Redes Complejas - Trabajo Computacional No.2

Carlos Ríos Chavéz[1], Rodrigo Sieira[1], Andrés Troiano[1], and Marco Villagran[1]
[1] *UBA, CABA, Argentina*

22 de octubre de 2018

ABSTRACT
En el presente trabajo se aborda por diferentes metodos de verificación la relación entre la esencialidad, la centralidad y la letalidad de proteínas de la levadura *Saccharomyces cerevisiae* en base a cuatro diferentes redes de sus interacciones.

Key words: Proteínas – Escencialidad – Centralidad

1 INTRODUCCIÓN

Las proteínas esenciales son aquellas estrictamente necesarias para la vida de un microorganismo dado. Es decir, su remoción por mutación o deleción del gen que las codifica tiene consecuencias letales. La regla de la centralidad-letalidad (centality-lethality rule) fue postulada por Jeong et al. (2001) a partir de la observación de que en redes de interacciones de proteínas existe una mayor probabilidad de que los nodos más conectados (*hubs*) correspondan a proteínas esenciales de lo que cabría esperar por azar. Dichos autores interpretaron que el carácter funcional de las proteínas esenciales se debe al rol estructural que éstas ejercen en el mantenimiento de la conectividad de la red de interacciones. Sin embargo, esta idea fue posteriormente desafiada por los autores He & Zhang (2006) Zotenko et al. (2008) quienes propusieron explicaciones alternativas para la regla de centralidad-letalidad. Mediante distintas metodologías utilizadas por dichos grupos, en el presente trabajo computacional analizaremos cuatro redes de interacciones de proteínas de la levadura *Saccharomyces cerevisiae*: la red de interacciones binarias determinada por el método de doble híbrido de levaduras *yeast-two hybrid* (red Y2H), red de complejos multiproteicos AP-MS, y dos redes de interacciones relevadas a partir de datos de la literatura a las que llamaremos red LIT y red REG, respectivamente.

2 CARACTERÍSTICAS DE LAS REDES

En primer lugar, utilizando el script *1.Tabla_1_Zotenko.py* (ver apéndice) analizamos las características generales de las cuatro redes determinando para cada una de ellas el número de nodos, número de enlaces, grado medio $< k >$ y coeficiente de clustering $< C >$. Como se puede observar en la Tabla 1, la red AP-MS posee el mayor valor de grado medio y el mayor coeficiente de clustering, de acuerdo a lo esperado por el modo en el que fue relevada (inmunoprecipitación y asignación de enlaces con todas las proteínas pertenecientes a un mismo complejo). La red REG presentó el mayor número de nodos y enlaces, mostrando también un alto valor de $< k >$ comparable al de la red AP-MS. Por otro lado, la red LIT mostró el menor valor de $< C >$.

	Y2H	AP-MS	LIT	REG
No. nodos	2013	1619	1537	3292
No. enlaces	2928	9070	2925	11853
$< k >$	3.62	9.01	3.81	7.20
$< C >$	0.05	0.56	0.29	0.26

Tabla 1. Características principales de las redes evaluadas.

Posteriormente se realizó un estudio del solapamiento de enlaces entre las distintas redes, calculando la cantidad de interacciones que una red tiene en común con otra sobre el total de interacciones de la red utilizando el script *2.Tabla_2_Zotenko.py* (ver apéndice). Como se puede observar en la Tabla 2, el 9 % de los enlaces de la red Y2H están presentes también en la red AP-MS. Cabe destacar el bajo grado de solapamiento que la red Y2H tiene con las demás. Sin embargo, dado que cada una de las redes fue relevada mediante técnicas diferentes, no es sorprendente que los enlaces en común reportados difieran sustancialmente. Luego se analizó en qué medida se relaciona la esencialidad de las proteínas con el grado en el que sus nodos correspondientes se encuentran conectados en las distintas redes mediante el script *3.Figura_1_Zotenko.py* (ver apéndice). Como se puede observar en la Figura 1, de acuerdo con la regla de la letalidad-centralidad, la proporción de proteínas esenciales de las redes AP-MS, LIT y REG, aumenta a medida que se incrementa el grado de sus nodos. En la red Y2H, sin embargo, se observa que la fracción de proteínas esenciales se mantiene constante durante la mayor parte de la curva y con valores sustancialmente más bajos que otras redes relevadas por distintos métodos. La regla de la centralidad-letalidad se verifica en numerosas redes de interacción de proteínas de levaduras y de otros organismos (He & Zhang (2006)), y por lo tanto el comportamiento observado para la red Y2H en la Fig. 1 podría ser un indicador de defectos en el releva-

Y2H	0.09	0.09	0.17
0.03	AP-MS	0.14	0.28
0.09	0.44	LIT	0.99
0.04	0.22	0.24	REG

Tabla 2. Solapamiento de las redes analizadas. En cada fila se observa el grado de solapamiento de una red con respecto a las demás, indicando la fracción de nodos contenida en el resto de las redes.

