red KEGG2 y los 71 restantes se pueden clasificar con cualquiera de las dos redes. Esto indica que es posible predecir más genes esenciales con la red YIPD.KEGG2 además de los que ya se podían predecir con la red KEGG2 debido, probablemente, a que se establecieron nuevas conexiones.

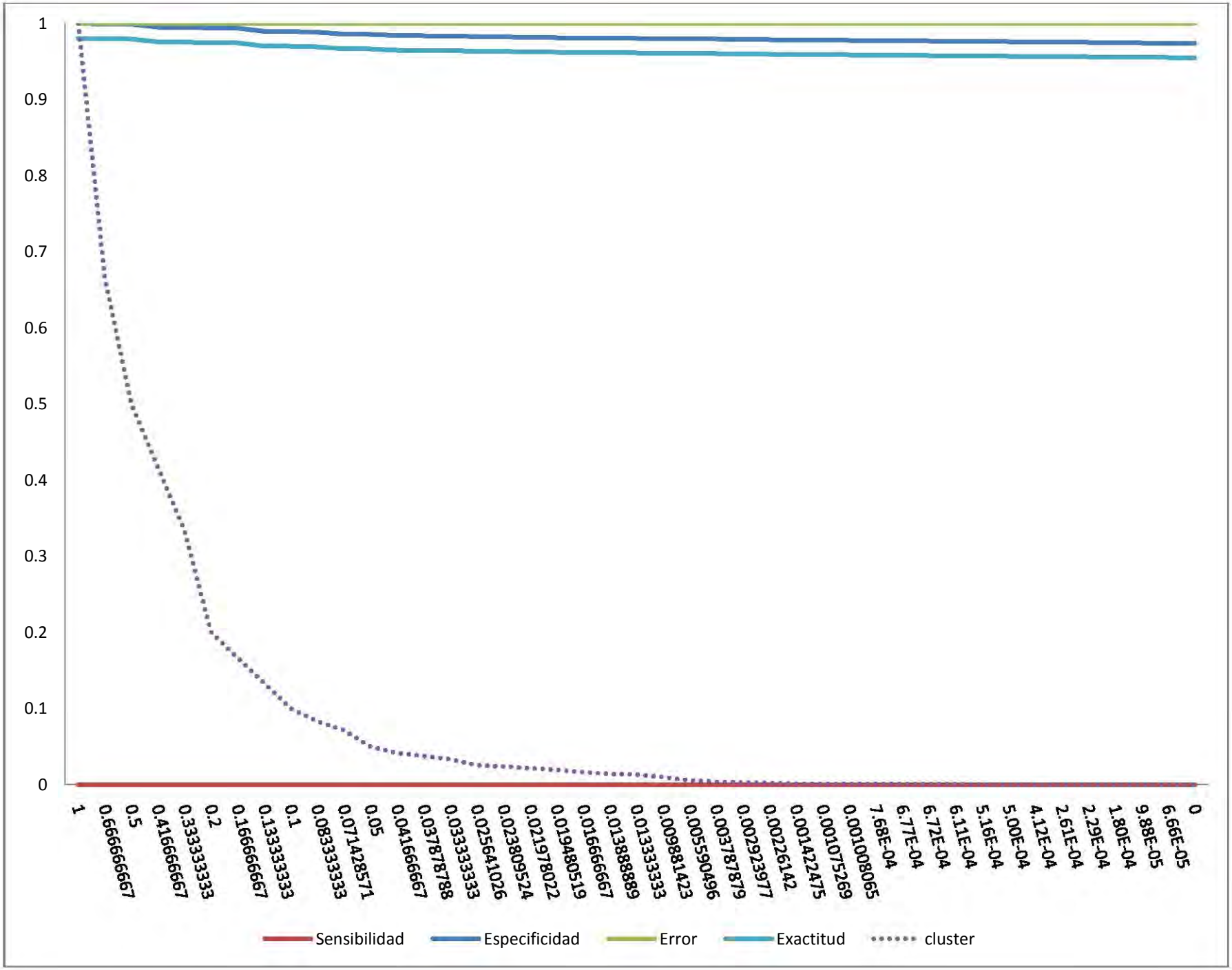


Figura 9 Comparación entre la sensibilidad, especificidad, error y exactitud para cada valor de corte usando el coeficiente de empacamiento de la red YIPD. Debajo de la gráfica se muestran los valores de cada punto de corte.

