UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Instituto de Investigaciones Biomédicas

Instituto de Fisiología Celular

“Evaluación funcional de la red del metabolismo de *Saccharomyces cerevisiae*”

T E S I S

QUE PARA OBTENER EL GRADO DE:

LICENCIADA EN INVESTIGACIÓN B I O M E D I C A B Á S I C A

P R E S E N T A:

AURORA LABASTIDA MARTÍNEZ

CIUDAD UNIVERSITARIA Julio, 2009

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

“Evaluación funcional de la red del metabolismo de *Saccharomyces cerevisiae*”

TESIS QUE PARA OBTENER EL GRADO DE:

LICENCIADA EN INVESTIGACIÓN BIOMÉDICA BÁSICA

PRESENTA:

AURORA LABASTIDA MARTÍNEZ

ASESOR DE TESIS:

DR. GABRIEL DEL RÍO GUERRA

CIUDAD UNIVERSITARIA Julio, 2009

AGRADECIMIENTOS

A la Universidad Nacional Autonoma de Mexico a quien pertenezco con orgullo y respeto. A ella le debo mi formación en innumerables aspectos y la esperanza en el progreso de México.

A mis maestros, especialmente a mis jefes de laboratorio y a los miembros de mi jurado por su dedicación y por las variadas experiencias con las que impulsaron mi formación.

A mis abuelos Juana y José y mis papás por su apoyo constante e incondicional, por llevarme de la mano al preescolar, pasearme en el camellón, comprarme dulces, acompañarme por café, llevarme y traerme del instituto correspondiente a cualquier hora, viajar a Cuernavaca dos veces en un día, por las vacaciones, los cumpleaños, etc. Esta tesis la hicimos entre todos, poniendo nuestro máximo esfuerzo.

Gracias Juana por tu paciencia, por acompañarme en las comidas y por construir conmigo esa serie interminable conceptos hilarantes e interpretaciones alternativas del mundo.

Le agradezco mucho a mi tía Reina, quien además de vacacionar y comer conmigo ha confiado en mi y me ha ayudado a estudiar con ánimo y actitud científica. Gracias también a Eloisa, cuyo sentido del humor, apoyo y compañía me dan ánimos para trabajar.

A mis compañeros de generación: Alexa, Amanda, Raquel, Mariana, Osvaldo, Aydé, Pablo, Omar y David por estar cerca de mi en las clases, los estudios comunitarios, las comidas de fin de año, Oaxtepec, los pasillos del IFC, etc. He tenido una gran suerte al encontrarme con ustedes, su aprecio y compañía me impulsaron durante seis años.

Raquel, muchas gracias por tu amistad con lo mucho que ha implicado: Diversión, apoyo, paciencia, comida, consejos, estudio, aprecio, asesoría académica, etc. Estuviste cerca de mi en todas las dificultades y en muchas ocasiones me llevaste a la solución.

A mis compañeros de diferentes laboratorios: Tere, por las buenas pláticas y por ir conmigo a Canadá; Marco, Raúl y Jonathan por la compañía, las pláticas y el café; Bety, por ser tan cariñosa, sabia y divertida; Caty, por su amabilidad, buenos consejos y, en definitiva, por lo bien que nos la pasamos en el congreso de Michoacán; Lalo, Adan, Vale, Mauri, Tania, Chuchi, Hugo, Yolis … muchas gracias por el tiempo y experiencias compartidos; Ana E., por sus enseñanzas y buen humor.

Si alguna de las personas que contribuyeron en mi tesis no está en estos agradecimientos es por falla de mi memoria y no de mi gratitud.

ÍNDICE

RESUMEN

La biología de sistemas implica el uso de herramientas computacionales para generar modelos (redes) sobre las interacciones moleculares de un ser vivo, casi siempre unicelular. Una red tiene como fin explicar y predecir el comportamiento del organismo representado. Para validar ó corregir una red celular se sigue un proceso de evaluación funcional, donde el modelo se usa para predecir el fenotipo celular y las predicciones se comparan con datos experimentales.

La red metabólica de *Saccharomyces cerevisiae* es un modelo muy estudiado sobre el metabolismo eucarionte. Para hacer su evaluación funcional se ha predicho que genes del metabolismo son esenciales o dispensables, lo que ha permitido corregir y ampliar el modelo. Aún así, los errores en la predicción siguen siendo frecuentes. Esto sugiere que la red todavía es inexacta y/o que las técnicas que se han usado para extraer información a partir de ella no capturan toda la información que contiene.

Se ha observado la capacidad de algunas medidas topológicas para capturar información biológica en las redes celulares de diferentes organismos, incluyendo a *S. cerevisiae*. Con base en estos antecedentes, esta tesis presenta un estudio sobre la capacidad de 210 medidas topológicas para predecir la esencialidad de los genes en la red metabólica de *S. cerevisiae*. Las medidas topológicas se aplicaron en dos versiones de la red metabólica, ambas creadas en este mismo estudio. Ante la posibilidad de que las redes fueran inexactas, afectando la predicción, se realizó una inspección visual de dos de sus vías metabólicas (la glucólisis y el ciclo de Krebs).

Dos medidas topológicas (los inversos aditivos del grado saliente o k_{out} y el KatzR) generaron predicciones significativamente mejores que una predicción aleatoria, si bien, la eficiencia de estas predicciones es limitada. Este defecto puede ser consecuencia de la inexactitud de los modelos, evidenciada en su inspección visual, o de la incapacidad de los inversos aditivos del k_{out} y el KatzR para capturar los parámetros cinéticos del metabolismo. Para conocer el verdadero potencial de las medidas topológicas como herramientas de predicción será necesario aplicarlas en modelos más exactos, así como compararlas y combinarlas con técnicas de predicción basadas en parámetros cinéticos.

I. INTRODUCCIÓN

I.1 Biología de sistemas

Muchas de las funciones de los seres vivos no pueden explicarse estudiando el comportamiento de sus componentes aislados. Estas características dependen de una compleja red de interacciones, donde la función de cada biomolécula tiene efectos sobre la función de las demás. Bajo este enfoque los seres vivos son sistemas complejos cuyas características trascienden a las de sus componentes por separado.

Esta visión integrativa no es nueva en la biología molecular, un ejemplo de ello es el estudio de la regulación del operón *lac*, sin embargo, la aparición de la secuenciación automatizada de genomas provocó un aumento muy importante en la cantidad de datos que podían integrarse. Esto se acentuó con la aparición de técnicas para recabar datos sobre diversos componentes celulares en forma simultánea y automatizada. La información generada se almacenó progresivamente en bases de datos haciendo necesario el uso de la bioinformática para clasificar y hacer el análisis estadístico de los datos y posteriormente integrarlos en modelos que permitieran comprender las funciones celulares (1). En este marco surge la biología de sistemas, la cual es un campo de la biología que pretende descifrar cómo el comportamiento orquestado de las partes de un organismo da lugar a las propiedades que observamos en el sistema pero no en sus elementos aislados (propiedades emergentes)(2).

I.1.a Objeto y método de estudio

Un sistema es un conjunto de dos o más elementos que se relacionan entre sí. En él no pueden existir subgrupos independientes, es decir: si un elemento afecta al sistema pero no es afectado por él entonces este elemento está fuera del sistema. Así mismo, si un elemento es afectado por el sistema pero no afecta a ningún elemento del sistema está fuera de él (3). Así pues, un sistema biológico puede ir desde un grupo pequeño de proteínas necesarias para una función específica pasando por la célula hasta la biosfera entera (4). El límite depende del objetivo particular de quien estudia al sistema. Por varias razones, entre ellas la cantidad de datos disponibles y el poder de cómputo actual, la biología de sistemas se enfoca principalmente en la célula o alguno de sus sub-sistemas.

Cada estudio de biología de sistemas comienza con una serie de datos que deben integrarse para generar un modelo que explique el comportamiento celular, Para ello hay dos aproximaciones metodológicas (Figura 1) (5):

- Top-down (de arriba abajo): Comienza con la obtención de datos a gran escala (high-troughput) (6) que describen la presencia, abundancia en diferentes condiciones o interacciones físicas de los componentes celulares (metabolitos, DNA, RNA y/o proteínas). El análisis de estos datos permite definir que moléculas están relacionadas entre si e integrar

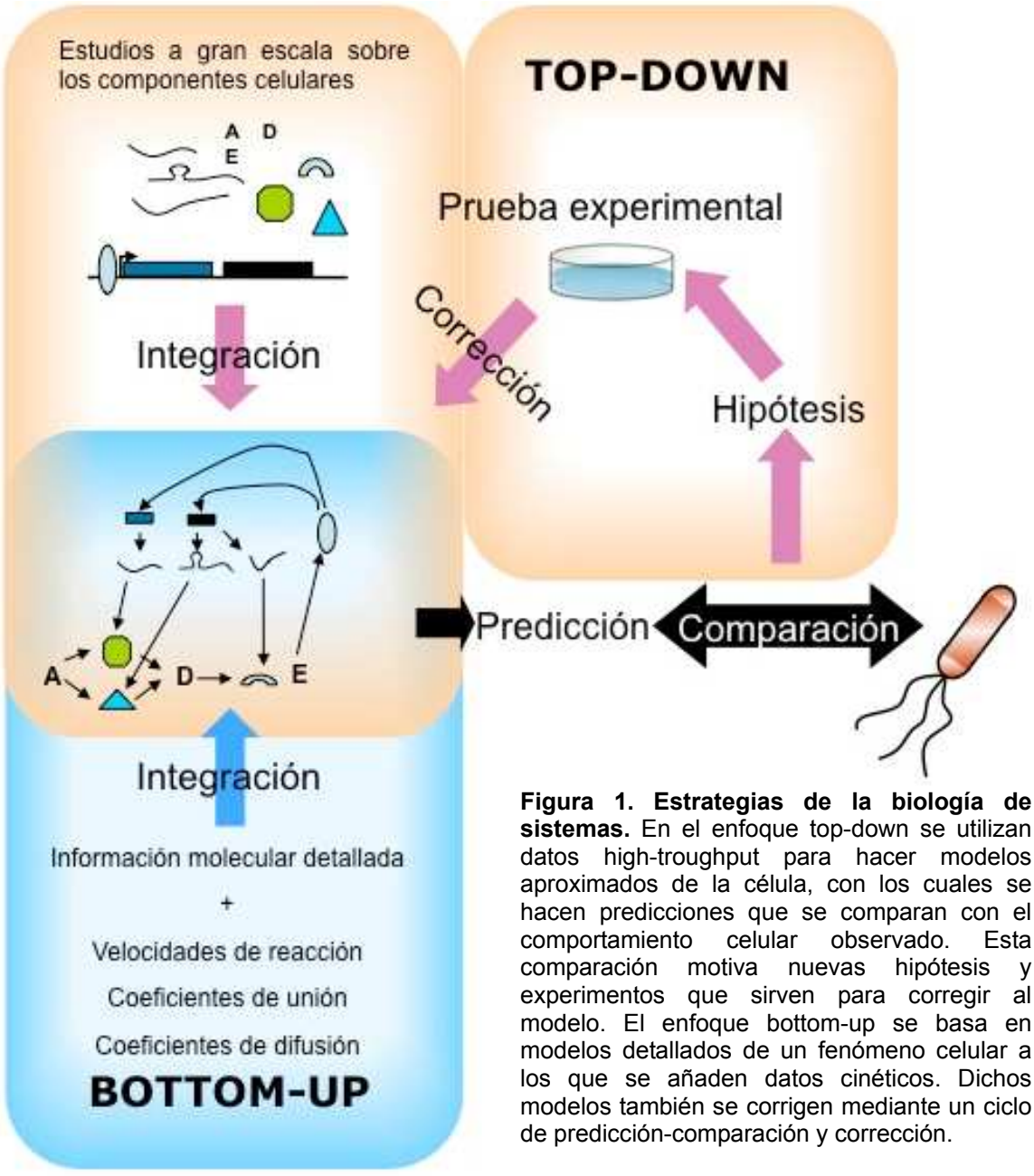


Figura 1. Estrategias de la biología de sistemas. En el enfoque top-down se utilizan datos high-troughput para hacer modelos aproximados de la célula, con los cuales se hacen predicciones que se comparan con el comportamiento celular observado. Esta comparación motiva nuevas hipótesis y experimentos que sirven para corregir al modelo. El enfoque bottom-up se basa en modelos detallados de un fenómeno celular a los que se añaden datos cinéticos. Dichos modelos también se corrigen mediante un ciclo de predicción-comparación y corrección.

una red[1] donde los nodos corresponden a los componentes moleculares unidos a través de uno o varios tipos de interacción. Esta red es sólo un andamio que debe corregirse para

[1] Una red o gráfico es un conjunto de objetos llamados nodos, que se unen a través de enlaces llamados aristas. Las aristas representan relaciones entre los nodos de la red (7).

mejorar su parecido con el sistema biológico real, para ello se lleva a cabo un proceso iterativo de predicción-comparación-corrección. El modelo se utiliza para hacer predicciones sobre el comportamiento celular que se comparan con datos experimentales. Las diferencias entre las predicciones del modelo y el fenotipo observado generan hipótesis que se comprueban o rechazan mediante nuevos experimentos. La aceptación de una hipótesis provoca cambios en los nodos o aristas del modelo. Con cada cambio del modelo se comienza de nuevo con la evaluación de sus predicciones.

- Bottom-up (de abajo a arriba): Se aplica a sistemas biológicos cuyo funcionamiento ha sido caracterizado a profundidad. Aquí, el objetivo es crear modelos detallados que permitan predecir el comportamiento dinámico del sistema, para lo cual se añaden datos cinéticos (velocidades de reacción, coeficientes de difusión, etc.) al modelo. Los datos cinéticos pueden obtenerse en forma experimental y/o estimarse mediante algoritmos basados en la estructura del modelo y en principios químicos o biológicos generales. Las predicciones obtenidas con los modelos del enfoque bottom-up también se comparan con datos experimentales para validar o mejorar el modelo.

I.1.b Utilidad de la biología de sistemas

La reconstrucción de una red celular nos permite abordar la forma en que las interacciones entre componentes dan origen a las propiedades emergentes. Gracias a las herramientas de la teoría de gráficos (7) puede estudiarse la topología de redes, que es el arreglo de los

nodos y aristas de las redes biológicas. A través de la topología se han descrito características como el nivel de tolerancia a errores, el sistema evolutivo de construcción y la tendencia de las redes biológicas a organizarse en grupos funcionales o módulos (8). Por otro lado durante la reconstrucción de una red celular salen a relucir las sub-maquinarias que participan en una función dada y la forma en que la llevan a cabo (9-11).

