Practico 2 de Introducción a las Redes Complejas con Aplicaciones a la Biología:

Relación entre algunas propiedades topológicas de una proteína y su importancia en el funcionamiento de microorganismos

Integrantes
Yuditsabet Burgos
Alan Givré
Lucía Pedraza

Introducción

El buen funcionamiento de microorganismos depende de factores claves como son las acciones individuales de las proteínas; éstas actúan como moléculas de señalización y bloques de construcción a través de sus propias interacciones moleculares. Algunas proteínas juegan una posición topológica más importante que otras y resultan de una consecuencia fenotípica clave si son eliminadas.

Para tratar cuantitativamente la importancia de algunas proteínas una posibilidad es estudiarlas a través del análisis estructural de las redes que conforman. En esta representación las proteínas son nodos conectados por interacciones físicas directas identificadas. Para esto, el primer paso es identificar la topología de la red, esto es: si presentan una topología uniforme (mismo número de enlaces para cada proteína) o si presenta una topología sin escala altamente heterogénea (conectividades diferentes).

En gran parte de la bibliografía (también en el trabajo práctico previo) ha sido demostrado que las redes de interacciones de proteínas siguen una ley de potencias. Esto indica alta inhomogeneidad, en la que pocas proteínas altamente conectadas desempeñan un papel central en la mediación de las interacciones entre numerosas proteínas menos conectadas (Jeong el al, 2001). Si se analiza cuánto afecta a la fragilidad de la red la eliminación de proteínas altamente conectadas, se obtiene que el diámetro de la red aumenta rápidamente al eliminar estos nodos altamente conectados (Hubs) de manera mucho más rápida que si se eliminan nodos de manera aleatoria. Esto conduce a asumir que la eliminación de las proteínas menos conectadas deberían ser menos esenciales que las altamente conectadas. Esta hipótesis es probada correlacionando la conectividad con la indispensabilidad de las proteínas. Este trabajo ha sido realizado por Jeong (Jeong el al, 2001), en donde observa que efectivamente las proteínas esenciales son con una probabilidad mayor aquellas que tienen mayor grado en la red de proteínas y son aquellas que al ser eliminadas atentan contra la robustez del funcionamiento de los microorganismos.

Posteriormente en el trabajo de He (He el al, 2006) se postula que esta correlación entre el grado de los nodos y su esencialidad biológica en el funcionamiento no depende directamente del grado, sino que responde a la existencia de enlaces esenciales. Cuando un enlace es esencial los dos nodos que están involucrados también lo son. De esta

manera, los nodos con mayor cantidad de enlaces tendrán más probabilidad de ser parte de un enlace esencial, y por lo tanto de ser esenciales, correspondiendo con el fenómeno observado por Jeong. Para comprobar esta hipótesis los autores se basan en el hecho de que de existir estos enlaces esenciales, el número de enlaces entre nodos esenciales debería ser alto en comparación con una red similar en el grado de sus nodos, pero recableada de manera random. De esta manera observan que efectivamente el fenómeno de centralidad y letalidad puede explicarse con la existencia de enlaces esenciales, donde alrededor del 3% de los enlaces de las redes que analizan son esenciales, y donde el 43% de los nodos esenciales son explicados de esta manera.

Por último, Zotenko (Zotenko, et al, 2008) demuestra que el fenómeno de centralidad y letalidad es en realidad explicado debido a que la mayor parte de los nodos esenciales forman parte de un complejo biológico de proteínas, que cumple una función esencial para el funcionamiento del microorganismo. Para esto demuestra que remover los nodos esenciales de una red es menos disruptivo (en relación a el tamaño de la red) a la remoción de una cantidad similar de nodos con un nivel de centralidad similar (ya sea el grado de los nodos u otras medidas de centralidad). De esta forma el grado de vulnerabilidad de la red no depende necesariamente de la esencialidad de los nodos, sino de la centralidad de estos, y que por lo tanto la hipótesis de que la esencialidad biológica se relaciona con su rol en la conectividad de la red no es cierta. Por lo tanto el trabajo desarrolla una nueva hipótesis que explica la esencialidad de un nodo por su pertenencia a un complejo biológico de proteínas.

Con estos estudios se verifica que de la organización de las interacciones y las posiciones topológicas de las proteínas individuales se obtendrá una mejor comprensión de la dinámica celular y la robustez a partir de un enfoque integrado que incorpora simultáneamente las propiedades individuales y contextuales de todos los constituyentes en redes celulares complejas.