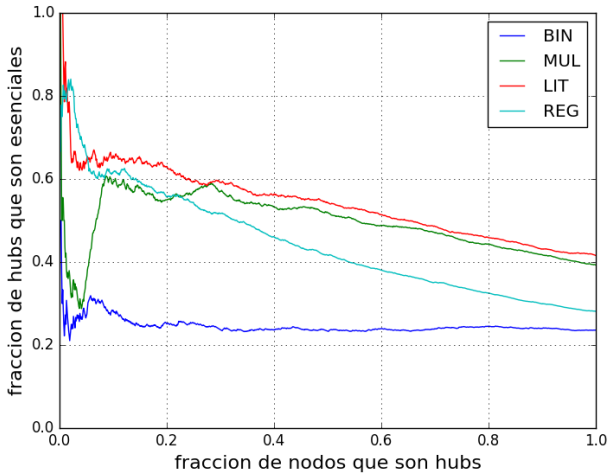

Figura 1. Relación entre esencialidad y grado. Para cada una de las redes se analizó la proporción de proteínas esenciales contenida en los hubs, definidos como la proporción de nodos que contienen los grados de mayor valor de la red.

miento de la misma que podrían haber derivado en una alta proporción de falsos positivos y/o falsos negativos.

3 ANÁLISIS DE VULNERABILIDAD

Además del grado, existen otros coeficientes de centralidad que indican la medida en que un nodo determinado es importante para mantener la conectividad global de la red. Siguiendo los pasos del trabajo de Zotenko et al. (2008), para cada una de las redes se determinaron distintos índices de centralidad mediante el script *4.Figura_2_Zotenko.ipynb* (ver apéndice). Los mismos incluyen índices de centralidad local que reportan la influencia de los vecinos inmediatos, tales como *eigenvector* (indica en qué medida un nodo está conectado con nodos de alto grado) y *subgraph* (indica si un nodo está conectado con caminos cerrados como triángulos o cuadrados), y por otro lado se analizaron índices de intermediatez (*betweenness*) tales como *shortest-path* (indica el promedio de la distancia del camino más corto entre un nodo y el resto de los nodos de un grafo) y *current flow* (la mayor cantidad de geodésicas que pasan a través de un nodo dado). Para estudiar el rol de los nodos en la conectividad de las distintas redes se analizó el impacto de la remoción de los mismos cuantificando el tamaño relativo de la componente gigante resultante.

Como se puede observar en la Figura 2, en las redes Y2H, LIT y REG el índice de intermediatez *shortest-path* mostró ser más importante para mantener la conectividad de todas las redes analizadas junto con el índice grado. En la red AP-MS, sin embargo, el índice *current flow* mostró la mayor capacidad disruptiva, mostrando un comportamiento notoriamente distinto al del resto de las redes, probablemente indicativo de sesgos introducidos por el método de relevamiento. En todos los casos se observó que la remoción de nodos al azar tuvo un impacto menor que la remoción por índices de centralidad, confirmando que dichos indicadores efectivamente expresan la importancia del rol de los nodos para mantener la conectividad de una red en función de su topología. Por último, la Figura 2 también mostró que la remoción de la totalidad de los nodos correspondientes a proteínas esenciales tuvo un impacto menor que la remoción de nodos al azar sobre el mantenimiento de la integridad de todas las redes. Esta última observación representa un argumento en contra de la interpretación de Jeong et al. (2001) acerca de la regla de la centralidad-letalidad, ya que en todos los casos las proteínas esenciales no mostraron tener un rol importante en el mantenimiento de la arquitectura de las redes.