Se pretende que parte del conocimiento generado por la biología de sistemas tenga un impacto en la sociedad en general. La industria farmacéutica, por ejemplo, puede beneficiarse de modelos que le permitan identificar blancos terapéuticos óptimos y de simulaciones que predigan los efectos causados por una droga sobre la maquinaria celular. Pueden predecirse modificaciones que conviertan a un organismo en una máquina de síntesis ó procesamiento útil para la industria química ó de alimentos. Finalmente, la comprensión de las redes metabólicas microbianas reforzará a la toxicología y ecología de microorganismos, prediciendo métodos y riesgos para la introducción de nuevos microorganismos en el ambiente y nuevas formas de biorremediación (12).

I.2 Biología de sistemas de *Saccharomyces cerevisiae*

I.2.a La levadura como modelo para la biología de sistemas

El uso de organismos modelo en biología se basa en la conservación de los mecanismos bioquímicos durante el proceso evolutivo, lo que nos permite extrapolar lo observado en una

especie a otra. *S. cerevisiae* es el modelo eucarionte más utilizado para describir al sistema celular y generar herramientas computacionales para su análisis. Hay varias razones por las que la levadura ocupa este lugar:

- Para empezar, los procesos celulares fundamentales son muy conservados entre los eucariontes desde la levadura hasta el humano y de hecho varios de ellos se describieron inicialmente en *S. cerevisiae*. El genoma de este microorganismo tiene aproximadamente 6200 genes de los cuales más de 2000 tienen ortólogos en el humano (13).

- En el ámbito experimental hay ventajas importantes. La levadura puede cultivarse en condiciones muy controladas, favoreciendo la reproducibilidad de los experimentos. Además, su manipulación genética es sencilla, lo que permitió la deleción sistemática de sus más de 6000 genes (14) y generar colecciones de fusiones cromosomales útiles para estudios de expresión (15) y localización de proteínas (16) o para la purificación de complejos proteicos (17).

- Hay una gran cantidad de información sobre las biomoléculas de *S. cerevisiae* y sus interacciones. El genoma eucarionte más estudiado es el de este organismo y la mayoría de las técnicas high-troughput para la obtención de transcriptomas, proteomas, interactomas y metabolomas, así como las herramientas bioinformáticas para el análisis de estos datos comenzaron en la levadura (12).

- La información molecular obtenida en los estudios de *S. cerevisiae* se encuentra en

bases públicas de datos (18).

- Finalmente, la comunidad científica que estudia a *S. cerevisiae* a nivel mundial es muy grande y organizada, lo que facilita el intercambio de materiales y datos (12).

I.2.b Avances en el área

La mayoría de los estudios de biología de sistemas en *S. cerevisiae* siguen la metodología top-down para hacer reconstrucciones a nivel celular. Estos modelos no alcanzan a resolver todos los detalles cualitativos sobre las interacciones que ocurren en el sistema y menos aún su dinámica, pero son de gran ayuda para abordar estructuras y comportamientos globales. Como ejemplos de estas aproximaciones:

- Hay estudios enfocados a describir una vía o comportamiento en particular. En uno de ellos se construyó una red celular para explicar el crecimiento filamentoso de la levadura, donde se identificaron grupos funcionales que explican las propiedades celulares observadas en este proceso (9).

- A partir de datos del proteoma y del interactoma pueden construirse redes de interacción proteína-proteína donde los nodos son las proteínas de la levadura y las aristas corresponden a sus interacciones físicas (que reflejan relaciones funcionales). Estos modelos se han utilizado para explicar la forma en que las proteínas se agrupan en

grandes complejos funcionales y para poner en contexto proteínas de función desconocida (10, 11, 17).

- Usando datos de la unión de factores transcripcionales al DNA y de cambios de expresión transcripcional se construyen redes de regulación transcripcional. Al estudiar la estructura de estas redes se han identificado formas globales de regulación génica que conectan a diversas funciones celulares y participan en la adaptación a diferentes ambientes (20, 21).

- Las redes de interacción genética se construyen con datos de interacciones epistáticas entre genes, que ayudan a identificar locus que participan en la misma vía o proceso celular. Actualmente estas interacciones pueden evaluarse en forma automatizada y aportan información sobre la función de genes individuales, la relación entre rutas metabólicas y la forma en que la célula resiste errores genéticos y ataques ambientales, entre otros aspectos (22).

- Las redes metabólicas representan los procesos bioquímicos de un organismo y son una herramienta poderosa para investigar la organización química y genética de la célula. Los modelos del metabolismo de *S. cerevisiae* se han utilizado con varios fines, entre ellos, estudiar la evolución (23), la regulación de la expresión (24, 25), la redundancia genética (26) y la aparente existencia de genes dispensables (27, 28) en el metabolismo. Más aún, la capacidad de estos modelos para predecir el impacto de mutaciones individuales en la viabilidad de la levadura se ha estudiado extensamente (ver I.3.b).

También se han hecho estudios bottom-up en *S. cerevisiae*, entre ellos están los estudios sobre la glucólisis, el ciclo celular, la vía de respuesta a feromonas y la osmoregulación (12). Para que estos estudios se realicen a nivel celular aún se requiere mucha información cuantitativa.

La descripción total de *S. cerevisiae* como sistema biológico requerirá de toda la información estructural y dinámica generada por diferentes estudios. Las redes generadas por métodos top-down deberán integrarse en un solo andamio para describir el panorama completo de las interacciones celulares. Además, las descripciones cinéticas detalladas del enfoque bottom-up permitirán que sub-maquinarias de la célula puedan integrarse en el andamio celular. Con esto se obtendrá una visión espacial y temporal de las funciones celulares.

Como ejemplo sobre la integración de los enfoques top-down y bottom-up (19): la utilización de galactosa es un proceso bien estudiado en *S. cerevisiae*. En base a los estudios sobre esta vía se construyó un modelo que representa la transformación de galactosa en glucosa-6-fosfato y la regulación transcripcional del proceso (un modelo bottom-up). Luego, se midió experimentalmente el efecto de mutar cada uno de los genes incluidos en el modelo sobre la expresión transcripcional de 6200 genes de *S. cerevisiae*. Con estos datos se definieron nuevas relaciones entre los genes de la utilización de galactosa y de estos genes con los de otros procesos celulares. Como consecuencia, se refinó el modelo de utilización de galactosa y se ubicó dentro de la red de interacción proteína-proteína y proteína-DNA de *S. cerevisiae* (modelos top-down).

I.3 Evaluación funcional de la red metabólica de *S. cerevisiae*

Como ya se mencionó, tras modelar un sistema biológico es necesario validarlo o corregirlo usando datos experimentales. Para ello se utilizan técnicas que hacen predicciones a partir del modelo, las cuales se comparan con el comportamiento biológico observado. A este procedimiento nos referimos como "evaluación funcional de la red". Esta sección trata de las redes metabólicas de la levadura y los avances en su evaluación funcional.

I.3.a Redes metabólicas de S. cerevisiae

Los primeros esfuerzos por modelar el metabolismo de *S. cerevisiae* culminaron en la construcción de redes bioquímicas, donde las reacciones enzimáticas se representan como conexiones entre metabolitos. Se realizaron varios estudios sobre la topología de los modelos bioquímicos (29-33), pero no se hizo énfasis en comprobar su fidelidad. Más tarde, con la construcción de redes genéticas se inició formalmente el proceso de evaluación funcional.

Las redes genéticas se basan en el modelo propuesto por la base de datos KEGG (34), pero incluyen información de otras bases de datos y correcciones derivadas de la literatura bioquímica. Se trata de redes bipartitas[2], en las que, tanto los genes que codifican enzimas,

[2] En una red bipartita los nodos se dividen en dos conjuntos y no existen aristas entre nodos del mismo conjunto (35). En este caso, cada arista une a un gen con un metabolito, representando la producción del metabolito y/o su consumo en la reacción enzimática corespondiente.

como los metabolitos que participan en cada reacción se representan como nodos. En ellas, las reacciones enzimáticas están organizadas por rutas metabólicas y por su localización en diferentes compartimentos celulares y se toma en cuenta la existencia de reacciones irreversibles, enzimas multiméricas y complejos proteicos, así como mecanismos de transporte molecular entre organelos. El modelo iFF708 (36) fue la primera red genética de *S. cerevisiae*. Sus modificaciones o mejoras han dado origen a otros modelos, incluyendo a iND750 (37) e iLL672 (26).

I.3.b Enfoques de la evaluación funcional

En general, la evaluación funcional de una red metabólica consiste en determinar que tan útil es el modelo para diferenciar a los genes esenciales para la viabilidad de los no esenciales. Por ejemplo, se usa a la red para predecir qué mutaciones afectarán el crecimiento de la levadura y se compara la predicción con datos experimentales.

Las técnicas de predicción más comunes se conocen como análisis basados en constricciones (CBA por sus siglas en inglés). En ellos se calcula el potencial de la célula para producir biomasa, para lo cual se estiman las velocidades de reacción (flujos) de la red metabólica. Para evaluar si un gen es esencial el cálculo se realiza con la red completa y con la red en ausencia del producto del gen, (una enzima o parte de una enzima). Si la eliminación de dicho producto disminuye la capacidad para producir biomasa, entonces se predice que el gen es esencial. El análisis de balance de flujos (FBA) (38) y la minimización

del ajuste metabólico (MOMA) (39) son ejemplos de este enfoque. Otras técnicas combinan parámetros químicos y topológicos en la predicción. Como ejemplo, la accesibilidad sintética mide la cantidad de pasos enzimáticos que se necesitan para producir un grupo de metabolitos a partir de sus precursores. Se predice que un gen es esencial cuando su ausencia aumenta la cantidad de pasos a realizar (40).

I.3.c Eficiencia de las técnicas de predicción

Ante la escasez de datos cinéticos, la alternativa para abordar a la célula como maquinaria química es usar herramientas como el CBA, que estima los parámetros desconocidos y los usa en la predicción de la viabilidad. En muchos casos los parámetros cinéticos calculados por el CBA son correctos, no obstante, los errores en la predicción de la esencialidad siguen siendo frecuentes (36, 41). Esto se debe en parte, a que en el CBA es necesario describir qué metabolitos están presentes en el medio de cultivo y especificar cuales son necesarios para la formación de biomasa, parámetros que son poco conocidos. Por otro lado, el CBA y la accesibilidad sintética enfocan a las enzimas únicamente como consumidoras o productoras de metabolitos, pero ignoran varios fenómenos que podrían determinar la importancia de algunas enzimas, por ejemplo, sus interacciones físicas con otras proteínas o a la forma en la que se regula su expresión. Es así que se han observado fallas en las predicciones del CBA originadas por su incapacidad para capturar fenómenos como la regulación transcripcional (41).

I.3.d Las medidas de centralidad como alternativa en la evaluación funcional.

Las medidas de centralidad son un grupo de algoritmos topológicos que miden la importancia relativa de un nodo para una red (su centralidad) (59). Para ello capturan la relación o posición de un nodo con respecto a otros nodos y aristas. Estas medidas pueden usarse en todo tipo de redes, sin importar la naturaleza de sus componentes (genes, metabolitos, individuos, etc.). Es así que, a diferencia del CBA y la accesibilidad sintética que sólo miden relaciones bioquímicas, las medidas de centralidad se han utilizado en varios tipos de redes celulares, donde se ha observado su capacidad para capturar diferentes tipos de información sobre los genes o proteínas. Como ejemplos:

- El grado, que es el número de aristas en las que participa un nodo *n* y la intermediación, que mide cuantas veces se pasa por *n* al ir de un nodo a otro, son medidas de centralidad comunes (ver Figura 3b y f en la sección V.3). En las redes de interacción proteína-proteína de *S. cerevisiae, Drosophila melanogaster y Caenorhabditis elegans* se ha observado una correlación positiva entre el grado o intermediación de las proteínas y su esencialidad o su conservación evolutiva (42).

- En redes de coexpresión genética de *S. cerevisiae* (donde los nodos son genes que se unen si están coexpresados) también se ha observado una correlación positiva entre el grado de un gen y su esencialidad o conservación evolutiva (43).

- Los capacitores fenotípicos, genes que estabilizan el fenotipo ante la variación genética o ambiental, pueden identificarse por su alto grado en la red de interacciones genéticas de *S. cerevisiae* (44).

- En la red metabólica iFF708 de *S. cerevisiae* se ha observado una correlación positiva entre el grado y la conservación evolutiva de las enzimas (23).

Se han propuesto varias explicaciones para la relación entre la centralidad y la esencialidad o conservación evolutiva de un gen o proteína. Por ejemplo, las proteínas o genes con alto grado en la redes celulares podrían conservarse por que influyen en varias funciones celulares, lo que impone una constricción sobre su evolución (42). También se ha propuesto que las proteínas con alta intermediación funcionan como puentes entre diferentes secciones de las redes celulares, por lo que son necesarias para la propagación de información en la célula (45, 46).

La capacidad de las medidas de centralidad para capturar información biológica abre la posibilidad de usarlas en la evaluación funcional de diferentes tipos de redes celulares. Para ello será necesario cuantificar y comparar la capacidad de diferentes medidas de centralidad para predecir la esencialidad y otras características biológicas. Cabe mencionar que existen muchas medidas de centralidad cuya capacidad para capturar información biológica no se ha estudiado y que podrían ser útiles como técnicas de predicción en redes celulares.

II. JUSTIFICACIÓN

La capacidad de una red para predecir el fenotipo celular sugiere que dicho modelo representa correctamente a la célula y que sus predicciones pueden facilitar tareas, cuya realización a nivel experimental es difícil o costosa (Ej. La identificación de blancos farmacéuticos (47)). No obstante, el éxito en la predicción depende tanto del modelo como de la técnica que extrae información a partir de él. Un error puede deberse a la inexactitud del modelo, o bien, a la incapacidad de la técnica de predicción para capturar cierto tipo de información. Así pues, al mismo tiempo que se enriquecen las redes biológicas, es necesario mejorar las técnicas de predicción existentes para que midan todos los fenómenos representados en ellas. La capacidad de las medidas de centralidad para capturar información biológica en diferentes tipos de redes celulares de varios organismos sugiere que alguna de estas medidas podría predecir la esencialidad de los genes de la red metabólica de *S. cerevisiae*, contribuyendo en su evaluación funcional.

III. HIPÓTESIS

Si una red representa correctamente al metabolismo de *S. cerevisiae* entonces los genes esenciales ocuparán un lugar importante en su estructura y podrán identificarse usando medidas topológicas.

IV. OBJETIVOS

IV.1 Objetivo general

Evaluar la capacidad de varias medidas topológicas en la identificación de los genes esenciales para la supervivencia en la red metabólica de *S. cerevisiae.*

IV.2 Objetivos particulares

a. Reconstruir a la red metabólica de *S. cerevisiae*.

b. Analizar la calidad del modelo obtenido.

c. Probar la capacidad de:

- 10 medidas de centralidad.
- Los inversos aditivos de las 10 medidas de centralidad.
- Combinaciones entre las 20 medidas anteriores

para predecir la esencialidad de los genes en la red metabólica de *S. cerevisiae.*

V. MÉTODOS

Para la realización de este trabajo se utilizaron programas de Java, comandos o scripts de la terminal de Linux o programas de AWK. Además, como se menciona más adelante se usaron datos y programas provistos por otros programadores o disponibles en Internet.