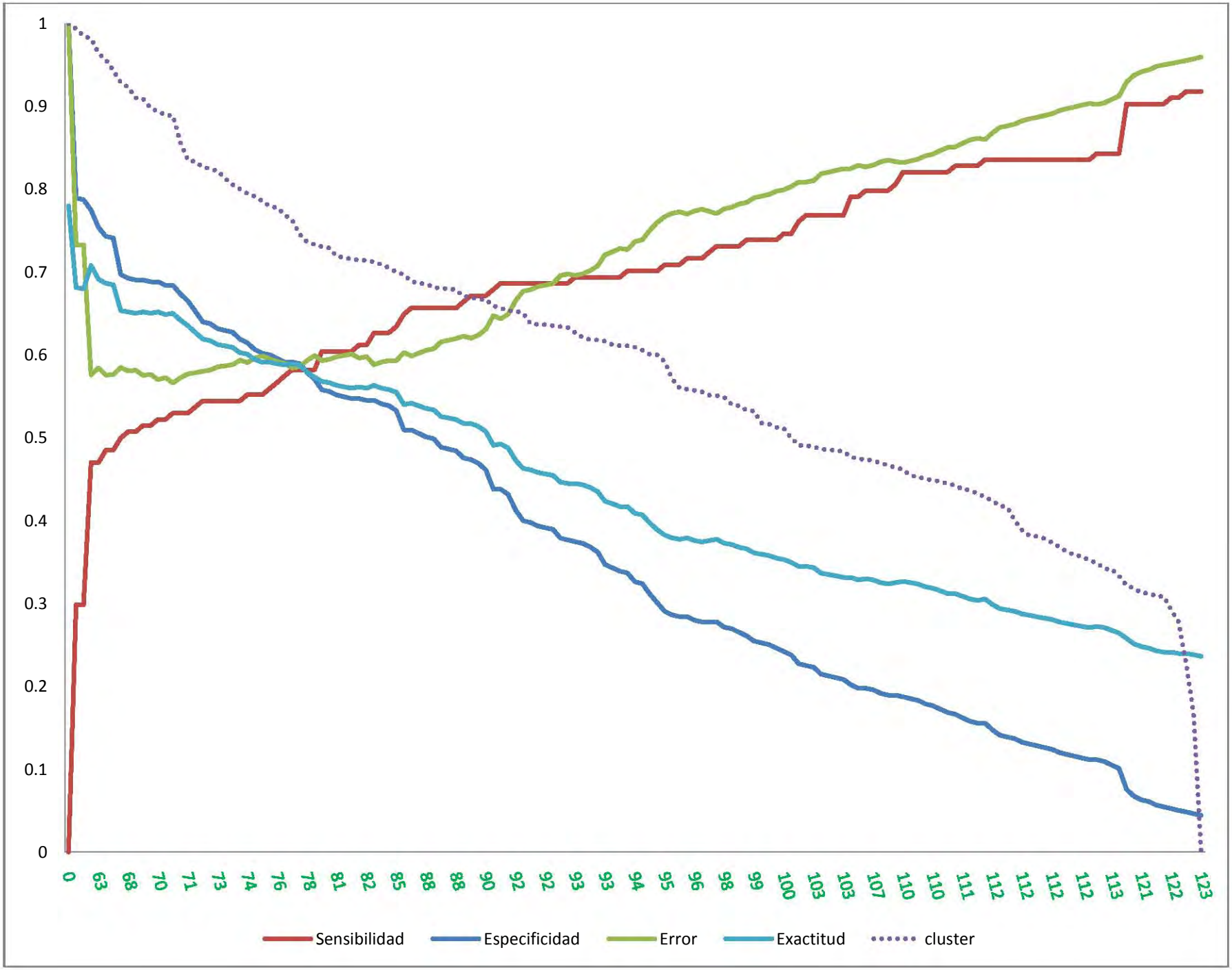


Figura 10 Comparación entre la sensibilidad, especificidad, error y exactitud para cada valor de corte usando el coeficiente de empacamiento de la red KEGG2. Debajo de la gráfica se muestran los verdaderos positivos en verde y los falsos negativos en rojo.