En la Tabla 3 se analiza el efecto de la remoción de nodos no esenciales siguiendo la misma distribución de grado que la de los nodos esenciales utilizando el script *5.Tabla_3_Zotenko.py* (ver apéndice) en 100 iteraciones. Como resultado se observó que la remoción de la totalidad de nodos esenciales de la red LIT produjo como consecuencia una componente gigante que contiene un 21,2 % de la totalidad de nodos de la red, mientras que la remoción de nodos no esenciales derivó en una componente gigante un 30 % más grande (29,9 % de la totalidad de nodos) (Tabla 3). Por lo tanto, para la red LIT los resultados del presente análisis coinciden con la interpretación de la centralidad-letalidad de Jeong *et al.* (2001), dado que la remoción de nodos esenciales tuvo una capacidad disruptiva claramente mayor que la remoción de nodos no esenciales. Sin embargo, en la red REG se observó lo contrario dado que, aún con escasas diferencias no mayores a un 9 %, la remoción de nodos esenciales resultó en una componente gigante de mayor tamaño en comparación con la remoción de nodos no esenciales. Si tomamos en cuenta que la red REG constituye un relevamiento de datos de la literatura aún más exhaustivo que el de la red LIT, podemos asumir que esta última observación constituye un resultado más confiable,opuesto a la centralidad-letalidad de Jeong *et al.* (2001). En el resto de las redes, Y2H y AP-MS, se observaron tendencias en ambos sentidos también con escasas diferencias no mayores a un 12 % (Tabla 3). Cabe destacar que el desvío estándar observado mostró valores muy bajos que oscilan en el rango de un 0,8 y 3 % con respecto al valor medio. Esto pudo deberse a que la restricción de remover nodos no esenciales en mismo número y distribución que los esenciales deja pocas combinaciones posibles para la extracción de nodos de alto grado. Por lo tanto, los valores de desvío estándar probablemente sufrieron un sesgo hacia valores muy bajos, y las diferencias observadas en las redes Y2H, AP-MS, y REG podrían no ser significativas.