V.1 Construcción de las redes R-iND750 y RSH-iND750

Como parte de este trabajo se construyeron dos redes metabólicas de la levadura (R-iND750 y RSH-iND750) basadas en el modelo iND750 (37), cuyos datos se obtuvieron del sitio web del Systems Biology Research Group, de la Universidad de California en San Diego (http://geneticcircuits.ucsd.edu/).

El modelo iND750 lista a las enzimas de la levadura, indicando que genes las codifican, las reacciones metabólicas en las que participan y la vías metabólicas y compartimientos celulares a los que pertenecen. El modelo también describe los mecanismos de transporte entre compartimientos e incluye 339 reacciones (enzimáticas o de transporte) para las cuales no se conocen los genes y enzimas responsables (reacciones hipotéticas), pero que son potencialmente necesarias para garantizar el crecimiento de la levadura.

Para construir a R-iND750 y RSH-iND750 a partir de iND750 se siguió un proceso como el

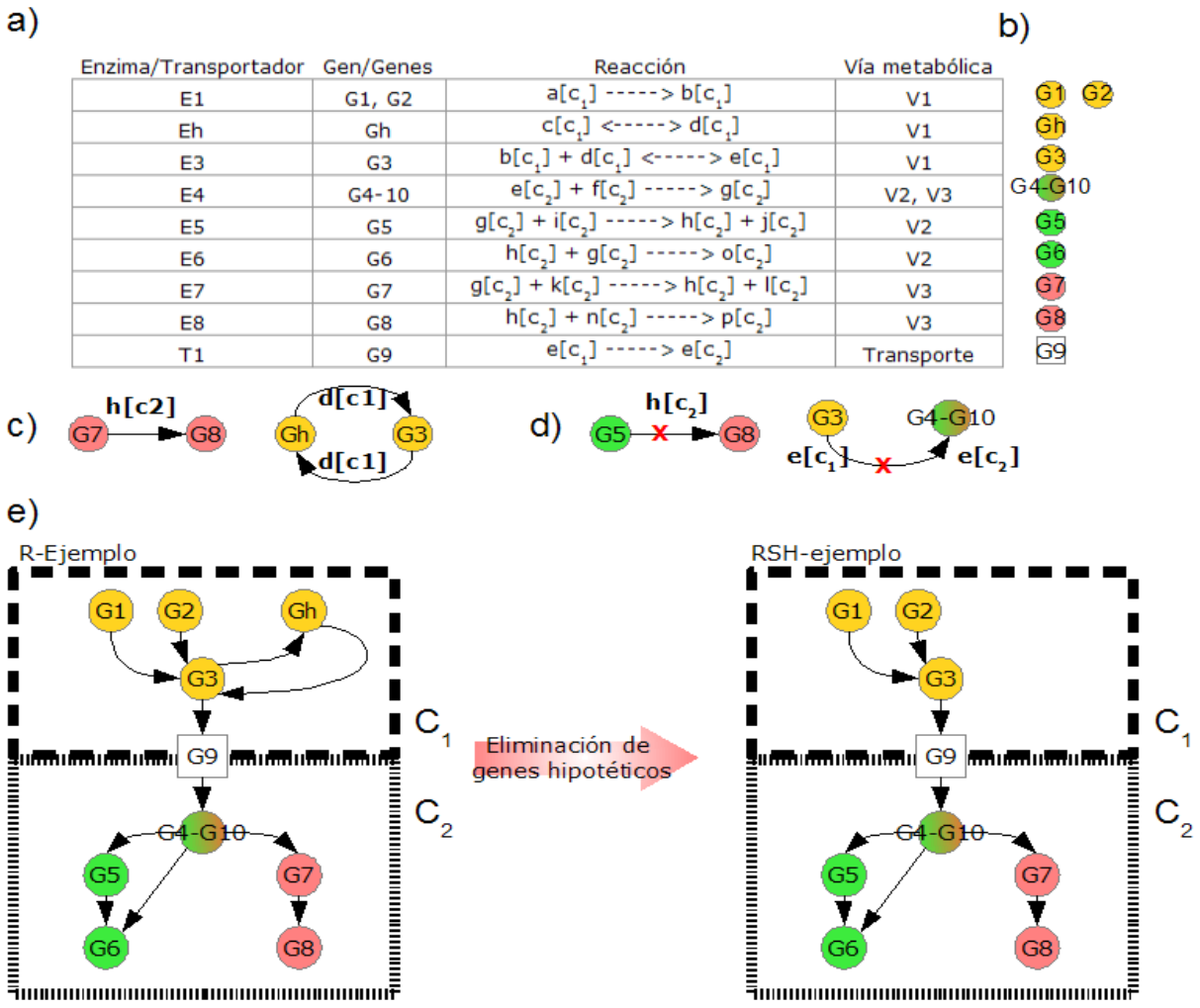


Figura 2. Construcción de las redes metabólicas. Los nodos que pertenecen a la misma vía metabólica tienen el mismo color. El nodo G4-G10 tiene dos colores, correspondientes a las vías V2 y V3. e) Las líneas punteadas delimitan a los compartimientos c_1 y c_2. Ver explicación en el texto

que se ilustra en la Figura 2. La Figura 2a supone un conjunto de siete enzimas (E1 y E3-E8) y un transportador (T1). Para representar a las reacciones hipotéticas se agregó una reacción catalizada por la enzima Eh (de la misma forma a cada reacción hipotética de iND750 se le asoció una enzima y un gen hipotéticos). Al igual que en iND750 las enzimas de la Figura 2a pertenecen a una o varias vías metabólicas y los transportadores están asociados al la función “transporte”. Cada enzima o transportador participa en una reacción

con una localización celular determinada, por ejemplo

$$c[c_1] <---> d[c_1]$$

implica que la el metabolito **c** del compartimiento $\mathbf{c_1}$ se convierte en el metabolito **d** del compartimiento $\mathbf{c_1}$ en forma reversible. En cambio

$$e[c_1] <---> e[c_2]$$

se refiere al transporte reversible del metabolito **e** del compartimiento $\mathbf{c_1}$ al compartimiento $\mathbf{c_2}$**.** Aquí, el metabolito **e** se interpreta como sustrato y producto de la reacción.

Cada enzima o transportador está relacionado con un gen o con varios cuando existen isoenzimas (Ej. E1) o se trata de un hetero-multímero o complejo proteico (Ej. E4). Para construir a las redes metabólicas se eligieron como nodos a los genes que codifican a las enzimas (Figura 2b). Se creó un nodo por cada gen excepto en el caso de los genes que codifican a un hetero-multímero o complejo proteico, donde todos los genes implicados se agruparon en un sólo nodo (Ej. el nodo formado por G4 y G10).

Para establecer las conexiones (aristas) entre los nodos de la red se siguió el siguiente criterio: Se establece una arista que va del nodo *X* al *Y* si la enzima o transportador codificado por *X* produce un metabolito *m* y la enzima o transportador codificada por *Y* toma a *m* como sustrato. La aplicación de este criterio tiene tres condiciones:

1. *X* debe producir a *m* en el mismo compartimiento celular donde *Y* lo toma como sustrato.

2. El metabolito *m* no puede ser H_20, un protón, fosfato inorgánico, ATP, ADP ó L-grutamato (Los sustratos y productos más abundantes en iND750). Esto se hizo para evitar la aparición de conexiones sin aparente significado funcional (30).

3. Las enzimas codificadas por *X* y *Y* deben coincidir en al menos una ruta metabólica. En iND750 las enzimas están relacionadas con 49 vías metabólicas de importancia funcional conocida (Ej. glucólisis y síntesis de ácidos grasos) (37).

Como ejemplos sobre el establecimiento de aristas (Figura 2c): El metabolito h es producto de E7 y sustrato de E8, por lo que se establece una arista de G7 (que codifica a E7) a G8 (que codifica a E8). Por otro lado, Eh y E3 llevan a cabo reacciones reversibles, lo que les permite tomar a d como sustrato y también producirlo. Así pues, se crea una conexión de Gh a G3 y otra de G3 a Gh.

En cuanto a las condiciones para la creación de aristas (Figura 2d): El metabolito h es producto de E5 y sustrato de E8, pero E5 y E8 están relacionados con diferentes vías metabólicas, lo que impidió la conexión de G5 a G8. Por otro lado, el metabolito e es producto de E3 y sustrato de E4, pero E3 y E4 no coinciden en alguna vía metabólica y mientras que la primera produce a e en el compartimientos celular c_1, la segunda consume a e en c_2. Es así que se descartó la conexión de G3 a G4-G10.

La Figura 2e muestra la red R-ejemplo obtenida mediante los principios enumerados anteriormente. En ella el gen G9, que codifica al transportador T1 comunica a los nodos asociados con el compartimiento c1 con los nodos asociados con el compartimiento c2 y el nodo G4-G10, que está asociado con las vías V2 y V3 sirve como puente entre los nodos de ambas vías. La red RSH-ejemplo (red sin hipotéticos ejemplo) de la Figura 2e se obtuvo de la misma forma, pero excluyendo al gen hipotético Gh. Se siguió el mismo procedimiento con los datos de iND750 para obtener la red R-iND750 y la red sin hipotéticos iND750 (RSH-iND750).

V.2 Valoración de la calidad de las redes

Las interacciones de la glucólisis y ciclo de Krebs de R-iND750 se modelaron con el software Cytoscape (http://www.cytoscape.org/) (48).

V.3 Medidas de centralidad

Los valores de centralidad para los nodos de R-iND750 y RSH-iND750 fueron provistos por Dirk Koschützki[1] (grado entrante, grado saliente, intermediación, Katz, KatzR, PageRank y PageRankR) (8, 49, 50. 51) ó calculados con programas codificados por Boris Thibert[2]

[1] Research Fellow, Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research, (IPK), Gatersleben, Alemania.

[2] Assistant Professor in mathematics, Laboratoire Jean Kuntzmann, Joseph Fourier University, Grenoble, Francia.

(exentricidad, distancia promedio y coeficiente de empacamiento) (8, 52, 53). La Figura 3 introduce los conceptos de la ruta más corta y la distancia geodésica (52), relacionados con varias de las medidas de centralidad y explica el cálculo de cada una de estas medidas.

V.4 Inversos aditivos de las medidas de centralidad

El valor de un nodo según el inverso aditivo de una medida de centralidad se define como

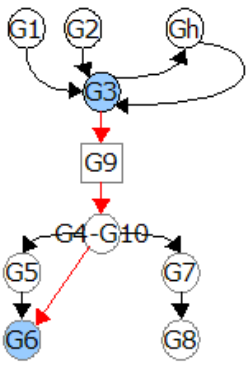
$$C_{ia}(n) = -C(n)$$

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. Se calcularon los inversos aditivos para los valores obtenidos con cada medida de centralidad en R-iND750 y RSH-iND750.

Los nodos de una red pueden ordenarse de los más centrales a los menos centrales según una medida de centralidad (ver V.7). Si los nodos se ordenan con respecto al inverso aditivo de la misma medida de centralidad el orden de los nodos se invierte (Figura 4). Así pues, una centralidad y su inverso aditivo hacen estimados opuestos sobre la importancia relativa de los nodos.

Figura 3. Medidas de centralidad evaluadas en este trabajo (Páginas 26-29). a) Conceptos generales relacionados con el cálculo de las medidas de centralidad, b) Grado entrante (k_{in}) y grado saliente (k_{out}), c) Coeficiente de empacamiento (CE), d) Excentricidad (E), e) Distancia Promedio (DP), f) Intermediación (I), g) Katz, h) KatzR, i) PageRank (PR), j) PageRank R (PR_R).

a) Conceptos generales



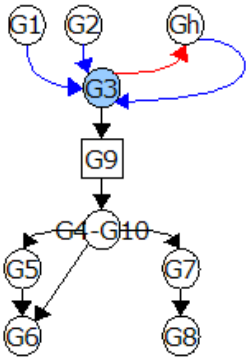
Ruta más corta
La ruta más corta entre un nodo *n* y un nodo *m* es la sucesión de aristas más pequeña que va de *n* a *m*.

Distancia geodésica
La distancia geodésica entre *n* y *m* o dist(*n*,*m*) es el número de aristas que hay en ruta más corta correspondiente.

ej.

dist(G3,G6)=3

b) Grado entrante k_{in} y grado saliente k_{out}



El $k_{in}(n)$ y el $k_{out}(n)$ son el número de aristas que llegan a un nodo *n* y que salen de un nodo *n* respectivamente.

ej.

$k_{in}(G3)=3$
$k_{out}(G3)=2$

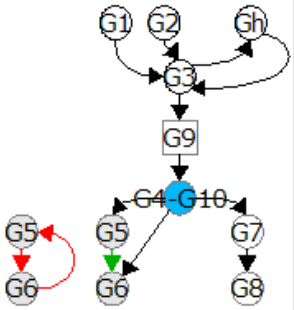
Nota: Si no se define el grado como "entrante" o "saliente" entonces se trata del número de aristas en las que participa *n*, sin importar su dirección.

c) Coeficiente de empacamiento (CE)

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

Número de conexiones entre los vecinos del nodo *n* (los nodos que reciben conexiones de *n*)

Máximo número de conexiones posible entre los vecinos de *n*

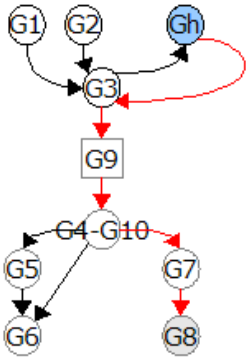


ej. G4-G10 tiene como vecinos a G5 y G6. Hay dos conexiones posibles entre G5 y G6. De ellas sólo existe una en la gráfica. Entonces:

CE(G4-G10)= 1/2 = 0.5

Nota: Para los nodos con 0 o 1 conexiónes CE=0

d) Excentricidad (E)



La E(*n*) es la distancia geodésica entre *n* y el nodo más alejado de *n*.

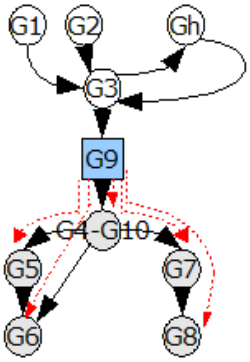
ej.