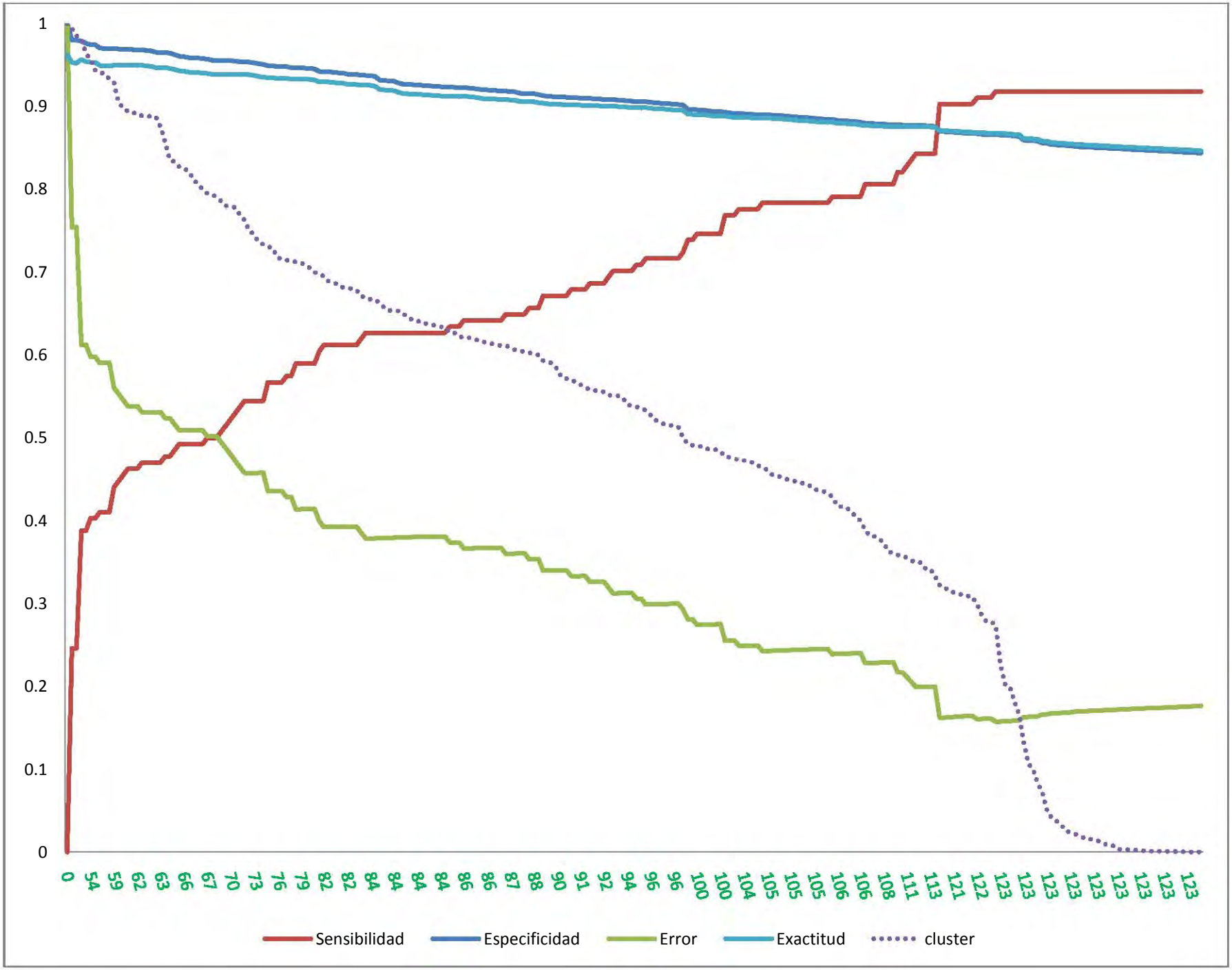


Figura 11 Comparación entre la sensibilidad, especificidad, error y exactitud para cada valor de corte usando el coeficiente de empacamiento de la red YIPD.KEGG2. Debajo de la gráfica se muestran los verdaderos positivos en verde.

VII. DISCUSIÓN

Las redes construidas en este trabajo tratan de resolver las limitaciones observadas anteriormente al cuestionar la posibilidad de que la información incluida en las redes metabólicas está incompleta y/o de que los genes esenciales que componen éstas redes no se pueden distinguir por no estar conectados con el CG. Al evaluar los CGs de las redes metabólicas y compararlos con las redes completas se observó cierta mejoría en su capacidad de predecir genes esenciales. Dicha mejoría fue superada al unir la información de las redes metabólicas con la de las redes de IM en las llamadas Int.MetNets y evaluarlas según los genes esenciales metabólicos que las componen (Sección VI.1.c; Figura 4). Así se obtuvo la red REMG llamada YIPD.KEGG2 a partir de las redes KEGG2 (MetNet) y YIPD (IntNet) de manera que no contiene los metabolitos más conectados, no está restringida a los genes cuyo producto proteico tiene clasificación EC y le fue añadida información obtenida mediante un experimento de dobles híbridos a gran escala.

Debido a la importancia de los genes esenciales en campos como la genómica funcional y la farmacología existen varios intentos por predecirlos. Algunos de ellos utilizan conjuntos de datos que también se usan en este trabajo y su análisis mediante el mismo método permite una comparación. Saha y Heber (2006) compararon características genómicas, filogenéticas y topológicas para predecir los genes descritos como esenciales en el YDP y obtuvieron un ABC=0.82 combinando el uso de métodos de inteligencia artificial sobre un conjunto balanceado de datos con número igual de genes esenciales y no esenciales. Acencio y Lemke (2009) construyeron una red para *S. cerevisiae*[*] y, utilizando características topológicas, información de localización subcelular y procesos biológicos obtuvieron, en el mejor de los casos, un ABC=0.808. Ambos valores de ABC son superados por el que se obtiene con la red YIPD.KEGG2

[*] a partir de los datos de IPP de la base de datos BioGRID e interacciones de regulación transcripcional experimentalmente determinadas presentes en la base de datos TRANSFAC

ABC=0.905±0.023 usando un intervalo de confianza del 99% (Sección VI.3; Tablas complementarias).

Es importante destacar que existe una diferencia fundamental entre este estudio y sus antecedentes. El ABC reportado aquí se obtuvo usando el conjunto de los genes esenciales del metabolismo (ver sección V.2.a), mientras que el de los trabajos anteriores fue conseguido usando todos los genes esenciales. Esto hace imprecisa la comparación pero es útil para contrastar los distintos enfoques. Aquí se espera predecir los genes esenciales para después poder analizar su función dentro del metabolismo. Mientras tanto, los esfuerzos anteriores buscan, por un lado, predecir genes críticos en forma confiable buscando facilitar su identificación en organismos en los que éstos no están descritos experimentalmente y por el otro, validar en forma empírica los conjuntos de genes esenciales verificando los que ya están descritos y proponiendo nuevos.