El objetivo de este trabajo es explorar la vulnerabilidad y esencialidad en las redes de interacción de proteínas de levaduras: Y2H, APMS, YID-LIT y LIT-REGULARY mediante los análisis reportados en los trabajos de (Zotenko et al, 2008 y He et al, 2006).

Características de las redes analizadas

Nuestras redes de interacciones de proteínas en levadura provienen de diferentes fuentes. La red de purificación por afinidad-espectrometría de masas (APMS) se basa en la purificación bioquímica de proteínas a partir de extractos celulares. Después de la purificación, las proteínas que están unidas (presas) a la proteína purificada (cebo) se determinan usando espectrometría de masas (https://medicapage.com/index.php?newsid=2644). Las redes YID_LIT y LIT_REGULARY fueron construidas a partir de interacciones reportadas en literatura y por último la levadura dos híbrida (Y2H) resulta de datos experimentales de rendimiento con un alto nivel de confianza en las interacciones (Zotenko et al, 2008).

En la tabla 1 se muestra un resumen de las principales características extraídas de cada una de las redes. La red LIT_REGULARY condensa la mayor cantidad de nodos y enlaces de todas las redes, sin embargo, presenta un coeficiente de clustering local bajo lo que sugiere que es una red muy dispersa con pocas conexiones locales. Por otro lado la red APMS presenta un grado medio alto que se corresponde con la densidad de enlaces en la red y además con un coeficiente de clustering local mayor a 0.5. Esto es consecuencia de que el método APMS recoge las relaciones entre las proteínas a través de complejos encontrados en conjunto, lo cual genera estructuras de cluster. En cambio, la red Y2H, al ser extremadamente selectiva, tiene particularmente una baja densidad, y una extremadamente baja clusterización.

Tabla 1. Propiedades estructurales de las redes de proteínas

ID	clusteringlocal	enlaces	kmean	nodos
apms	0.554636	9070	11.183724	1622
YID_lit	0.292492	2925	3.808594	1536
y2h	0.046194	2930	2.903865	2018
lit_regulary	0.261134	11858	7.171454	3307

Para analizar cuántos enlaces en común comparten las redes entre sí calculamos la superposición entre ellas. En la tabla 2 se muestra qué porcentaje de enlaces de una red está presente en otra. Por empezar, se señala que, como las YID_lit y y2h tienen una menor cantidad de enlaces, es mucho más probable que un enlace de estas redes pertenezca a la redes de alta cantidad de enlaces, que viceversa. Por ejemplo, 16% de los enlaces de y2h están presentes en la red lit_regularly. Pero sólo 4% de los enlaces de la red lit_regularly están presentes en la red y2h. Se puede observar que, en tanto que el 98% de los enlaces de YID_lit están presentes en lit_regularly, esto significa que la red lit_regularly prácticamente incluye a la red YID_lit. La red APMS contiene muchos enlaces, muchos de los cuales son espurios, lo cual es necesario tener en cuenta. Más adelante analizaremos cómo se comporta la esencialidad y letalidad para cada una de estas redes.

Tabla 2. Fracción de interacciones en común entre las redes

apms	*0.443761*	0.0887372	0.212515
0.143109	**YID_lit**	0.0887372	0.241187
0.0286659	0.0888889	**y2h**	0.0403947
0.277839	*0.977778*	0.163481	**lit_regulary**

Relación entre el grado de los nodos y su esencialidad

Como vimos antes, dada una red podemos identificar los hubs (aquellos nodos con mayor grado) y podemos identificar cuáles de esos nodos son esenciales biológicamente. En esta sección vamos a demostrar que esas dos condiciones están correlacionadas para las redes que estamos analizando, cumpliendo así el fenómeno de centralidad y letalidad.
Definimos la fracción de nodos esenciales como la cantidad total de nodos esenciales entre la cantidad de hubs totales (donde la definición de hubs depende del grado que decidamos tomar a partir del cual clasificamos los nodos como hubs). Graficamos y comparamos la fracción de nodos esenciales para cada grado, verificando que cuanto más son los nodos considerados hubs, menos es la fracción de esenciales. Para eso, para cada posible límite para definir un hubs (comenzando por la situación donde sólo el nodo de mayor grado es hub), calculamos la fracción de nodos esenciales. Para realizar esta operación en tiempos más rápidos ordenamos primero los nodos de mayor a menor y en cada iteración calculamos solo cuántos de los nuevos posibles hubs son esenciales, utilizando la información del paso anterior. En la figura 1 se muestra este resultado para cada red, donde claramente se puede observar que el grado de cada nodo en las redes de levaduras se correlaciona con el efecto fenotípico de su eliminación, como se demuestra en los trabajo citados anteriormente. Esto muestra en esencia, la relación que había mostrado Jeong, es decir, que las proteínas esenciales tienden a tener un valor de grado alto en casi todas las redes. Podemos observar en cambio que este fenómeno no es notorio para la red Y2H en donde la esencialidad es similar para los nodos de todos los grados.