El nodo más alejado de Gh es G8, por lo tanto:

E(Gh)=dist(Gh, G8)=5

Nota: Si no existe una ruta desde *n* y hacia un nodo *m*, el tamaño de la ruta entre n y m es cero.

e) Distancia Promedio (DP)



Promedio de las distancias geodésicas entre *n* y todos los nodos a los que se pueda llegar desde *n* (conjunto A)

$$DP(n)=\frac{\sum_{m\in A} dist(n,m)}{a}$$

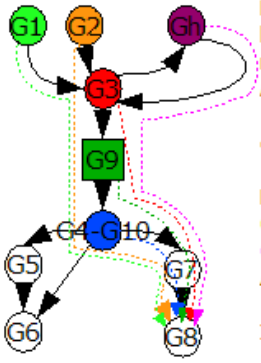
Cantidad de nodos en el conjunto A

ej. Partiendo de G9 se puede llegar a G4-G10, G5, G6, G7 y G8, por lo tanto:

DP(G9)= [1+2+2+2+3]/3.

Nota: Si $k_{out}(n)=0$ se asume que DP(*n*)=0.

f) Intermediación (I)



La I(*n*) es la cantidad de las rutas más cortas de la red que atraviezan el nodo *n*.

ej.

Las seis rutas que van de G1, Gh, G3, G2, G9 y G4-G10 a G8 pasan por G7. Así pues:

I(G7)= 6

g) Katz

Constante fraccionaria

$$C_{katz} = \sum_{a=0}^{\infty} \alpha^a (A^T)^a \mathbf{1}$$

Matriz de adyacencias de la red R, donde cada elemento a_{ij} es el número de conexiones que van del nodo i al nodo j.

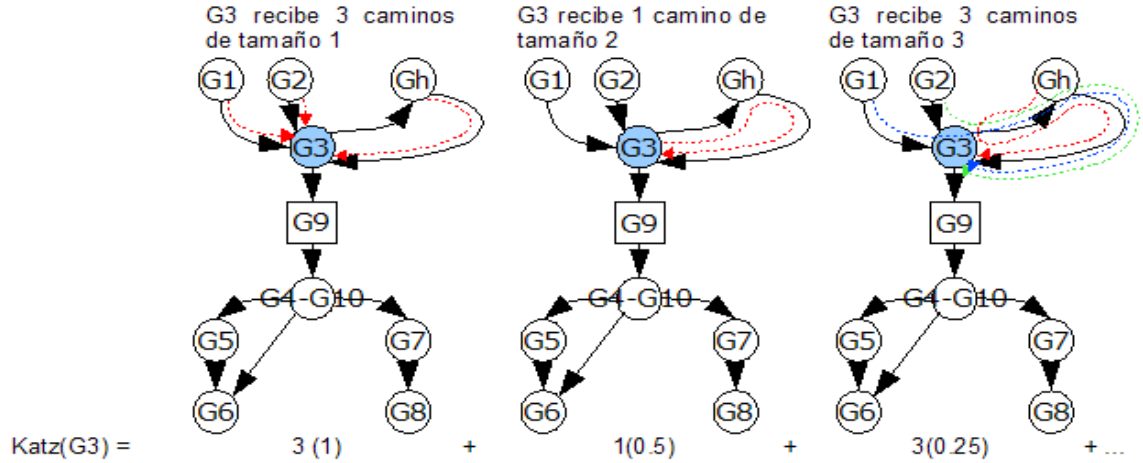
Vector con los valores de Katz para los nodos de R

Vector cuyos k elementos son iguales a 1 (k es el número de nodos de R)

El Katz(*n*) depende de los caminos que llegan al nodo *n* desde otros nodos. Dichos caminos se buscan de los más cortos a los más largos. Se toma en cuenta cualquier camino posible, incluso si pasa por las mismas aristas repetidamente o si sale de *n* y regresa a *n*. Los caminos encontrados reciben un valor que es menor entre más largos son (La disminución del valor con respecto al tamaño depende de α). El Katz(*n*) es la suma de los valores de los caminos encontrados.

ej. Si $\alpha=0.5$

Tamaño del camino	1	2	3
Valor del camino	$\alpha^0=1$	$\alpha^1=0.5$	$\alpha^2=0.25$



Katz(G3) = 3 (1) + 1(0.5) + 3(0.25) + ...

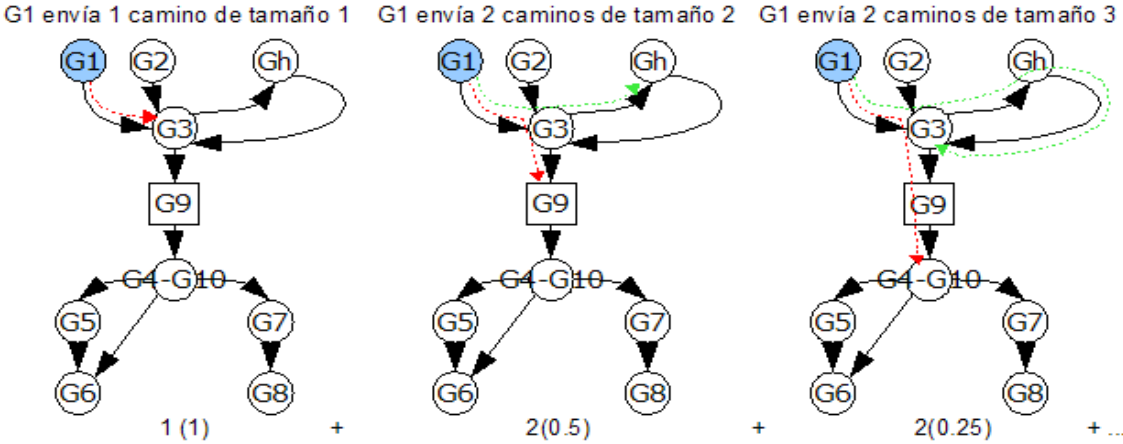
---

h) KatzR

$$C_{katzR} = \sum_{a=0}^{\infty} \alpha^a (A)^a \mathbf{1}$$

El cálculo del KatzR(*n*) es similar al cálculo del Katz(*n*), pero se basa en los caminos que salen de *n hacia* otros nodos.

ej.
$\alpha=0.5$ y el valor de caminos de tamaño 1, 2 y 3 es 1, 0.5 y 0,25, respectivamente.



Katz(G1) = 1 (1) + 2(0.5) + 2(0.25) + ...

i) PageRank (PR)

$$PR(n) = (1-d) + d \sum_{m \in N_n} \frac{PR(m)}{k_{out}(m)}$$

Constante fraccionaria

EL PR(n) mide el valor de las aristas que recibe el nodo n y para calcularlo se toman en cuenta a todos los nodos m que envían aristas a n. Cada arista que va de m hacia n se interpreta como un voto que m otorga a n. El voto de m tiene mayor valor entre más grande es el PR de m y su valor disminuye al aumentar el número de votos emitidos por m.

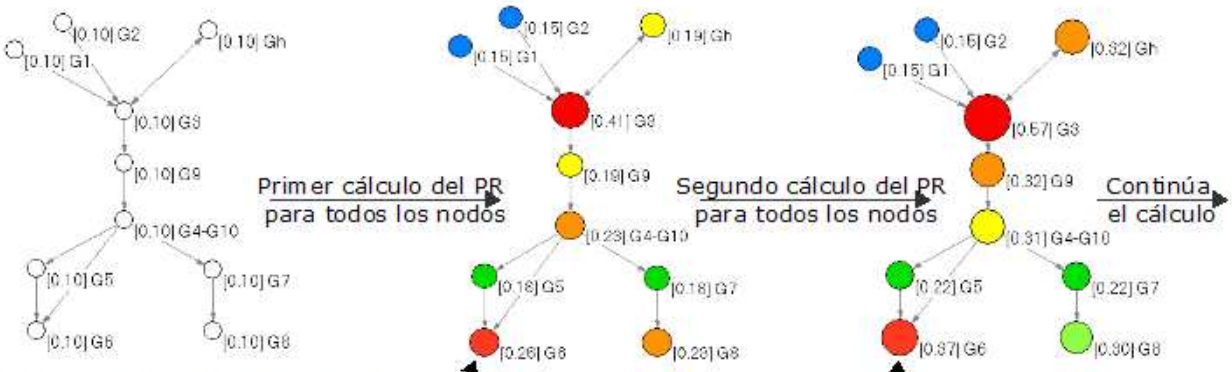
Para calcular el PR de los k nodos de una gráfica se comienza asumiendo que los k nodos tienen el mismo PR (PR(n)=1/k). Los valores de PR se re-calculan varias veces hasta que se vuelven estables.

ej. El nodo G6 recibe aristas de los nodos G5 y G4-G10. Si d = 0.85.

$$PR(G6) = 0.15 + 0.85 \left(\frac{PR(G4\text{-}G10)}{3} + \frac{PR(G5)}{1} \right)$$

d; $k_{out}(G4\text{-}G20)$; $k_{out}(G5)$

En la gráfica hay 10 nodos. El PR inicial para todos los nodos es 1/10



$$PR(G6) = 0.15 + 0.85 \left(\frac{0.1}{3} + \frac{0.1}{1} \right)$$

$$PR(G6) = 0.26$$

$$PR(G6) = 0.15 + 0.85 \left(\frac{0.23}{3} + \frac{0.18}{1} \right)$$

$$PR(G6) = 0.37$$

j) PageRank R (PR_R)

$$PR_R(n)=(1-d)+d\sum_{m\in N,n}\frac{PR_R(m)}{k_{in}(m)}$$

Conjunto de nodos que reciben aristas de *n*

El $PR_R(n)$ mide el valor de las aristas (votos) que salen del nodo *n* hacia otros nodos *m*. Un voto de *n* vale más entre más alto es el PR_R del nodo electo (*m*) y su valor disminuye al aumentar el número de nodos que eligieron a m.

El cálculo del PR_R para los k nodos de una gráfica sigue la misma mecánica que el cálculo del PR. Al principio $PR_R(n)=1/k$ para todos los nodos y los valores de PR_R se re-calculan varias veces.

ej. El nodo G4-G10 envia aristas a los nodos G5, G6 y G7. Si d = 0.85.

$$PR_R(G4\text{-}G10) = 0.15 + 0.85\left(\frac{PR_R(G5)}{1} + \frac{PR_R(G6)}{2} + \frac{PR_R(G7)}{1}\right)$$

d; $k_{in}(G5)$; $k_{in}(G6)$; $k_{in}(G7)$

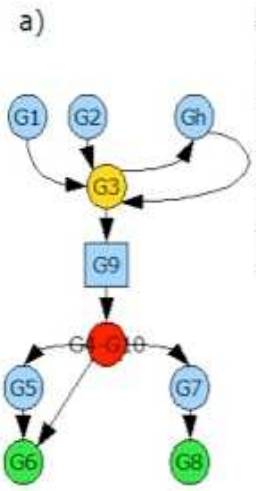
En la gráfica hay 10 nodos. El PR_R inicial para todos los nodos es 1/10

$$PR_R(G4\text{-}G10) = 0.15 + 0.85\left(\frac{0.1}{1} + \frac{0.1}{2} + \frac{0.1}{1}\right)$$

$PR(G6) = 0.36$

$$PR_R(G4\text{-}G10) = 0.15 + 0.85\left(\frac{0.19}{1} + \frac{0.15}{2} + \frac{0.23}{1}\right)$$

$PR(G6) = 0.58$



b)

k_{out}	Nodos
3	G4-G10
2	G3
1	G1, G2, G5, G7, G9, Gh
0	G6, G8

$K_{out\ ia}(n) = -k_{out}(n)$

c)

$k_{out\ ia}$	Nodos
0	G6, G8
-1	G1, G2, G5, G7, G9, Gh
-2	G3
-3	G4-G10

Figura 4. Inverso aditivo de las medidas de centralidad. a) Los nodos de la red R-ejemplo (ver Figura 2) que tienen el mismo grado saliente (k_{out}) aparecen con el mismo color. b) La tabla muestra a los genes de R-ejemplo (excepto el gen hipotético Gh) ordenados por categorías, de los de mayor k_{out} a los de menor k_{out}. c) Si los nodos de la misma red se ordenan con respecto al inverso aditivo de su k_{out} ($k_{out\ ia}$) se obtiene una tabla donde las categorías de nodos siguen un orden opuesto

V.5 Medidas topológicas combinadas

Se generaron 190 medidas topológicas combinadas, usando dos medidas topológicas a la vez (de centralidad e inversos de la centralidad). El valor de un nodo según una medida combinada se define como

$$MC(n) = 0.5\ \left(\frac{C_1(n) - \min(C_1)}{\max(C_1) - \min(C_1)} + \frac{C_2(n) - \min(C_2)}{\max(C_2) - \min(C_2)}\right)$$

donde $C_1(n)$ y $C_2(n)$ son los valores del nodo *n* calculados con las medidas de centralidad C_1 y C_2 respectivamente; $\max(C_1)$ y $\min(C_1)$ son el máximo y el mínimo valor obtenidos con la medida de centralidad C_1 en toda la red y $\max(C_2)$ y $\min(C_2)$ son el máximo y el mínimo valor obtenidos con la medida de centralidad C_2 en toda la red.

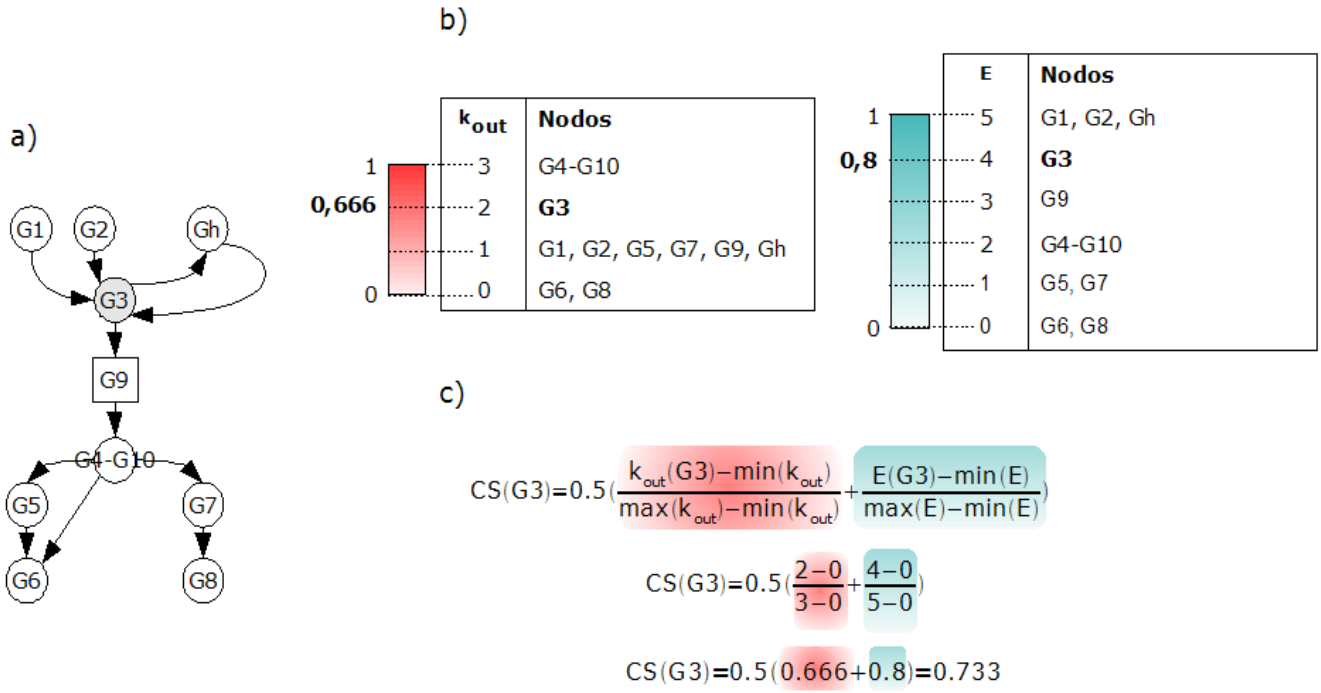


Figura 5. Calculo de las medidas topológicas combinadas. Ver explicación en el texto

La Figura 5 explica la aplicación de una medida que combina al grado saliente (k_{out}) y excentricidad (E) en el nodo G3 la red R-ejemplo (Figuras 5a). Los valores de k_{out} para la red van de cero a 3. Si se normalizan los valores de k_{out} de tal forma que los valores vayan de cero a 1, el nodo G3 obtiene un valor de 0.666. Al normalizar los valores de E el valor del nodo G3 es 0.8 (Figura 5b). El valor de G3 para la medida combinada es el promedio de sus valores de k_{out} y E normalizados (Figura 5c). Las 190 medidas combinadas se aplicaron en los nodos de R-iND750 y RSH-iND750.