Entre las redes de IM, aquellas construidas a partir de las bases de datos BioGrid y DIP fueron las que obtuvieron mejores resultados al ser analizadas en sí mismas mediante su contenido de genes esenciales y usando una centralidad local como el grado esférico (BioGrid sphereDeg ABC=0.681; DIP sphereDeg ABC=0.676; Tablas complementarias). Si bien estos resultados son mejores que cualquiera de los obtenidos a partir de las redes metabólicas (KEGG2path clos ABC=0.586 y PnH4 clos ABC=0.331 siendo el máximo y mínimo respectivamente; Tablas complementarias), no se puede concluir que la función de los genes del metabolismo se deba representar exclusivamente a partir de las IM o de las relaciones entre enzimas y sus metabolitos. Por ello, las IM resultan una buena opción para enriquecer la capacidad predictiva de las redes metabólicas más allá considerarlas individualmente. Esta unión entre distintos tipos de redes arroja resultados que no se pueden intuir a partir de los conjuntos usados por separado. Un ejemplo de este fenómeno es el hecho de que la mejor de las redes metabólicas y la mejor de las redes de IM no componen la mejor de las redes Int.Met. Tampoco sucede que la mejor de las centralidades anteriores (sphereDeg

o clos) es la que distingue a la red REMG ni mucho menos que este valor sea resultado de la simple suma de las ABC de cada red.

Siguiendo con la discusión de las mejores redes de IM, la base de datos BioGrid incluye tanto interacciones genéticas como IPP obtenidas ya sea por estudios de alto rendimiento como por co-citación en la literatura y curadas manualmente mientras que la base de datos DIP cuenta con las mismas consideraciones sin tomar en cuenta las interacciones genéticas. En cuanto a las mejores redes Int.Met, éstas incluyen datos de las redes YIPD y MpAct que contienen IPP a gran escala curadas manualmente y por co-citación sólo en el segundo caso. El hecho de que las interacciones de la base de datos Mpact sean del tipo IPP corrobora lo observado con las redes BioGrid y DIP revelando limitaciones al usar co-citación pero apoyando el uso de IPP curadas manualmente. Que la información de la red Yeastnet no figure en ninguno de los casos se suma a la observación de que las interacciones genéticas de la red BioGrid no mejoran visiblemente la predicción de genes esenciales metabólicos (Sección V.2.c, Tabla 2; Tablas complementarias).

También es necesario resaltar que existen dificultades que emergen al aumentar el tamaño de las redes (Bonholdt 2005). De tal forma, un conjunto gigantesco que presente todas las interacciones posibles puede terminar siendo prohibitivamente grande como para permitir su análisis y no necesariamente tendría que resultar mejor que uno pequeño. Un fenómeno de este tipo se puede observar al comparar la red Union con las demás, ya que contiene más interacciones que cualquiera pero no mejora la capacidad de distinguir a los genes esenciales (Figuras 3, 4 y 6).

Con lo anterior se muestra que el tamaño podría ser un factor irrelevante, por ello, al modelar sistemas biológicos mediante relaciones genéticas se debe estar consciente de su gran diversidad y determinar cuáles son las más relevantes para el propósito que se persigue. La experiencia obtenida en otros campos ha

demostrado que se pueden construir modelos poderosos y útiles sin tener que conocer todo sobre el sistema, de hecho, una parte fundamental en la construcción de modelos consiste en determinar qué es lo que se necesita para construir una representación útil (Palsson 2000). Una manera de simplificación podría resultar de ignorar detalles moleculares (como las interacciones genéticas experimentales, probabilísticas o las de co-citación; ver arriba) mientras sea posible.

Resultados previos demostraron que no existen medidas de centralidad que por sí mismas sean capaces de identificar genes esenciales de forma confiable en redes metabólicas comúnmente usadas revelando la naturaleza compleja de la función de genes esenciales para la viabilidad (del Río 2009). Las redes presentadas aquí cuentan con una mayor probabilidad de identificar genes esenciales usando una sola medida de centralidad y sugieren que la función de los genes críticos del metabolismo depende de relaciones químico-genéticas y de interacciones proteína-proteína que no habían sido consideradas en el estudio anterior. En este trabajo no se evaluó el efecto de combinar medidas de centralidad como en el anterior y por ello aún no se puede descartar la utilidad de tal enfoque.

El análisis topológico de las redes identifica las propiedades cualitativas del sistema mediante las distintas medidas de centralidad descritas dentro de la teoría y el análisis de redes. Un ejemplo característico es la interconectividad de la red que puede ser caracterizada mediante el diámetro. Se ha demostrado que el diámetro incrementa ante la remoción de los nodos (sustratos) más conectados (Jeong 2000). Esto mismo sucede con las redes utilizadas aquí, siendo más notorio en las redes Palsson que en sus derivadas (Tablas complementarias).

Si la conectividad promedio de una red es fija su diámetro incrementaría logarítmicamente con la adición de nuevos nodos. Para las redes metabólicas esto implica que una bacteria más compleja tendría un diámetro mayor que una

bacteria simple. No obstante, el diámetro de la red metabólica de por lo menos 43 organismos es el mismo independientemente del número de sustratos (Jeong 2000). Esto podría ser evidencia de presión selectiva (como restricción física), sin embargo, considerando que el procedimiento de reconstrucción de esas redes está basado en homología, podría pensarse que las limitaciones al crecimiento de este parámetro son consecuencia del proceso de reconstrucción y no una característica de los organismos estudiados.

En el caso de la red KEGG2, su unión con la red YIPD aumentó su diámetro (Tablas complementarias) y la volvió más predictiva (Sección VI.3; figura 8), sugiriendo una vez más que la magnitud de este parámetro no necesariamente está restringido entre las redes metabólicas. No obstante, ya que pueden existir interacciones de la red YIPD que no participan en el metabolismo, es posible que muchas de estas relaciones y genes debieran eliminarse. En este trabajo no se realizó dicha depuración de genes ya que no es un proceso trivial y la forma de hacerlo sigue siendo una pregunta abierta (como se comenta arriba). Este análisis puede justificar la realización de dicho proceso de eliminación para permitir una evaluación objetiva del diámetro de las redes metabólicas.