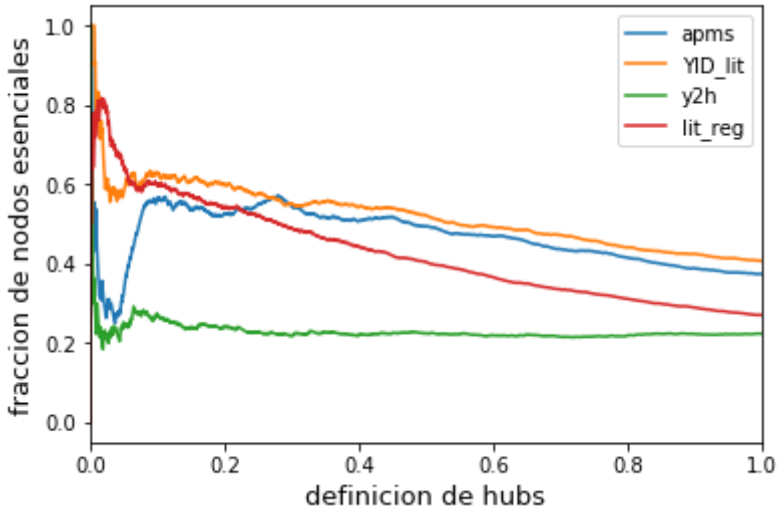

Figura 1. Relación entre el grado y la esencialidad en las redes. Fracción de nodos esenciales

Análisis de vulnerabilidad

Si la esencialidad de los nodos estuviera relacionada con la desconexión en la red que generan al ser eliminados, este efecto debería manifestarse siendo mayor la desconexión que al eliminar otros nodos de grados similares. Sin embargo, según los trabajos de

Zotenko esto no es así. Siguiendo este trabajo, comparamos la eliminación de estos nodos con la eliminación de una misma cantidad de nodos, elegidos con distintos criterios de centralidad e incluso al azar.
Según la bibliografía existen varias formas de medir la centralidad de un nodo dentro de una red. Índices como centralidad de grado, intermediación, vector propio, subgrafos y flujos serán aplicados a nuestras redes para estudiar diferentes aspectos topológicos en las mismas. Estos índices serán comparados con una eliminación aleatoria de nodos con lo cual será posible hacer un evaluación. Primeramente se eligió la componente principal de cada red y se extrajo un nodo aleatorio para eliminar de esta componente; se recalculó la componente principal a partir de la componente que queda después de la extracción del nodo y en cada extracción se va calculando la fracción de nodos extraídos y el tamaño de la componente más grande. El mismo procedimiento es aplicado para el resto de los índices, sólo difiere en la elección del nodo que será extraído.
Para garantizar que los tiempos computacionales sean posibles, durante la mitad de las iteraciones se sacaron 10 nodos por cada componente principal elegida, asumiendo que en cada paso el cambio de cuál es la componente principal no es tan grande. Durante la segunda mitad la componente principal fue calculada en cada iteración. Se extrajeron nodos hasta que el tamaño de la red sea la mitad de la original.
Como se puede observar, nuestro resultado es significativamente diferente del de Zotenko. Suponemos que nuestro método, al recalcular la componente principal ante cada extracción de los nodos y remover nodos solamente de la componente principal genera diferencias con el método de Zotenko.