V.6 Genes esenciales de R-iND750 y RSH-iND750

Se obtuvo la lista de los 1113 genes esenciales para la viabilidad detectados por el *Saccharomyces* genome deletion project hasta agosto de 2006 (14) y se buscaron a los genes de R-iND750 y RSH-iND750 que aparecían en dicha lista. Con esto se creó una lista de genes esenciales para cada red (las listas no se integraron en las redes sino que permanecieron como archivos independientes). Los genes de una red que no aparecían en su lista de genes esenciales se consideraron no esenciales.

V.7 Listas de genes ordenados por valor topológico

Los genes de cada red se ordenaron por categorías de los de mayor valor a los de menor valor según una medida topológica (como ejemplos ver las Figuras 4b y c, 5b y 6b). Al hacer esto con todas las medidas topológicas se obtuvieron 210 tablas para cada red: 10 para las medidas de centralidad, 10 para sus inversos y 190 para las medidas topológicas combinadas. En estas tablas no se incluyen a los genes hipotéticos. Como se verá a continuación, las tablas se relacionaron con las listas de genes esenciales para evaluar si una medida topológica era capaz de identificar a los genes esenciales.

V.8 Capacidad de una medida topológica para identificar a los genes esenciales

La capacidad de una medida topológica para identificar a los genes esenciales, es decir, para dar mayor valor a los genes esenciales que a los no esenciales se evaluó usando el área bajo la curva receptor-operador (ABC) que es una modificación de la estadística U de Mann-Whitney. El ABC se calcula relacionando una tabla T que contiene a los genes de una red ordenados por valor topológico (ver V.7) con la lista de genes esenciales para la misma red (GE) (ver V.6) (54). En las siguientes secciones se introducen las estadísticas que componen al ABC y se explica el cálculo del ABC.

V.8.a El uso de un valor de corte para comparar la centralidad y la esencialidad

Los valores de centralidad para un gen son variables cuantitativas (que se expresan mediante un número), en tanto que la esencialidad es una variable binaria (un gen es esencial o no es esencial). Una forma de comparar estas variables es establecer un valor corte en T, prediciendo que las categorías que se encuentran por encima del valor corte (fracción superior), contienen genes esenciales, mientras que las categorías que se encuentran por debajo del corte (fracción inferior) incluyen genes no esenciales (Figura 6b). Esta predicción se evalúa comparando las fracciones superior e inferior con la lista GE (ver V.8.b). Diferentes valores de corte producen diferentes fracciones superior e inferior, por lo que la calidad de la predicción es función del valor de corte. Como se explicará más adelante

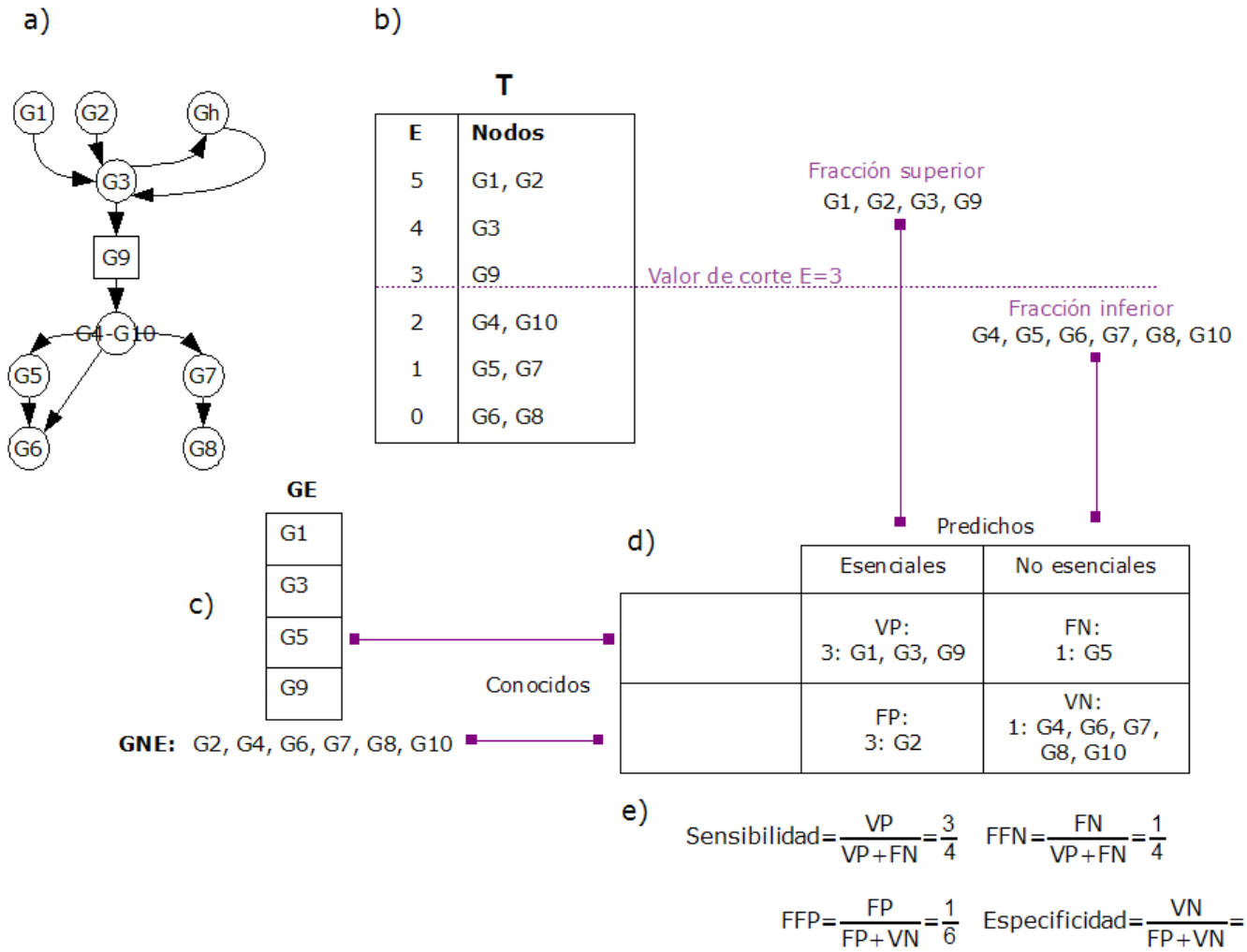


Figura 6. Utilización de un valor de corte para relacionar la centralidad y la esencialidad. a) Red R-ejemplo. b) Nodos de la red R-ejemplo ordenados en una tabla T, de los de mayor a los de menor excentricidad (E) y establecimiento de un valor de corte en T. c) Lista de genes esenciales (GE) y grupo de genes no esenciales (GNE) para la red R-ejemplo. d) Valores calculados al comparar las fracciones superior e inferior de T con GE y GNE. VP: verdaderos positivos, FP: falsos positivos, VN: verdaderos negativos, FN: falsos negativos. e) Estadísticas calculadas a partir de los valores presentados en d. FFP: fracción de falsos positivos, FFN: fracción de falsos negativos. Ver explicación en el texto.

(ver V.8.c) en el cálculo del ABC no se utiliza un sólo valor de corte, sino que se toman en cuenta todos los posibles valores de corte de una tabla T.

V.8.b Relación entre la centralidad y la esencialidad para un valor de corte

Como ya se mencionó, podemos considerar a las fracciones superior e inferior de la tabla T (obtenidas con un valor de corte cualquiera) como el grupo de genes predichos como esenciales y como no esenciales respectivamente. En cambio, los genes que están en la lista GE son genes cuya esencialidad ha sido comprobada. Por lo tanto, podemos considerar a los genes de GE como esenciales conocidos y proponer que los genes de la red que no aparecen en GE son no-esenciales conocidos. Esta clasificación nos permite calcular los siguientes valores (Figura 6c y d) (55):

1) El número de verdaderos positivos (VP) que es el número de genes predichos como esenciales que si aparecen entre los esenciales conocidos.
2) El número de falsos positivos (FP) que es el número de genes que a pesar de ser no-esenciales conocidos aparecen entre los predichos como esenciales.
3) El número de verdaderos negativos (VN) que es el número de genes predichos como no esenciales que si aparecen entre los no esenciales conocidos.
4) El número de falsos negativos (FN) que es el número de genes de la red que a pesar de ser esenciales conocidos aparecen entre los predichos como no esenciales.

Estos valores se utilizan a su vez en el cálculo de cuatro estadísticas que evalúan la bondad de la predicción (Figura 6e) (54, 55):

1) sensibilidad, que es la fracción de los genes esenciales conocidos que si aparecieron entre los predichos como esenciales.
2) especificidad, que es la fracción de los genes no esenciales conocidos que si aparecieron entre los predichos como no esenciales.
2) fracción de falsos positivos (FFP) que es la fracción de los genes no esenciales conocidos que aparecieron entre los predichos como esenciales.
4) fracción de falsos negativos (FFN) que es la fracción de los genes esenciales conocidos que aparecieron entre los predichos como no esenciales.

En el cálculo del ABC se utilizan únicamente la sensibilidad y la fracción de falsos positivos. Como se explicará en la siguiente sección, estas dos estadísticas son suficientes para evaluar la capacidad de predicción de una medida topológica en una red y el uso de otras combinaciones de estadísticas aportaría el mismo tipo de información.

V.8.c Cálculo del ABC y su IC de 99.99%

En el cálculo del ABC no se elige un sólo valor de corte arbitrario para la tabla T, sino que se toman en cuenta todos los cortes posibles de T y para cada corte se calcula la sensibilidad y la FFP. Al graficar la sensibilidad y la FFP obtenidas para cada corte se obtiene una curva

a)

T

E	Nodos		
5	G1, G2	E=5	Sensibilidad=1/4 FFP= 1/6
4	G3	E=4	Sensibilidad=2/4 FFP= 1/6
3	G9	E=3	Sensibilidad=3/4 FFP= 1/6
2	G4, G10	E=2	Sensibilidad=3/4 FFP= 3/6
1	G5, G7	E=1	Sensibilidad=4/4 FFP= 4/6
0	G6, G8	E=0	Sensibilidad=4/4 FFP= 6/6

Sensibilidad= 0/4 FFP= 0/6

b)

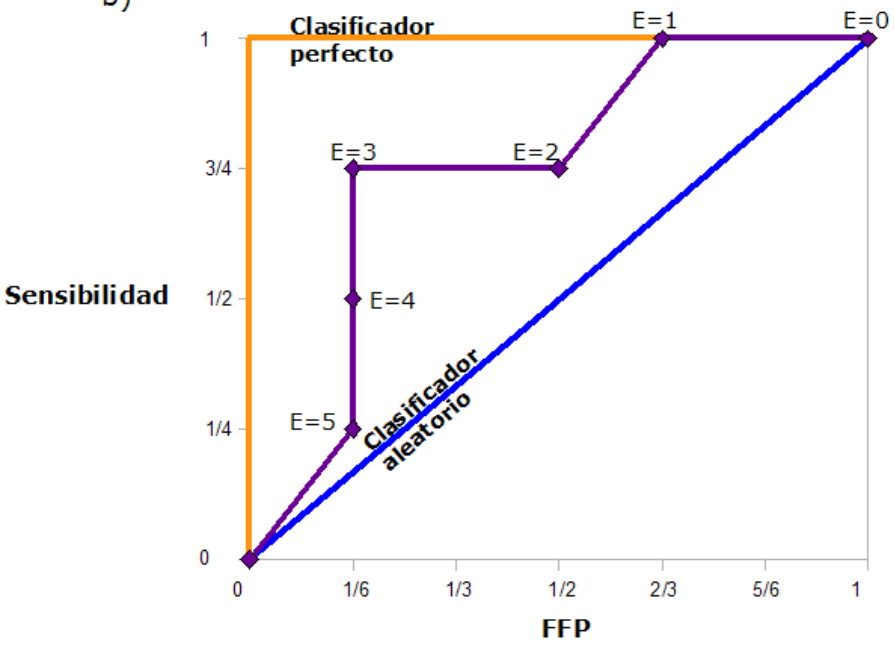


Figura 7. Calculo del área bajo la curva ROC (ABC). a) Se muestran todos los cortes posibles para la tabla T de la Figura 6b y la sensibilidad y FFP correspondiente a cada corte (tomando en cuenta la lista GE y grupo GNE de la Figura 6). b) Curva ROC para la excentricidad (E) en la red R-ejemplo (morado), obtenida al graficar los valores de sensibilidad (eje y) y la FFP (eje x) para cada corte de T. También se muestran las curvas ROC para un clasificador aleatorio (azul) que da la misma importancia a los genes esenciales que a los no esenciales, y para un clasificador perfecto (amarillo) que siempre da mayor valor a un gen esencial que a uno no esencial.

receptor-operador (ROC). El área debajo de esta curva (ABC) equivale a la probabilidad de que la medida topológica evaluada (en una red en particular) le de mayor valor aun gen esencial que a un gen no esencial (ambos elegidos al azar) (56). Si el ABC es igual a 0.5 la medida topológica es un clasificador aleatorio, que no tiende a dar mayor valor a los genes de algún grupo en particular. Una medida topológica que siempre da mayor valor a un gen

esencial que a uno no esencial genera un ABC igual a 1 (Figura 7).