Jeong y colaboradores (2001) han propuesto que si este fenómeno fuera debido a un componente topológico, las proteínas más conectadas deberían probar ser esenciales, por lo menos en promedio, al compararlas con las proteínas menos conectadas. Por eso, en ese mismo año, ordenaron todas las proteínas de la base de datos DIP según su grado para buscar correlación con el efecto fenotípíco de su remoción individual. De tal forma, las proteínas altamente conectadas dentro de una red de IPP resultaron tener una probabilidad de ser esenciales hasta tres veces mayor que las proteínas con pocas interacciones. En nuestro estudio, las medidas de centralidad que arrojaron mejores resultados no incluyen al grado pero comparten su sentido local. El hecho de que las mejores centralidades hayan sido el coeficiente de empacamiento y el grado esférico no apoya la hipótesis de que los genes con mayor número de conexiones sean

esenciales pero es congruente con la clasificación de los genes según sus interacciones locales para encontrar genes esenciales.

Estos resultados apoyan las observaciones de otros grupos (Pržulj 2004; Friedel y Zimmer 2006; Hormozdiari 2007) que, al analizar las redes de IPP, observaron que los valores altos de grado no son la mejor medida para clasificar genes esenciales y que no son caracterizadas por distribuciones puras de escalamiento libre. Ya se ha discutido que la cobertura alcanzada por las distintas aproximaciones para describir IPP es limitada además de la posible redundancia en la anotación funcional y la existencia de falsos positivos y falsos negativos como resultado de los análisis de donde se obtienen los conjuntos de datos (Dunker 2005). Por ello, aún no se puede decir que exista una prueba contundente de que las redes biológicas sigan de algún tipo particular de distribución sino hasta que se mejore la calidad de los datos y los métodos con los que las construyen y evalúan.

El hecho de que las medidas locales, como el coeficiente de empacamiento y el grado esférico, sirvan para encontrar genes esenciales en redes metabólicas y/o de IM contrasta con lo que ocurre con las redes de estructura de proteínas, donde una medida global resulta más predictiva (Thibert 2005). Además, sugiere que las bases de datos ricas en genes con interacciones locales pueden ser valiosas para reconstruir redes cada vez más coherentes con el fenotipo representado por un conjunto de genes esenciales. Lo anterior se puede corroborar al observar lo que ocurre al integrar la red KEGG2 con la base de datos YIPD, construida con interacciones anotadas según experimentos de dobles híbridos y, por tanto, locales (Tablas 1 y 2; Tablas complementarias).

El coeficiente de empacamiento refleja grupos de nodos altamente interconectados que pueden representar complejos proteicos (también llamados hubs al analizarse en redes de IPP). Experimentalmente se ha demostrado que los hubs tienden a estar compuestos de manera uniforme, ya sea por proteínas

esenciales o por proteínas no esenciales (Barabási y Oltvai 2004). Esto indica que cada nodo puede presentar distintas capacidades y funciones que dependen tanto de sí mismo como de sus vecinos a una y dos conexiones de distancia, por lo que su esencialidad no puede depender únicamente del grado tal como sucedió aquí.

Esta relación observada entre el coeficiente de empacamiento y la esencialidad de los genes apoya algunas otras conclusiones evolutivas interesantes realizadas anteriormente (Dunker 2005; Barabási y Oltvai 2004). Pero como cualquier otro intento por explicar la función de las proteínas presenta complicaciones entre las que se encuentra la anotación funcional realizada mediante conservación de dominios o firmas funcionales (The Gene Ontology Consortium 2004). También al intentar explicar fenotipos observados en base a su función se pueden presentar complicaciones al tratar con proteínas multifuncionales o con múltiples dominios (como las proteínas descritas como "moonlighting" (Jeffery 2003)).

Un tema común sobre la función de las proteínas es que generalmente son componentes de complejos que contienen otras macromoléculas y llevan a cabo procesos biológicos específicos conectados por redes de interacciones. Estos conceptos tienen varias implicaciones: 1)si la función de una proteína se conoce ésta puede servir para predecir la función de por lo menos algunos de sus vecinos, 2)cada proteína puede participar en distintos complejos asociados a diferentes funciones, 3)la comunicación entre procesos implica IPP que conectan complejos, 4)la función de un complejo descrito por primera vez puede ser definida por la función de la mayoría de sus miembros, 5)las redes biológicas pueden exhibir propiedades emergentes que se comprenden mejor después de que son descritas todas, o por lo menos la mayoría, de sus conexiones, 6)confiar en anotación o redes incompletas puede llevar a conclusiones sesgadas o erróneas (Cusick 2005). Así, la capacidad de predecir genes esenciales usando información químico-genética relaciona la esencialidad de los genes metabólicos con su función enzimática, pero el hecho de que las IM mejoren dicha capacidad añade

elementos para tratar de explicar la esencialidad de éstos genes. Esta explicación concuerda con el sentido local observado en otros estudios y con los seis puntos discutidos arriba.