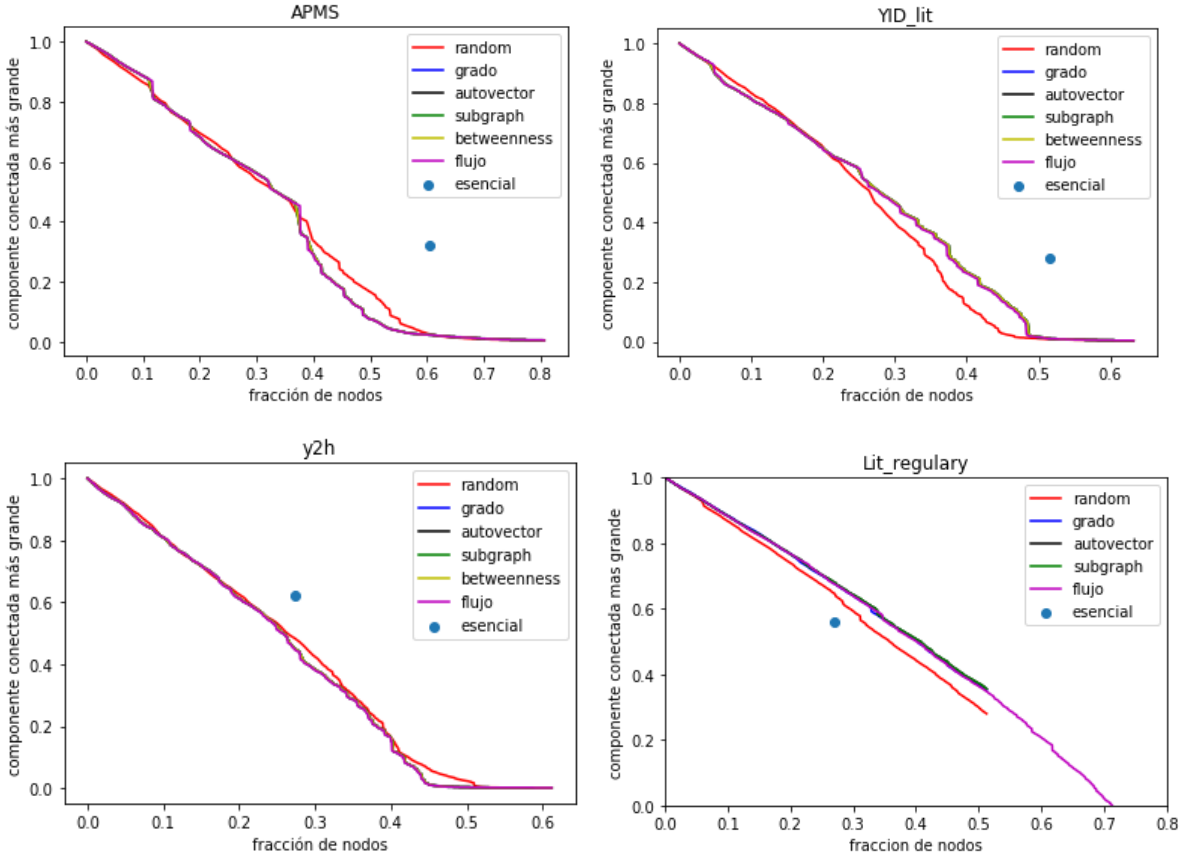

Figura 2. Vulnerabilidad de ataque contra las proteínas centrales. El impacto de la eliminación de nodos se cuantifica por la fracción de nodos en la componente conectado mayor. Hay una curva para cada medida de centralidad que muestra la fracción de nodos en el componente conectado más grande.

Seguidamente analizamos más concretamente el poder de interrupción de las proteínas más conectadas de los centros esenciales. Para esto se seleccionaron las proteínas esenciales y se extrajeron de la red, comparando el resultado con la extracción de proteínas de forma aleatoria respetando la distribución de grados de las proteínas esenciales. Realizamos 100 selecciones de proteínas random y tomamos el promedio y su desvío standard. Con este análisis vemos si, dejando fija la distribución de grado, repercute más en la topología de la red eliminar proteínas de forma aleatoria o eliminar proteínas esenciales. Como se muestra en la tabla 3 la eliminación de proteínas esenciales no es más destructivo que la eliminación de un número aleatorio equivalente de proteínas en tres de las redes. Sin embargo, en la red de YID_Lit sucede que sí se destruye más rápidamente la red con la eliminación de proteínas esenciales.