Para evaluar si el ABC obtenida con una medida topológica y red dadas era significativamente mayor que 0.5 se calculó el IC de 99.99% para cada ABC:

$$IC=\pm 3.87*SE(ABC)$$

Donde 3.87 es el valor z para una significancia de 0.01% y SE(ABC) es el error estándar para el ABC (56). Si una medida topológica es significativamente mejor que un predictor aleatorio el IC de 99.99% para su ABC no debe incluir a 0.5.

Cabe mencionar que aunque la curva ROC se construye usando únicamente la sensibilidad y la FFP, otras combinaciones de las estadísticas presentadas en la Figura 6e generan curvas que tienen la misma forma que la curva ROC y también aparecen en el primer cuadrante del plano cartesiano. Aunque estás curvas están reflejadas y/o rotadas con respecto a la curva ROC su integral es equivalente al ABC o bien, es igual a 1-ABC. Sólo las curvas obtenidas al graficar la sensibilidad contra la FFN (y viceversa) ó la especificidad contra la FFP (y viceversa) crean curvas cuya integral es siempre igual a 0.5.

VI. RESULTADOS

Medimos la capacidad de varias medidas topológicas para identificar a los genes esenciales en dos redes basadas en el modelo iND750.

VI.1 Características de los modelos

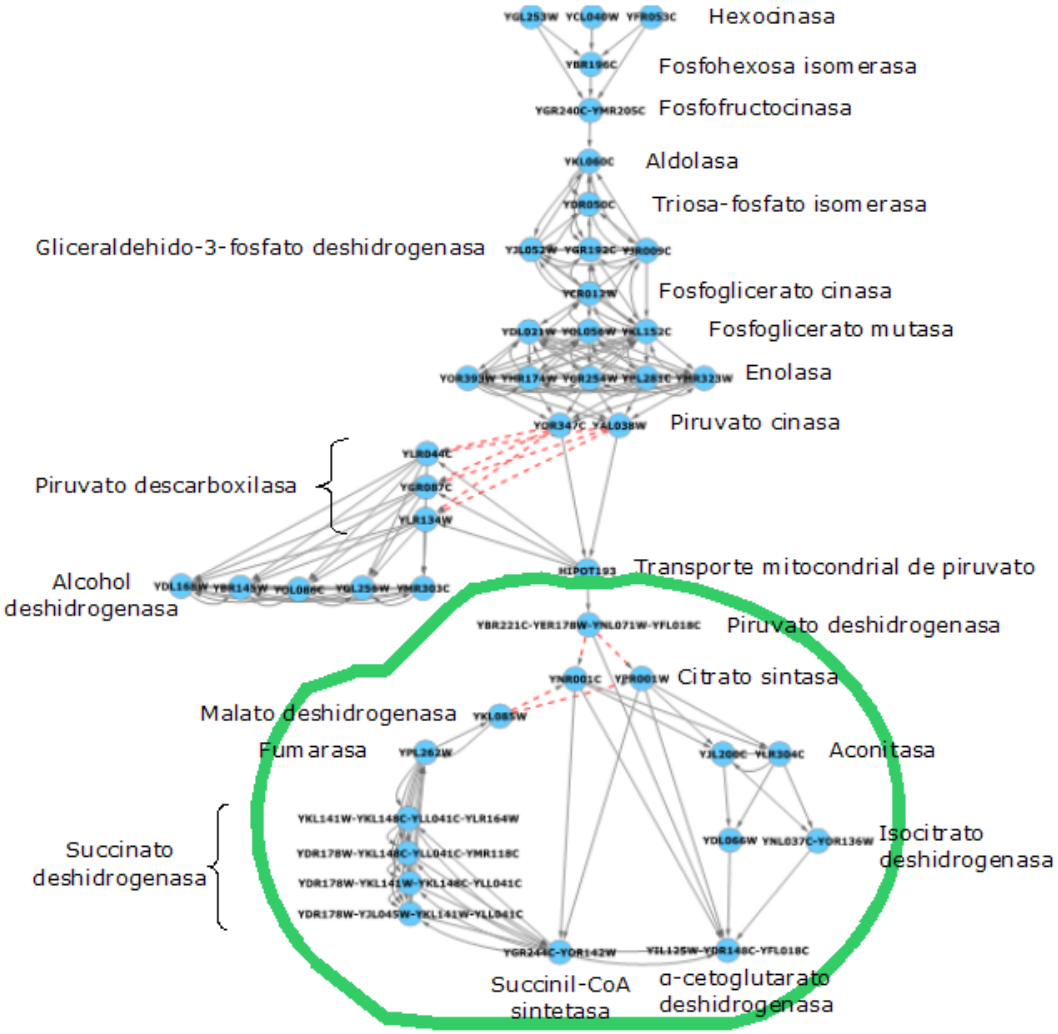
El modelo iND750 lista las reacciones enzimáticas que conforman el metabolismo de *S. cerevisiae* y relaciona cada reacción con una enzima y esta a su vez con un gen o grupo de genes. Se tomó la información de este modelo para crear la red R-iND750 donde los nodos son genes o grupos de genes, que se conectan entre si cuando codifican enzimas que catalizan reacciones sucesivas. R-iND750 incluye 692 genes conocidos y 325 hipotéticos y sus nodos se conectan a través de 5185 aristas.

También se creó la red RSH-iND750 donde se eliminaron los genes hipotéticos y 31 genes conocidos (los que sólo podían unirse a la red a través de un gen hipotético). Así pues, la red RSH-iND750 consta de 661 genes y tiene 2719 aristas.

VI.2 Calidad de los modelos

R-iND750 y RSH-iND750 presentan varios errores en la representación del metabolismo. En

Figura 8. Conexiones entre los nodos de la glucólisis y ciclo de Krebs en iND750. Cada nodo de la glucólisis y ciclo de Krebs (círculos azules) está formado por uno (enzimas monoméricas) o varios (enzimas multiméricas y complejos proteicos) genes. Varios nodos pueden relacionarse con la misma actividad enzimática (enzimas con varias isozimas). La línea verde representa la separación entre la mitocondria y el citoplasma. Algunas conexiones importantes para el funcionamiento de estas vías (líneas rojas punteadas) están ausentes en la red (ver texto).



ambos modelos la glucólisis y ciclo de Krebs carecen de varias conexiones importantes (Figura 8). La vía glucolítica oxida a la D-glucosa, produciendo piruvato, el cual puede destinarse a la fermentación alcohólica o al ciclo de Krebs (57). Sin embargo, en R-iND750 la última enzima de la glucólisis (piruvato cinasa) y la primera de la fermentación alcohólica

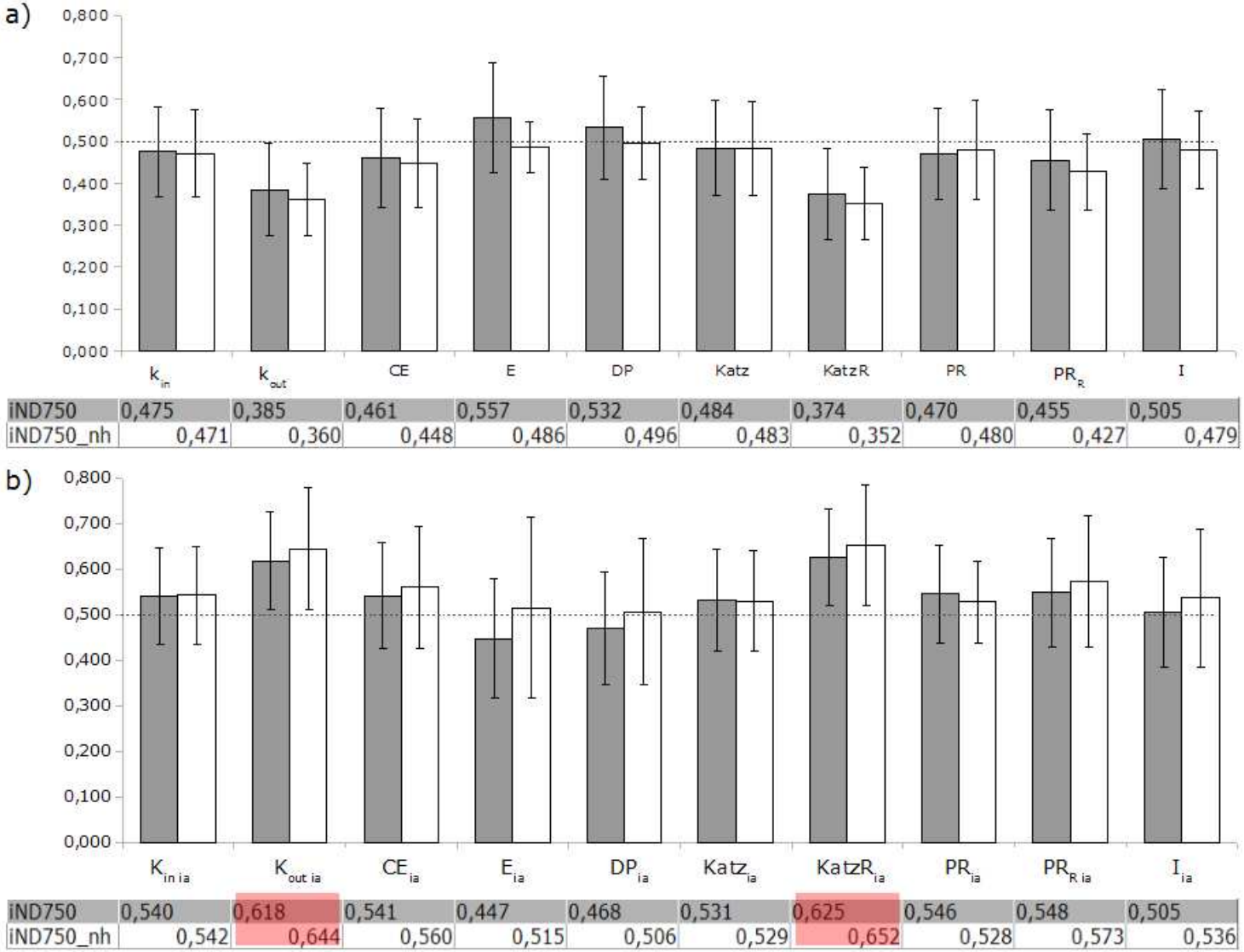
(piruvato descarboxilasa) están desconectadas. Más aún, el complejo piruvato deshidrogenasa, que debería conectar a la glucólisis y el ciclo de Krebs está desconectado de la primera enzima del ciclo de Krebs (citrato sintasa). Además, la última enzima del ciclo de Krebs (malato sintasa) está desconectada de la citrato sintasa, con lo que el ciclo queda incompleto. RSH-iND750 presenta los mismos errores y además carece de un transportador mitocondrial de piruvato, pues esta función está asociada a un gen hipotético.

VI.3 Capacidad de predicción de las medidas topológicas

Según el fenotipo de sus mutantes, hay 111 genes esenciales en R-iND750 y 107 en RSH-iND750. Medimos la capacidad de 10 medidas de centralidad para dar mayor valor a los genes esenciales que a los no esenciales (área bajo la curva ROC o ABC). También se probaron los inversos de las medidas de centralidad y 190 medidas obtenidas al combinar dos medidas de centralidad (o sus inversos) como identificadores de genes esenciales. Se consideró que una medida era capaz de identificar a los genes esenciales si producía un ABC mayor a 0.5 cuyo IC de 99.99% no incluyera a 0.5.

VI.3.a Medidas de centralidad

Como se observa en la Figura 9a, ninguna de las medidas de centralidad fue capaz de



iND750	0,475	0,385	0,461	0,557	0,532	0,484	0,374	0,470	0,455	0,505
iND750_nh	0,471	0,360	0,448	0,486	0,496	0,483	0,352	0,480	0,427	0,479

iND750	0,540	0,618	0,541	0,447	0,468	0,531	0,625	0,546	0,548	0,505
iND750_nh	0,542	0,644	0,560	0,515	0,506	0,529	0,652	0,528	0,573	0,536

Figura 9. Eficiencia de las medidas de centralidad en la predicción de genes esenciales. a) ABC (eje y) obtenida con cada medida de centralidad (eje x) en las redes R-iND750 (barras grises) y RSH-iND750_nh (barras blancas). Se indica el IC de 99.99% para cada caso y el valor de ABC correspondiente (debajo de cada barra). La línea punteada indica el valor de ABC de 0.5. b) Valores obtenidos con los inversos aditivos de las medidas de centralidad. Se indican en rojo los valores de ABC significativamente mayores que 0.5. k_{in}: grado entrante, k_{out}: grado saliente, CE: Coeficiente de empacamiento, E: Excentricidad, DP: Distancia promedio, PR: PageRank, PR_R: PageRankR, I: Intermediación. Los inversos aditivos de estas medidas tienen el mismo nombre seguido del sufijo "ia" en subíndice.