La función de un gene puede depender de 2 factores, la estructura de la red y la dinámica de la red. Por estructura de la red se entiende los genes y sus relaciones funcionales. Por dinámica de la red se entiende la variación que sufre la estructura a lo largo del tiempo y/o el flujo asociados a las conexiones en la red. El hecho de que se puedan predecir casi todos los genes esenciales del metabolismo a partir de la estructura de la red hace pensar que la función de los genes esenciales metabólicos depende poco o nada de la dinámica de la red, sin embargo, hasta que no se realice un análisis de este tipo no se puede hacer tal afirmación. Esto puede significar que, en el medio en el que se hicieron los experimentos para detectar a los genes esenciales aquí utilizados (YPD; medio rico), el metabolismo debe sufrir poca o ninguna regulación en su estructura: hay muchos metabolitos todo el tiempo que procesar, todos los genes y sus funciones se requieren la mayor parte del tiempo, por lo que no hay necesidad de cambiar flujos y/o estructura metabólica. Una pregunta que se deriva de estos resultados es si los genes que no se pueden predecir son aquellos cuya función si depende de la dinámica o son genes que no se pueden predecir todavía porque aún faltan relaciones funcionales en la red del metabolismo. Este aspecto podría ser relevante para continuar explorando esta línea de investigación.

VIII. PERSPECTIVAS

Este trabajo es un esfuerzo por entender, mediante el uso de herramientas matemáticas y computacionales, qué tanto influye la estructura de las redes metabólicas para determinar características importantes como la esencialidad de los genes. Aquí se abordó la optimización de las capacidades predictivas de modelos a escala genómica esperando seguir una tendencia por llenar los espacios vacíos en el conocimiento de forma sistemática. Con este fin se utilizaron bases de datos públicas con relaciones genéticas diversas (entre las que no existe una sola interacción reportada al mismo tiempo) para ampliar modelos de relaciones químico-genéticas del metabolismo. De tal modo, la información de las redes de IM mejoró la predicción de genes metabólicos al unirlas con las redes metabólicas, ya que las uniones resultaron más predictivas que cualquiera de los dos tipos de redes por sí solas. Esto relaciona la esencialidad de los genes metabólicos con la existencia de funciones enzimáticas asociadas mediante interacciones locales de diversos tipos (como las interacciones físicas directas reveladas en estudios de dobles híbridos).

Se logró observar que estas redes Int.MetNets incluyen genes esenciales para *S. cerevisiae* y que hay medidas locales de centralidad que son capaces, por sí mismas, de identificarlos con una probabilidad mayor a la que se obtendría con un clasificador aleatorio o con cualquiera de los probados anteriormente. Estos resultados son, en general, estadísticamente confiables ya que con este nivel de confianza se espera un 1% de error entre las 1870 predicciones, es decir, se podrían esperar hasta 19 predicciones correctas en forma aleatoria y se observaron muchas más de este número.

En este caso, el coeficiente de empacamiento y el grado esférico son las medidas con las que se obtuvieron mejores resultados. Estas observaciones también nos permiten hacer dos propuestas para mejorar la reconstrucción de redes metabólicas buscando lograr una mejor representación de un sistema

biológico. La primera tiene que ver con los genes esenciales que no se han logrado predecir y buscar nuevos tipos de interacciones para tratar de conectarlos en forma local esperando facilitar su clasificación. Así, una vez que se haya construido una red REMG, se pueden incluir las relaciones genéticas que les falten los genes esenciales no predichos o incluir nuevas interacciones para los genes esenciales predichos para lograr una mayor probabilidad de encontrarlos. Esto representa un número finito de experimentos y las nuevas relaciones genéticas se pueden entonces evaluar por criterios de centralidad, y este proceso puede continuar hasta que no se alcance mayor mejoría. La segunda consiste en explorar los cambios funcionales y dinámicos causados por distintas perturbaciones sobre los genes ordenados según estos últimos dos criterios locales. Esto debido a que aún no es posible afirmar o descartar con certeza el hecho de que los elementos estructurales sean el único factor que determina la función de una red biológica. Es de esperar que algunos de los genes esenciales solo se puedan predecir en modelos dinámicos, que no han sido considerados hasta el momento.

La biología celular está cambiando de asignar funciones a genes o proteínas individuales a considerar a las interacciones que dan lugar a los distintos programas celulares mediante las complejas redes de los sistemas biológicos (Hartwell 1999). Muchos de los esfuerzos en la Biología han sido inspirados por la observación de la existencia de semejanzas entre propiedades y mecanismos moleculares llevando a principios generales que permiten predecir y comprender a los sistemas vivos. Todos estos principios conducen a relaciones invariantes, aspecto de la biología de sistemas en el que pudieran parecerse a las leyes físicas y, por tanto, puede llevar al desarrollo de perspectivas fundamentales sobre los principios que subyacen a la Biología (Bruggeman y Westerhoff 2006). La combinación sinérgica y el desarrollo de experimentación cuantitativa, modelado y teoría es un acercamiento prometedor para llevar a la biología al nivel de sistemas.

IX. CONCLUSIONES

Se construyeron redes a partir de relaciones químico-genéticas e IM diversas que presentan características observadas en otras reconstrucciones de sistemas biológicos. Usando estas redes optimizadas se demostró que las limitaciones en las capacidades predictivas de las redes metabólicas no dependen de la presencia o ausencia de los genes esenciales en CG de la red sino de la falta de interacciones. La información más útil para encontrar los genes esenciales de una red metabólica se obtuvo a partir de estudios de dobles híbridos a gran escala (IPP). Las medidas de centralidad con las que se obtuvieron los mejores resultados son el coeficiente de empacamiento y el grado esférico. Dichos índices evalúan las conexiones locales de cada nodo, lo que es congruente con el tipo de interacciones añadidas y relaciona la esencialidad de los genes metabólicos con la existencia de funciones enzimáticas asociadas mediante interacciones locales.