Tabla 3. Impacto de remover proteínas esenciales comparado con remover un número equivalente de proteínas no esenciales de forma aleatoria con la misma distribución de grado

	esencial	**random**
apms	0.200370	0.210561+/-0.032005
YID_lit	0.222005	0.078268+/-0.032994
y2h	0.509415	0.488578+/-0.011024
lit_regulary	0.560629	0.513441+/-0.010841

Esencialidad: Módulos biológicos vs. Interacciones Esenciales

He et al, 2006 habían propuesto que si bien la correlación que habían mostrado Jeong et al, 2001 era correcta, la interpretación que había otorgado no lo era. Propone como una causa alternativa que existan pares de interacción proteína - proteína (PPI de aquí en más) que sean esenciales. Una interacción esencial entre dos proteínas hace que ambas proteínas sean esenciales, porque la eliminación de cualquiera de las dos proteínas causa letalidad o infertilidad debido a la interrupción de la interacción.

Según esta hipótesis, en las redes conviven nodos que son esenciales por ser parte de un par esencial, con nodos esenciales por otras razones. Asumiendo que la proporción de los primeros es α y de los segundos es β la probabilidad de que una proteína no sea esencial es que no esté en un enlace esencial, y que no sea esencial por otras razones. De esta manera, la probabilidad de tener un nodo esencial dado un nodo de grado k es

$$(1)\quad P_e = 1 - (1-\beta)(1-\alpha)^k$$

Por lo tanto al graficar ln(1-Pe) en función de k nos encontramos con una función lineal cuyos coeficientes son ln(1-\alpha) y ln(1-\beta). Solo calculamos los nodos con grados menores a 10, debido a la poca cantidad de nodos que tenemos disponibles de los otros grados, que hace que no podamos obtener resultados confiables.

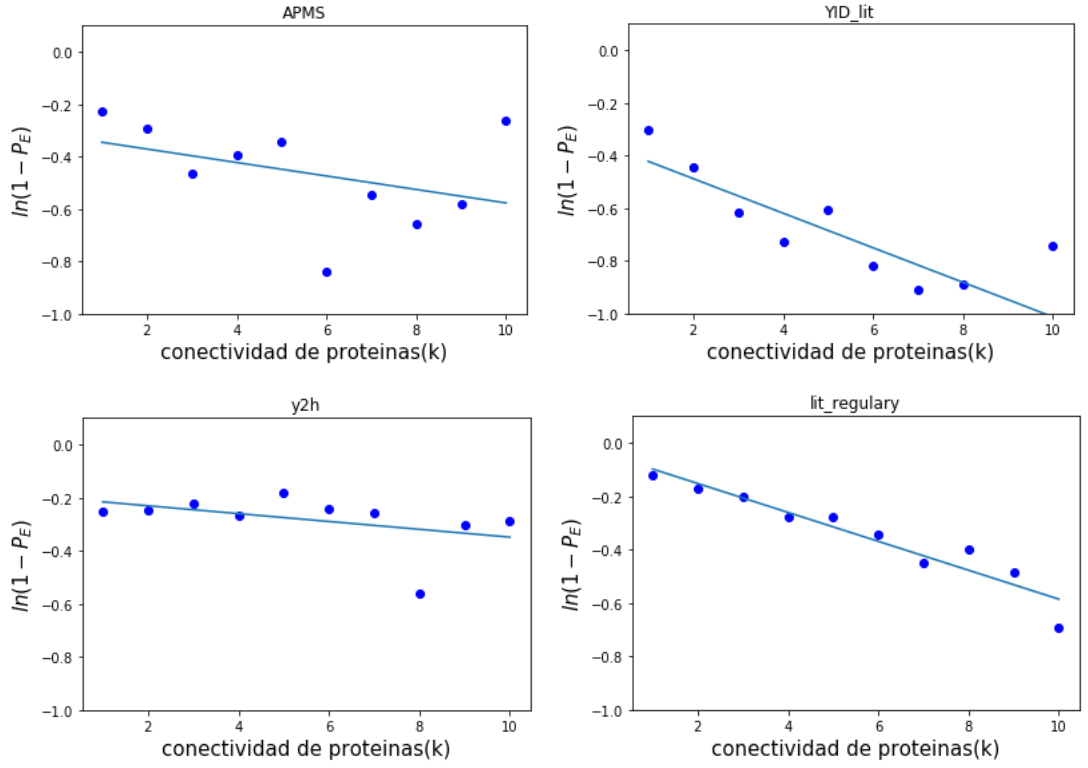

Figura 3. Regresión lineal entre la probabilidad de que una proteína sea no esencial (1- P_E) y la conectividad (k) de la proteína.