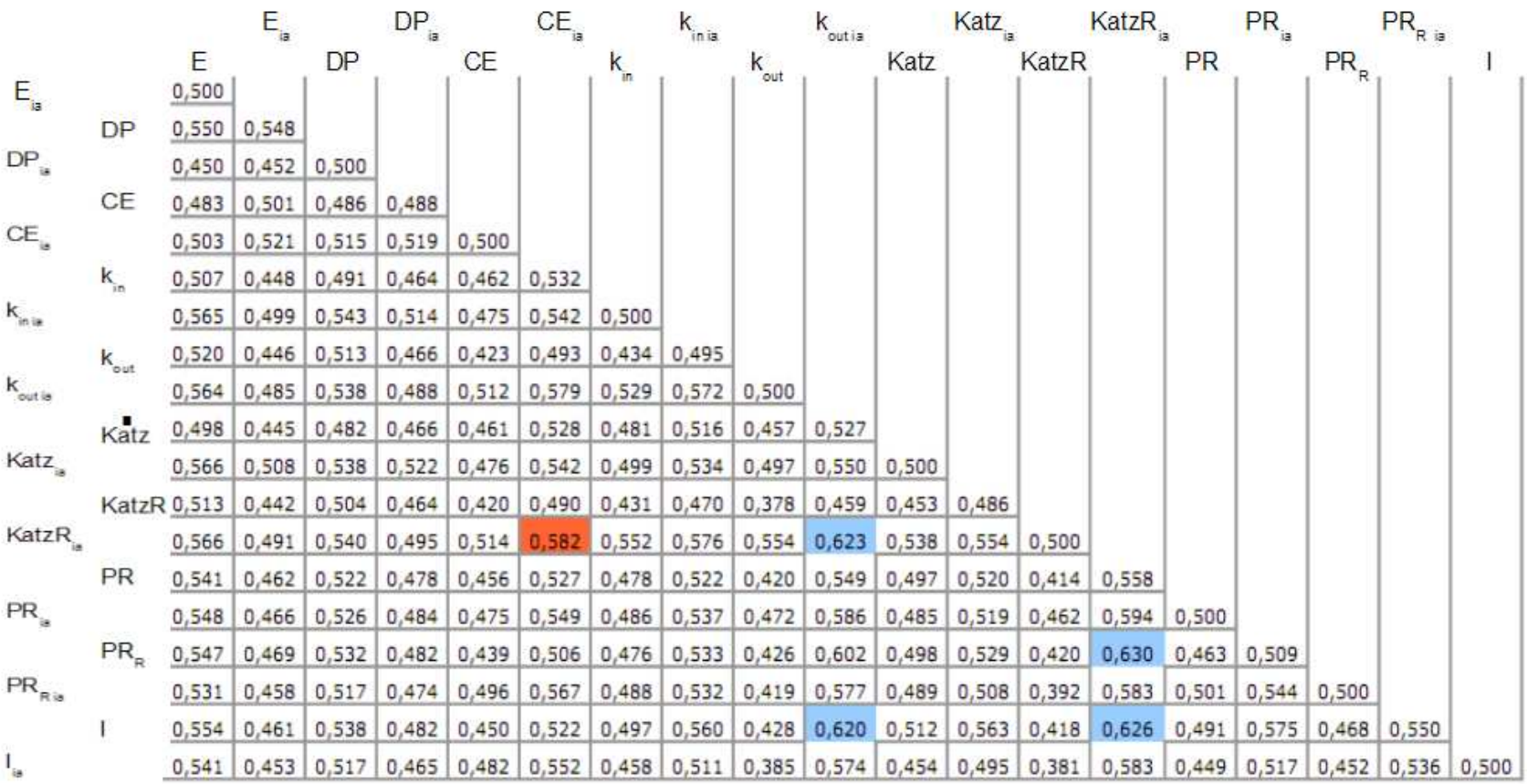
identificar a los genes esenciales en R-iND750 o RSH-iND750. No obstante las medidas de centralidad grado saliente y KatzR produjeron ABCs significativamente menores a 0.5 (observar los IC de 99.99%). Esto demuestra una tendencia para dar menor valor a los genes esenciales que a los no esenciales y es posible que los inversos aditivos de estas medidas tengan un comportamiento opuesto (le den mayor valor a los genes esenciales).

VI.3.b Inversos aditivos de las medidas de centralidad

Al aplicar los inversos de las medidas de centralidad en R-iND750 y RSH-iND750 se comprobó que los inversos aditivos del grado saliente y el KatzR ($k_{out\ ia}$ y $KatzR_{ia}$) son capaces de distinguir a los genes esenciales de los no esenciales (Figura 9b). Los valores más altos de ABC para R-iND750 y RSH-iND750 fueron de 0.625 y 0.652 respectivamente y se obtuvieron con el inverso de KatzR.

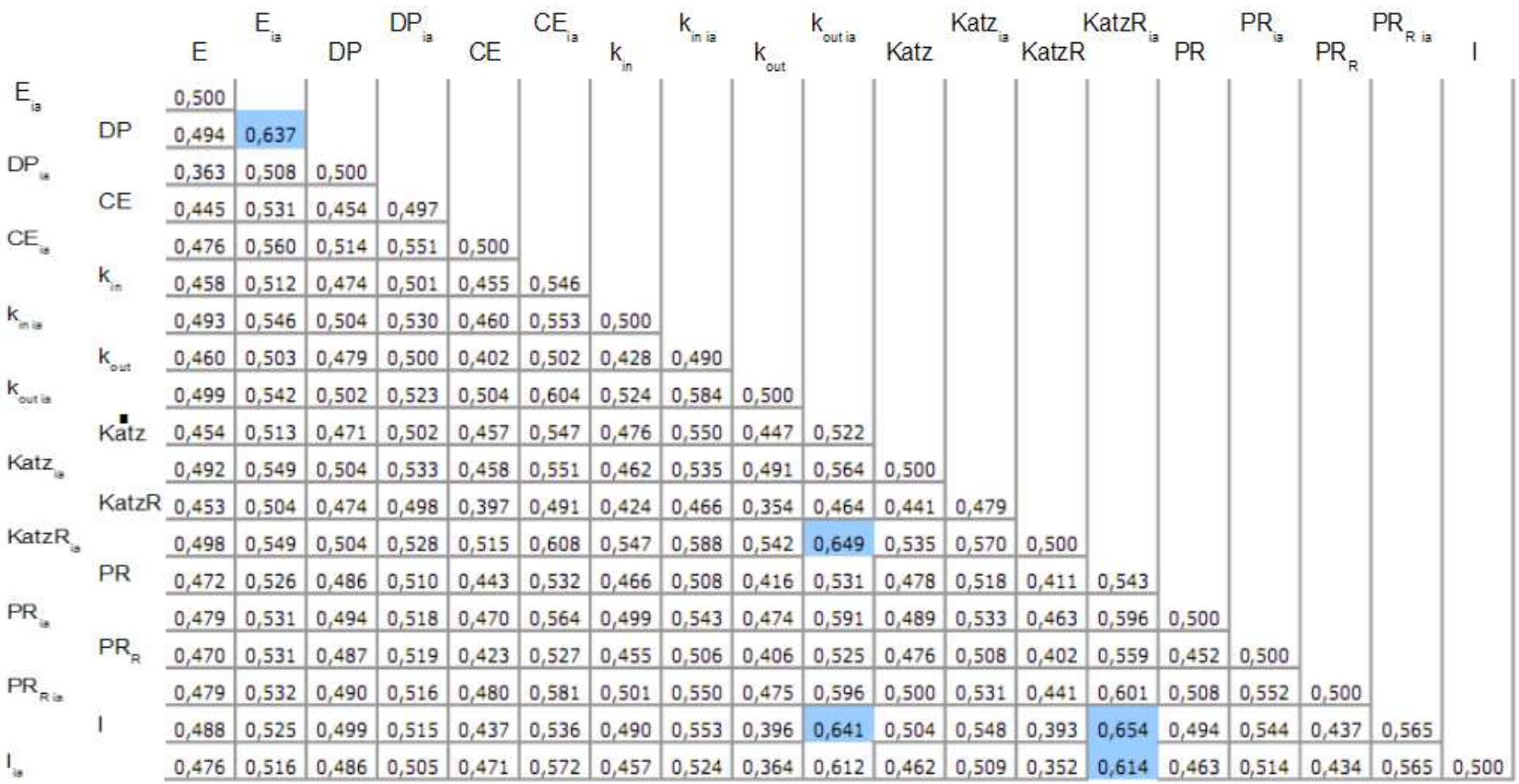
VI.3.c Medidas combinadas

Las medidas de centralidad y sus inversos se combinaron de dos en dos, creando 190 medidas combinadas, seis de las cuales fueron capaces de predecir a los genes esenciales en R-iND750 y/o RSH-iND750 (Figuras 10 y 11). La mayoría de estas medidas combinadas (5 de 6) son funciones del $k_{out\ ia}$ y/o a $KatzR_{ia}$. Los máximos valores obtenidos al aplicar una



	E	E_{ia}	DP	DP_{ia}	CE	CE_{ia}	k_{in}	$k_{in\,ia}$	k_{out}	$k_{out\,ia}$	Katz	$Katz_{ia}$	KatzR	$KatzR_{ia}$	PR	PR_{ia}	PR_R	$PR_{R\,ia}$	I	I_{ia}
E_{ia}	0,500																			
DP	0,550	0,548																		
DP_{ia}	0,450	0,452	0,500																	
CE	0,483	0,501	0,486	0,488																
CE_{ia}	0,503	0,521	0,515	0,519	0,500															
k_{in}	0,507	0,448	0,491	0,464	0,462	0,532														
$k_{in\,ia}$	0,565	0,499	0,543	0,514	0,475	0,542	0,500													
k_{out}	0,520	0,446	0,513	0,466	0,423	0,493	0,434	0,495												
$k_{out\,ia}$	0,564	0,485	0,538	0,488	0,512	0,579	0,529	0,572	0,500											
Katz	0,498	0,445	0,482	0,466	0,461	0,528	0,481	0,516	0,457	0,527										
$Katz_{ia}$	0,566	0,508	0,538	0,522	0,476	0,542	0,499	0,534	0,497	0,550	0,500									
KatzR	0,513	0,442	0,504	0,464	0,420	0,490	0,431	0,470	0,378	0,459	0,453	0,486								
$KatzR_{ia}$	0,566	0,491	0,540	0,495	0,514	0,582	0,552	0,576	0,554	0,623	0,538	0,554	0,500							
PR	0,541	0,462	0,522	0,478	0,456	0,527	0,478	0,522	0,420	0,549	0,497	0,520	0,414	0,558						
PR_{ia}	0,548	0,466	0,526	0,484	0,475	0,549	0,486	0,537	0,472	0,586	0,485	0,519	0,462	0,594	0,500					
PR_R	0,547	0,469	0,532	0,482	0,439	0,506	0,476	0,533	0,426	0,602	0,498	0,529	0,420	0,630	0,463	0,509				
$PR_{R\,ia}$	0,531	0,458	0,517	0,474	0,496	0,567	0,488	0,532	0,419	0,577	0,489	0,508	0,392	0,583	0,501	0,544	0,500			
I	0,554	0,461	0,538	0,482	0,450	0,522	0,497	0,560	0,428	0,620	0,512	0,563	0,418	0,626	0,491	0,575	0,468	0,550		
I_{ia}	0,541	0,453	0,517	0,465	0,482	0,552	0,458	0,511	0,385	0,574	0,454	0,495	0,381	0,583	0,449	0,517	0,452	0,536	0,500	

Figura 10. Eficiencia de las medidas topológicas combinados en la predicción de genes esenciales en la red R-iND750. Se muestran las ABC obtenidas al usar combinaciones de las medidas de centralidad y sus inversos aditivos. Como ejemplo, el valor resaltado en naranja se obtuvo al combinar el inverso aditivo del coeficiente de empacamiento (ver nombre de la columna) y el inverso aditivo del KatzR (ver nombre de la fila) en la red iND750. Las celdas en azul indican las ABC significativamente mayores que 0.5. La nomenclatura de las medidas topológicas es la misma que en la Figura 9.



	E	E_{ia}	DP	DP_{ia}	CE	CE_{ia}	k_{in}	$k_{in\ ia}$	k_{out}	$k_{out\ ia}$	Katz	$Katz_{ia}$	KatzR	$KatzR_{ia}$	PR	PR_{ia}	PR_R	$PR_{R\ ia}$	I
E_{ia}	0,500																		
DP	0,494	0,637																	
DP_{ia}	0,363	0,508	0,500																
CE	0,445	0,531	0,454	0,497															
CE_{ia}	0,476	0,560	0,514	0,551	0,500														
k_{in}	0,458	0,512	0,474	0,501	0,455	0,546													
$k_{in\ ia}$	0,493	0,546	0,504	0,530	0,460	0,553	0,500												
k_{out}	0,460	0,503	0,479	0,500	0,402	0,502	0,428	0,490											
$k_{out\ ia}$	0,499	0,542	0,502	0,523	0,504	0,604	0,524	0,584	0,500										
Katz	0,454	0,513	0,471	0,502	0,457	0,547	0,476	0,550	0,447	0,522									
$Katz_{ia}$	0,492	0,549	0,504	0,533	0,458	0,551	0,462	0,535	0,491	0,564	0,500								
KatzR	0,453	0,504	0,474	0,498	0,397	0,491	0,424	0,466	0,354	0,464	0,441	0,479							
$KatzR_{ia}$	0,498	0,549	0,504	0,528	0,515	0,608	0,547	0,588	0,542	0,649	0,535	0,570	0,500						
PR	0,472	0,526	0,486	0,510	0,443	0,532	0,466	0,508	0,416	0,531	0,478	0,518	0,411	0,543					
PR_{ia}	0,479	0,531	0,494	0,518	0,470	0,564	0,499	0,543	0,474	0,591	0,489	0,533	0,463	0,596	0,500				
PR_R	0,470	0,531	0,487	0,519	0,423	0,527	0,455	0,506	0,406	0,525	0,476	0,508	0,402	0,559	0,452	0,500			
$PR_{R\ ia}$	0,479	0,532	0,490	0,516	0,480	0,581	0,501	0,550	0,475	0,596	0,500	0,531	0,441	0,601	0,508	0,552	0,500		
I	0,488	0,525	0,499	0,515	0,437	0,536	0,490	0,553	0,396	0,641	0,504	0,548	0,393	0,654	0,494	0,544	0,437	0,565	
I_{ia}	0,476	0,516	0,486	0,505	0,471	0,572	0,457	0,524	0,364	0,612	0,462	0,509	0,352	0,614	0,463	0,514	0,434	0,565	0,500

Figura 11. Eficiencia de los algoritmos combinados en la predicción de genes esenciales en la red RSH-iND750. (ver Figura 10 para mayor información).

medida de centralidad combinada son de 0.630 en R-iND750 (combinación entre $KatzR_{ia}$ y PR_R) y de 0.654 en R-iND750 (combinación entre $KatzR_{ia}$ e Intermediación), muy parecidos a los que ya se habían obtenido al aplicar únicamente el KatzR en las mismas redes.

VII. DISCUSIÓN

En este trabajo se evaluó la capacidad de 210 medidas topológicas para distinguir a los genes esenciales de los no esenciales (predecir la esencialidad) en la red metabólica de *S. cerevisiae*. Para tal fin se crearon dos versiones de la red (R-iND750 y RSH-iND750), basadas en el modelo iND750 (37). Tras valorar la calidad de ambos modelos (observando la exactitud con la que representan a la glucólisis y el ciclo de Krebs) se aplicaron las 210 medidas topológicas en cada uno.

Aunque se encontraron varios errores de representación en los modelos, dos medidas topológicas, los inversos aditivos del grado saliente (k_{out}) y el KatzR ($k_{out\ ia}$ y $KatzR_{ia}$, respectivamente), generaron predicciones significativamente mejores que una predicción aleatoria. Este resultado también se obtuvo con otras medidas topológicas, que son función del $k_{out\ ia}$ y/o del $katzR_{ia}$. En las siguientes secciones se integran y analizan estos datos.

VII.1 Fundamento de los resultados obtenidos con el $k_{out\ ia}$ y el $KatzR_{ia}$

El grado saliente o k_{out} y el KatzR son medidas relacionadas. Ambas valoran a un nodo *n* de acuerdo a la posibilidad de alcanzar a otros nodos partiendo de *n*. La diferencia entre estas medidas topológicas radica en que el k_{out} cuenta las aristas salientes de *n* (aristas que van de *n* a otros nodos) y el KatzR cuenta los caminos que salen de *n* hacia otros nodos y que pueden tener una o varias aristas e incluso ser redundantes (ver definición del k_{out} y el KatzR

en la Figura 3b y h). Si en una red existen nodos sin aristas salientes estos tendrán el menor valor de KatzR y k_{out} para toda la red, pues no se puede ir de ellos a algún otro nodo.