En este trabajo evaluamos datos experimentales en forma iterativa permitiendo hacer propuestas para la construcción y evaluación de nuevos modelos. Mediante un ejercicio sistemático se identificaron nuevas propiedades de las redes metabólicas que no se habrían inferido mediante los enfoques tradicionales en los que se considera la función de un gene en un contexto aislado.

BIBLIOGRAFÍA

Acencio y Lemke. «Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information.» *BMC Bioinformatics*, 2009: 10:290.

Albert, R. «Scale-free networks in cell biology.» *Journal of Cell Biology*, 2005: 118: 4947–4957.

Arigoni, F, et al. «A genome based approach for the identification of essential bacterial genes.» *Nature Biotechnology*, 1998: 16: 851–856.

Barabási y Oltvai. «Network biology: understanding the cell´s functional organization.» *Nature Reviews Genetics*, 2004: 101-113.

Becker, R , et al. *The New S Language.* New York: Chapman & Hall, 1988.

Becker y Palsson. «Three factors underlying incorrect in silico predictions of essential metabolic genes.» *BMC Systems Biology*, 2008: 2:14.

Bonholdt, S. «Less is more in modelling large genetic networks.» *Science*, 2005: 310: 449-450.

Bruggeman y Westerhoff. «The nature of systems biology.» *TRENDS in Microbiology*, 2006: 15: 45-50.

Chalker y Lunsford. «Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach.» *Pharmacology & Therapeutics*, 2002: 95(1):1-20.

Clayton, R, et al. «The first genome form the third domain of life.» *Nature*, 1997: 387: 459-462.

Costanzo, M, et al. «The Genetic Landscape of a Cell.» *Science*, 2010: 327: 425-431.

Cusack, M, et al. «Efficient identification of critical residues based only on protein structure by network analysis.» *PLoS ONE*, 2007: 2(5):e421.

Cusick, M, et al. «Interactome: gateway into systems biology.» *Human Molecular Genetics*, 2005: 14(2):R171–R181.

De Keersmaecker, S, et al. «Integration of omics data: how well does it work for bacteria?» *Molecular Microbiology*, 2006: 62(5):1239–1250.

del Río, G, et al. «How to identify essential genes from molecular networks?» *BMC Systems Biology*, 2009: 3:102.

Duarte, N, et al. «Reconstruction and validation of Saccharomyces cerevisiae metabolic network iND750, a fully compartmentalized genome-scale metabolic model .» *Genome Research*, 2004: 13(2):244-253.

Dunker, A, et al. «Flexible nets. The roles of intrinsic disorder in protein interaction networks.» *FEBS Journal*, 2005: 272:5129–5148.

Feist, A, et al. «Reconstruction of biochemical networks in microorganisms.» *Nature Reviews Microbiology*, 2009: 7:129-143.

Fell, D. *Understanding the Control of Metabolism.* Portland Press, 1997.

Förster, J, et al. «Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.» *Genome Research*, 2003: 13: 244-254.

Förster, J, et al. «Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae.» *OMICS*, 2003b: 7(2): 193-202.

Francke, C, et al. «Reconstructing the metabolic network of a bacterium from its genome.» *TRENDS in Microbiology*, 2005: 13(11):550-558.

Friedel, C y Zimmer, R. «Inferring topology from clustering coefficients in protein-protein interaction networks.» *BMC Bioinformatics*, 2006: 7:519.

Goffeau, A, et al. «The yeast genome directory.» *Nature*, 1997: 387: 5-6.

Güldener, U, et al. «MPact: the MIPS protein interaction resource on yeast.» *Nucleic Acids Research*, 2006: 34:436-441.

Hanley y Mc Neil. «The meaning and use of the area under the receiver characteristic (ROC) curve.» *Radiology*, 1982: 143: 29-36.

Hartwell, L, et al. «From molecular to modular cell biology.» *Nature*, 1999: 402 supp C47-C51.

Hasty y Collins «Protein interactions. Unspinning the web.» *Nature*, 2001: 411:30-31.

He, X y Jianzhi, Z. «Why Do Hubs Tend to Be Essential in Protein Networks?» *PLoS Genet*, 2006: 2(6): e88.

Hermjakob, H, et al. «IntAct: an open source molecular interaction database.» *Nucleic Acids Research*, 2004: 32:452-455.

Herrgård, M, et al. «Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae.» *Genome Research*, 2006: 16: 627-635.

Hoffmann y Valencia. «Protein interaction: same network, different hubs.» *TRENDS in Genetics*, 2003: 19(12):681-683.

Hormozdiari, F, et al. «Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution.» *PLoS Computational Biology*, 2007: 3(7):e118.

Huber, W, et al. «Graphs in molecular biology.» *BMC Bioinformatics*, 2007: 8(Suppl 6):S8.

Hwang, Y-H, et al. «Predicting essential genes based on network and sequence analysis.» *Molecular BioSystems*, 2009: 5:1672-1678.

Ideker, T, et al. «A new approach to decoding life: systems biology.» *Annual Revisions on Genomics and Human Genetics*, 2001: 2: 343-372.