Por último, si la hipótesis de He es correcta, la esencialidad de dos nodos que no interactúan entre sí deberían ser independientes, aún cuando esos nodos tengan varios vecinos en común. En el trabajo de Zotenko (et al, 2008), este es el argumento que utiliza para refutar la hipótesis de He. Siguiendo ese razonamiento para cada red calculamos la cantidad total de pares no vecinos, pero con 3 o más vecinos en común. Luego calculamos cuántos de esos pares son ambos escenciales, o ambos no esenciales.
Ese valor, asumiendo como cierta la hipótesis de He, puede ser calculado a través de la ecuación (1) y utilizando los valores de α y β utilizados en el punto anterior puede ser calculado como

$$A = e^{ln(1-\beta)+ln(1-\alpha)k_1}\, e^{ln(1-\beta)+ln(1-\alpha)k_2} + (1 - e^{ln(1-\beta)+ln(1-\alpha)k_1})(1 - e^{ln(1-\beta)+ln(1-\alpha)k_2})$$

Para calcular el error obtenido asumimos una aproximación lineal. De manera que el error es equivalente a la derivada del valor obtenido en el ajuste dando como resultado que

$err = (e^{ln(1-\beta)+ln(1-\alpha)k_1} + e^{ln(1-\beta)+ln(1-\alpha)k_2}) * \xi$ donde ξ es el error residual del ajuste.

En la tabla 4 podemos ver el resultado obtenido. En el trabajo de Zotenko, estos valores indican que la cantidad de pares reales es mayor, escapando al margen de error calculado. De esta manera Zotenko concluye que la hipótesis de He no es correcta, y que en la esencialidad de las proteínas no sólo está involucrado su vecindad, sino que hay proteínas esenciales por ser parte de complejos biológicos esenciales.

En el caso APMS la cantidad de pares del mismo tipo es menor al ajuste. Sin embargo esa red no fue analizada tampoco en el trabajo original, dado que la vecindad está determinada por pertenecer al mismo complejo.
En el caso YID_lit y y2h se obtiene un valor de nodos del mismo tipo mayor al del ajuste, en concordancia con los resultados de Zotenko. Sin embargo los resultados obtenidos están dentro del margen de error. Debido a ser una red esparza en el trabajo original se utilizaron pares con más de uno vecino en lugar de 3, pudiendo ser esta la explicación a la poca cantidad de pares obtenidos en nuestro caso, y por lo tanto la diferencia en el resultado.
La última red en cambio, al igual que el análisis de Zotenko, nos lleva a concluir que el número de nodos del mismo tipo es mayor que el esperado, incluso por fuera del rango de error, falseando de esta manera la hipótesis de He y dando lugar al desarrollo de la teoría de Zotenko por la cual la esencialidad de los nodos también está relacionada con la pertenencia a complejos biológicos que son esenciales para el funcionamiento del individuo y no solamente responde a una distribución de nodos ni de aristas esenciales.

Tabla 4. Comparación de la cantidad de nodos de un mismo tipo entre nodos no vecinos pero con 3 o más vecinos en común en las redes y en según el resultado teórico obtenido en el punto anterior.

ID	número del mismo tipo	número total	Ajuste	Error
APMS	11814	23226	12861	2570
YID_lit	778	1460	771	143
y2h	704	1044	583	123
lit_reg	12374	21554	11558	809

Realizamos el mismo procedimiento, pero utilizando nodos que no sean vecinos pero tengan al menos un nodo en común, especialmente para observar las redes YID_lit y y2h que resultaban muy esparzas y por lo que la cantidad de pares encontrados no era significativos. Sin embargo en este caso los resultados no concuerdan con los obtenidos por

Zotenko, ya que en todos los casos, si bien el número de pares presentes en la red es la mayor al ajuste, está contemplada en el margen de error, por lo que sería concordante con la hipótesis de He.

ID	número del mismo tipo	número total	Ajuste	Error
YID_lit	11970	20526	10366	2540
y2h	30174	46146	28486	6065

Bibliografía

- H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai (2001): Lethality and centrality in protein networks; Nature Brief Bommunications.

- Xionglei He, Jianzhi Zhang (2006): Why Do Hubs Tend to Be Essential in Protein Networks; Plos genetics.

- Elena Zotenko, Julian Mestre, Dianne P. O'Leary, Teresa M. Przytycka (2008): 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.