El $k_{out\ ia}$ y $KatzR_{ia}$, tienen un comportamiento opuesto al k_{out} y el KatzR. Le dan mayor valor a un nodo *n* entre menos aristas salientes tiene ($k_{out\ ia}$) o menos caminos hay de *n* hacia otros nodos ($KatzR_{ia}$). Ambas medidas le dan el mayor valor a los nodos sin aristas salientes.

En las redes R-iND750 y RSH-iND750 hay nodos sin aristas salientes. En el caso de RSH-iND750 se trata de 85 nodos que incluyen 27 genes esenciales (el 25 % de los genes esenciales de la red) y 64 no esenciales (el 12 % de los genes no esenciales de la red). La mayoría de estos genes esenciales (20 de 27) codifican enzimas que están al final de una ruta biosintética, entre ellas 16 aminoacil-tRNA sintetasas, que en la red aparecen al final de las rutas de síntesis de aminoácidos y enzimas que finalizan la síntesis de heme, tiamina pirofosfato (TPP) y quitina (un componente importante de la pared celular) (58).

Importantemente, la capacidad de predicción del $k_{out\ ia}$ y el $KatzR_{ia}$ dependen principalmente del valor que le dan a los nodos sin aristas salientes, pues al evaluar el resto de los nodos ambas medidas se aproximan al comportamiento de un clasificador aleatorio (la sensibilidad y la FFP aumentan a la par en su curva ROC) (ver Figura 7).

En cuanto a las medidas combinadas. La eficiencia de estas medidas en la predicción es menor o supera por muy poco a los valores obtenidos con el $k_{out\ ia}$ y el $KatzR_{ia}$. Mas aún, la mayoría de estas medidas incluyen al $k_{out\ ia}$ y/o al $KatzR_{ia}$, lo que sugiere que el potencial de

predicción observado no es producto de la combinación entre medidas topológicas, sino de la inclusión del $k_{out\ ia}$ o el $KatzR_{ia}$.

VII.2 Factores que limitan la calidad de las predicciones

De las 210 medidas topológicas estudiadas, sólo el $k_{out\ ia}$ y el $KatzR_{ia}$ muestran una capacidad limitada para identificar a los genes esenciales en R-iND750 y RSH-iND750. Tanto la inexactitud de las redes construidas como las limitantes de las medidas topológicas pueden ser responsables de este resultado.

Se espera que la calidad de R-iND750 y RSH-iND750 determine en gran medida la calidad de las predicciones obtenidas con estas redes. Los errores en la representación de la glucólisis y el ciclo de Krebs (y posiblemente en otras vías metabólicas) podrían imponer un límite al poder de predicción independiente de la capacidad de la técnica de predicción usada para detectar a los genes esenciales.

Por otro lado, las medidas topológicas no capturan los parámetros cinéticos del metabolismo (Ej. velocidades de reacción), que son determinantes en la función de este sistema celular. Además, el $k_{out\ ia}$ y el $KatzR_{ia}$ se enfocan principalmente en los genes que están al final de las rutas biosintéticas y podrían ser inadecuados para identificar genes esenciales en otras regiones de la red metabólica.

Para conocer el verdadero potencial de las medidas topológicas como identificadores de genes esenciales en la red metabólica de *S. cerevisiae*, será necesario estudiar qué tanto afectan cada uno de estos factores la calidad de la predicción.

VIII. CONCLUSIONES

a. Dos medidas topológicas, los inversos aditivos del KatzR y el grado saliente ($k_{out\ ia}$ y $KatzR_{ia}$), tienen una capacidad limitada para diferenciar a los genes esenciales de los no esenciales en las redes R-iND750 y RSH-iND750 de *S. cerevisiae*.

b. La combinación del $k_{out\ ia}$ y $KatzR_{ia}$ entre ellos mismos o con otras medidas topológicas no mejora la predicción de genes esenciales en la redes R-iND750 y RSH-iND750.

c. La baja calidad de las predicciones puede ser consecuencia de la inexactitud de los modelos R-iND750 y RSH-iND750 y/o de las deficiencias del $k_{out\ ia}$ y el $KatzR_{ia}$, que detectan genes esenciales en una región limitada de la red y no capturan los parámetros cinéticos del metabolismo.

REFERENCIAS

1. Westerhoff, H. V.; Palsson, B. O. "The evolution of molecular biology into systems biology". *Nature Biotechnology*. 2004, Vol. 22, No. 10, pp. 1249-52.
2. Fromm, J. *The Emergence of Complexity*. Alemania:, Kassel University Press, 2004. ISBN:978-3-89958-069-3.
3. Backlund, A. “The definition of system”. *Kybernetes*. 2000, Vol. 29, No. 4, pp. 444-51.
4. Wetson A. D.; Hood, L. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, “Preventative, and Personalized Medicine”. *Journal of Proteome Research*. 2004, Vol. 3, No. 2, pp. 179-96.
5. Bruggeman, F. J.; Westerhoff, H. V. “The nature of systems biology”. *TRENDS in Microbiology*. 2006, Vol. 15, No. 1, pp. 45-50.
6. Joyce, A. R.; Palsson, B. O. “The model organism as a system: integrating ‘omics’ data sets”. *Nature Reviews Molecular Cell Biology*. 2006, Vol. 7, No. 3, pp. 198-210.
7. Diestel, R. *Graph Theory*. 2a ed. Berlin, New York: Springer-Verlag, 2000. ISBN: 978-3-540-26183-4.
8. Barabási, A. L.; Oltvai, Z. N. “Network biology: Understanding the cell’s functional organization”. *Nature Reviews Genetics*. 2004, Vol. 5, No. 2. pp. 101–13.
9. Pinz, S., *et al*. “Cotrol of yeast filamentous-form growth by modules in an integrated molecular network”. *Genome Research*. 2004, Vol. 14, No. 3, pp. 380-90.
10. Ho, Y. “Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry”. *Nature*. Vol. 415 (10 de Enero de 2002) pp. 180-83.
11. Uetz, P., *et al*. “A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae”*. *Nature*. Vol. 403 (10 de Febrero de 2000) pp. 623-27.
12. Mustacchi, R.; Hohmann, S.; Nielsen, J. “Yeast systems biology to unravel the network of life”. *Yeast*. 2006, Vol. 23, No. 3, pp. 227-38.
13. Stockholm University. InParanoid: Eukaryotic Ortholog Groups [en linea]. Versión 6.0. [Stockholm, Sweden]: Stockholm Bioinformatics Center, Agosto de 2007, [Consulta: 28 de Octubre de 2007]. Disponible en Web: <http://inparanoid.sbc.su.se/>.
14. Giaever, G., *et al*. “Functional profiling of the *Saccharomyces cerevisiae* genome”. *Nature*. Vol. 418 (25 de Julio de 2002) pp. 387–91.
15. Ghaemmaghami, S., *et al*. “Global analysis of protein expression in yeast”. *Nature*. Vol. 425 (16 de Octubre de 2003) pp. 737–41.
16. Huh, W. K., *et al*. “Global analysis of protein localization in budding yeast”. Nature. Vo. 425 (16 de Octubre de 2003) pp. 686-91.
17. Gavin, A.C., *et al*. “Functional organization of yeast proteome by systematic analysis of protein complexes”. *Nature*. Vol. 415 (10 de Enero de 2002) pp. 141–47.
18. Hohmann, S. “The Yeast Systems Biology Network: mating communities”. *Current Opinion in Biotechnology*. 2005, Vol. 16, No. 3, pp. 356-360.
19. Ideker, T., *et, al*. “Integrated genomic and proteomic analyses of a systematically perturbed. metabolic network”. *Science*. Vol. 292, No. 5518, pp. 929–934.

20. Harbison, C. T., *et al*. “Transcriptional regulatory code of a eukaryotic genome”. *Nature*. Vol. 431 (2 de Septiembre de 2004) pp. 99-104.
21. Lee, T.I. “Transcriptional regulatory networks in *Saccharomyces cerevisiae”. Science.* Vol. 228, No. 5594, pp. 799-804.
22. Boone, C.; Bussey, H.; Andrews, B.J. “Exploring genetic interactions and networks with yeast”. *Nature Reviews Genetics*. 2007, Vol. 8, No. 6, pp. 437-49.
23. Vitkup, D,; Kharchenko, P.; Wagner, A. "Influence of metabolic network structure and function on enzyme evolution". *Genome Biology*. 2007, Vol. 7, No. 5, R39.
24. Kharchenko, P.; Church, G.M.; Vitkup, D. "Expression dynamics of a cellular metabolic network". *Molecular Systems Biology*. 2005, Vol. 1, 2005.0016.
25. Patil, K. R.; Nielsen, J. "Uncovering transcriptional regulation of metabolism by using metabolic network topology". *Proceedings of the National Academy of Sciences USA*. 2005, Vol. 102, No. 8, pp. 2685-89.
26. Kuepfer, L.; Sauer, U.; Blank, L. M. "Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*". *Genome Research*. 2005, Vol. 15, No. 10, pp. 1421-30.
27. Papp, B.; Pál, C.; Hurst, L. D. "Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast" *Nature*. Vol. 429 (10 de Junio de 2004) pp. 661-4.
28. Harrison, R., *et al*. "Plasticity of genetic interactions in metabolic networks of yeast". *Proceedings of the National Academy of Sciences USA*. 2007, Vol. 104, No. 7, pp. 2307-12.
29. Jeong, H., *et al*. "The large-scale organization of metabolic networks". *Nature*, Vol. 407 (5 de Octubre de 2000) pp. 651-4.
30. Ma, H.; Zeng, A. P. "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms". *Bioinformatics*. 2003, Vol. 19, No. 2, pp. 270-7.
31. Fell, D. A.; Wagner, A.; "The small world of metabolism". *Nature Biotechnology*. 2000, Vol. 18, No. 11, pp. 1121-2.
32. Ravasz, E., *et al*. "Hierarchical organization of modularity in metabolic networks". *Science*. 2002, Vol. 297, No. 5586, pp. 1551-5.
33. Ma, H.; Zeng, A. P. "The connectivity structure, giant strong component and centrality of metabolic networks". *Bioinformatics*. 2003, Vol. 19, No. 11, pp. 1423-30.
34. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. "The KEGG resources for deciphering the genome". *Nucleic Acids Res.* 2004, Vol. 32 (Database Issue) pp. D277-80.
35. Diestel, R. *Graph Theory*. 2a ed. Berlin, New York: Springer-Verlag, 2000. ISBN: 978-3-540-26183-4.
36. Förster, J., *et al*. “Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network”. *Genome Research*. 2003, Vol. 13, No. 2, pp. 244-53.
37. Duarte, N. C.; Herrgård, M. J.; Palsson, B. Ø. "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model". *Genome Research*. 2004, Vol. 14, No. 7, pp. 298-309.
38. Kauffman, K. J.; Prakash, P.; Edwards, J. S. “Advances in flux balance analysis”. C*urrent Opinion in Biotechnology*. 2003, Vol. 14, No. 5, pp. 491-96.
39. Segrè, D.; Vitkup, D.; Church, G. M. “Analysis of optimality in natural and perturbed

metabolic networks". *Proceedings of the National Academy of Sciences USA*. 2002, Vol. 99, No. 23, pp. 15112-17.
40. Wunderlich, Z.; Mirny, L.A. "Using the topology of metabolic networks to predict viability of mutant strains". *Biophysical Journal*. 2006, Vol. 91, No. 6, pp. 2304-11.
41. Förster, J., *et al*. "Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae"*. *OMICS*. 2003, Vol. 7, No. 2, pp. 193-202.
42. Hahn, M. W.; Kern, A. D. "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks". *Molecular Biology and Evolution*. 2005. Vol. 22, No. 4, pp. 803-806.
43. Carlson, M. R., *et al*. "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks". *BMC Genomics*. 2006. Vol. 7, No. 40.
44. Levy, S. F.; Siegal, M. L. "Network Hubs Buffer Environmental Variation in *Saccharomyces cerevisiae*". *PLOS Biology*. 2008. Vol. 6, No. 11, pp. E2588-2604.
45. Yu. H. *et al*. "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics". *PLOS Computational Biology.* 2007. Vol. 3, No. 4, pp. 713-720.
46. Joy, M.P., et al. "High-betweenness proteins in the yeast protein interaction network". Journal of Biomedicine and Biotechnology. 2005. No. 2, pp. 96-103.
47. Butcher, E. C.; Berg, E. L.; Kunkel, E. J. "Systems biology in drug discovery". *Nature Biotechnology*. 2004, Vol. 22, No. 10, pp. 1253-59.
48. Shannon, P., *et al*. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *Genome Research*, Vol. 13, No. 11, pp. 2498-504
49. Junker, B. H.; Koschützki, D.; Schreiber, F. "Exploration of biological network centralities with CentiBin". *BMC Bioinformatics*. 2006, Vol. 6, No. 219, pp. 1-7.
50. Borgatti, S. P.; Everett, M. G. "A graph-theoretic perspective of centrality" *Social Networks*. 2006, Vol. 17, No. 4, pp. 466-484.
51. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank citation ranking: bringing order to the web [en linea]. Stanford InfoLab Publication Server, 1999 [fecha de consulta: 13 de febrero de 2008] Disponible en http://dbpubs.stanford.edu:8090/pub/1999-66
52. Buckley, F.; Harary, F. *Distance in Graphs*. Redwood, California: Addison-Wesley Publishing Company, 1990. ISBN: 0-201-09591-2.
53. Thibert, B.; Bredesen, D. E.; del Rio, G. "Improved prediction of critical residues for protein function based on network and phylogenetic analysis". *BMC Bioinformatics*. 2005, Vol. 6, No. 213, pp. 1-15.
54. Lasko, T. A., *et al*. "The use of receiver operating characteristic curves in biomedical informatics". Journal Of Biomedical Informatics. 2005, Vol. 38, No. 5, pp. 404-415.
55. Fawcett, T. "An introduction to ROC analysis". *Pattern Recognition Letters*. 2006, Vol. 27, No. 8, pp. 861-874.
56. Hanley, H. A.; McNeil, B. J. "The meaning and use of the area under a receiver operating characteristic (ROC) curve". *Radiology*. 1982, Vol. 143, pp. 29-36.
57. Nelson, D.L.; Cox, M.M. *Lehninger, principles of biochemistry*. 3a ed. New York: W.H. Freeman, 2000. ISBN: 1-57259-9316.
58. Stanford University. Saccharomyces Genome Database [en linea]. [Stanford, California]: Department of Genetics, Stanford School of Medicine, [Consulta: 24 de Junio

de 2008]. Disponible en Web: <http://www.yeastgenome.org/>.
59. Koschützki, D., *et al*. “Centrality Indices”. En: Brandes, U; Erlebach, T. (eds.) *Network Analysis: Methodological Foundations*. Berlin: Springer-Verlag, 2005. pp. 16-61.