Ito, T, et al. «A comprehensive two-hybrid analysis to explore the yeast protein interactome.» *Proccedings of the National Academy of Science*, 2001: 98(8): 4569–4574.

Ito, T, et al. «Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.» *Proccedings of the National Academy of Science*, 2000: 97(3): 1143-1147.

Han, J-D, et al. «Evidence for dynamically organized modularity in the yeast protein–protein interaction network.» *Nature*, 2004: 430:88-93.

Jeffery, C. «Moonlighting proteins: old proteins learning new tricks.» *Trends in Genetics*, 2003: 19:415–417.

Jeong, H, et al. «Lethality and centrality in protein networks.» *Nature*, 2001: 411:41-42.

Jeong, H, et al. «The large-scale organization of metabolic networks.» *Nature*, 2000: 407:651-654.

Joyce y Palsson. «The model organism as a system: integrating "omics" data sets.» *Nature*, 2006: 7:198-210.

Kitano, H. «Systems biology: a brief overview.» *Science*, 2002: 295: 1662-1664.

Lamichhane, G, et al. «A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis.» *Proccedings of the National Academy of Science*, 2003: 100: 7213–7218.

Lee, I, et al. «An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae.» *PLoS ONE*, 2007: 10: e988.

Ma y Zeng. «Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.» *Bioinformatics*, 2003: 19(2):270-277.

Ma y Zeng. «The connectivity structure, giant strong component and centrality of metabolic networks.» *Bioinformatics*, 2003b: 19(11):1423-1430.

Maslov y Sneppen. «Specificity and stability in topology of protein networks.» *Science*, 2002: 296:910-913 .

Ogata, H, et al. «KEGG: Kyoto Encyclopedia of Genes and Genomes.» *Nucleic Acids Research*, 1999: 27(1):29-34.

Oltvai y Barabási. «Life’s complexity pyramid.» *Science*, 2002: 298: 763-764.

Orwand, J, et al. *Mastering algorithms with PERL.* Sebastopol CA: O'Reilly & Associates, Inc., 1999.

Palsson, B. «In silico biology through "omics".» *Nature*, 2002: 20:649-650.

Palsson, B. «The challenges of in silico biology.» *Nature Biotechnology*, 2000: 18:1147-1150.

Pržulj, N, et al. «Modeling interactome: scale-free or geometric?» *Bioinformatics*, 2004: 20 (18): 3508–3515.

Reed, J, et al. «An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).» *Genome Biology*, 2003: 4:R54.

Rocha, I, et al. «SOofptwtarFe lux: an open-source software platform for in silico metabolic engineering.» *BMC Systems Biology* , 2010: 4:45.

Kerrien, Y, et al. «IntAct—open source resource for molecular interaction data.» *Nucleic Acids Research*, 2007: 35:561-565.

Saha y Heber. «In silico prediction of yeast deletion phenotypes.» *Genetics and Molecular Research*, 2006: 5(1):224-232.

Salwinski, L, et al. «The Database of Interacting Proteins: 2004 update.» *Nucleic Acids Research*, 2004: 32: D449-D451.

Savageau, M. *Biochemichal Systems Analysis: A Study of Function and Design in Molecular Biology.* Addison-Wesley, 1976.

Sawinski y Eisenberg. «Computational methods of analysis of protein-protein interactions.» *Current Opinion in Structural Biology*, 2003: 13: 377-382.

Seringhaus, M, et al. «Predicting essential genes in fungal genomes.» *Genome Research*, 2006: 16: 1126-1135.

Shannon, P, et al. «Cytoscape: a software environment for integrated models of biomolecular interaction networks.» *Genome Research*, 2003: 13(11): 2498-504.

Shoemaker, D, et al. «Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy.» *Nature Genetics*, 1996: 14: 450-456.

Stark, C, et al. «BioGRID: a general repository for interaction datasets.» *Nucleic Acids Research*, 2006: 34: 535-539.

The Gene Ontology Consortium. «The Gene Ontology (GO) database and informatics resource.» *Nucleic Acids Research*, 2004: 32:D258–D261.

Thibert, B, et al. «Improved prediction of critical residues for protein function based on network and phylogenetic analyses.» *BMC Bioinformatics*, 2005: 6:213.

Uetz, P, et al. «A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.» *Nature*, 2000: 623-631.

Viswanatan, G. «Getting started in biological pathway construction and analysis.» *PLoS Computational Biology*, 2008: 4 e16.

Von Bertalanffy, L. *General System Theory.* New York: Braziller, 1976.

Wiener, N. *Cybernetics: or the Control and Communication in the Animal and the Machine (2a edición).* The MIT Press, 1965.

Winzeler, E, et al. «Funtional caratcterization of the S. cerevisiae genome by gene deletion and parallel analysis.» *Science*, 1999: 285: 901-906.

Wunderlich y Mirny. «Using the topology of metabolic networks to predict viability of mutant strains.» *Biophysics Journal*, 2006: 91: 2304-2311.

Xenarios y Eisenberg. «Protein interaction databases.» *Current Opinion in Biotechnology*, 2001: 12: 334-339.

Zhang, S, et al. «Discovering functions and revealing mechanisms at molecular level from biological networks.» *Proteomics*, 2007: 7: 2856–2869.

Zotenko, E, et al. «Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential?: Reexamining the Connection between the Network Topology and Essentiality.» *PLoS Computational Biology*, 2008: 4(8):e1000140.