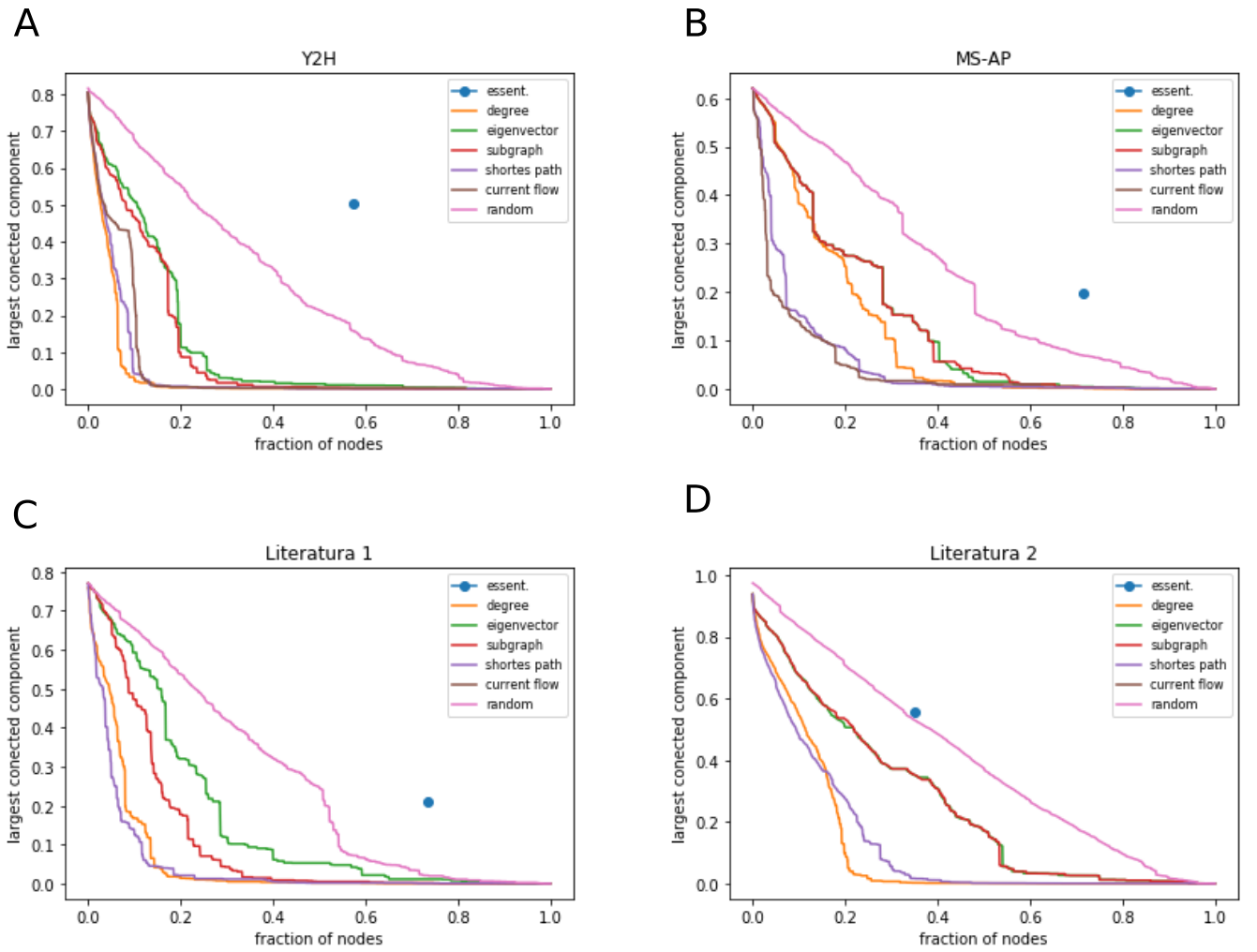

Figura 2. Análisis de vulnerabilidad. Se determinó el impacto de la remoción de nodos de las redes Y2H (A), AP-MS (B), LIT (C) y REG (D) de acuerdo a sus índices de grado, eigenvector, subgraph, shortest-path, current flow o removiendo nodos al azar (random) sobre el tamaño de la componente gigante resultante. Los puntos azules indican el tamaño de la componente gigante luego de remover la totalidad de los nodos esenciales.

Red	RemEs	RemNoEs	Std
Y2H	0.504	0.466	± 0.011
AP-MS	0.198	0.224	± 0.007
LIT	0.212	0.299	± 0.003
REG	0.555	0.508	± 0.004

Tabla 3. Comparación del impacto de la remoción de proteínas esenciales o no esenciales sobre la integridad de la red. Para cada red se indica el la proporción de la cantidad de nodos de la componente gigante luego de remover la totalidad de las proteínas esenciales (RemEs) o luego de remover una cantidad equivalente de proteínas no esenciales con misma distribución de grado (promedio RemNoEs más desvío standard luego de 100 iteraciones).

4 ESENCIALIDAD DE LAS INTERACCIONES ENTRE PROTEÍNAS

La verificación experimental de la esencialidad de una proteína en un sistema es un proceso complejo y que no da resultados completamente confiables, ya que la metodología del experimento muy seguramente llega a afectar el resultado al dañar el sistema. En He & Zhang (2006) los autores tratan el problema de la determinación de la esencialidad de una proteína planteando la existencia de interacciones esenciales entre proteínas. La hipótesis elemental del trabajo mencionado consiste en declarar a un nodo (o proteína) central como esencial, por el hecho de tener muchas interacciones, de las cuales un número importante son esenciales. El tener un nodo declarado como esencial por las condiciones ya expuestas relaja la dependencia que tiene la esencialidad de las proteínas con la topología de la red de sus interacciones, esto en comparación con trabajos anteriores, e. g. Jeong et al. (2001).

Teniendo una base de datos de I interacciones entre p proteínas y comparándolo con una lista confiable de proteínas esenciales (N_E) conocidas se puede obtener el número de interacciones entre N_Es en una red, llamemos a este número I_{PE}. Este número puede ser mayor al numero de interacciones esenciales reales del sistema, I_E , ya que no todas las interacciones entre N_Es son esenciales, pues la esencialidad de una proteína puede ser dada por factores ajenos a una I_{PE} (véase He & Zhang (2006)).

Para probar la hipótesis propuesta se generan redes de

Datos	Nodos Esenciales	Enlaces Esenciales	α	β
REG	897	6518	0.040	0.14
AP-MS	610	5786	0.035	0.19
Y2H	459	546	0.009	0.11
LIT	634	2100	0.040	0.20

Tabla 4. Cálculo de los valores simulados de α y β.

control re-cableando aleatoriamente (y repitiendo el proceso una cantidad significativa de veces) las conexiones entre proteínas cuidando que cada una de ellas conserve su grado k de conectividad. En cada re-cableado se puede calcular el número de interacciones entre proteínas esenciales en la red aleatoria, que llamaremos $\mathrm{RI_{PE}}$. Con este número en mente se define la cantidad α,

$$\alpha = \frac{\mathrm{I_{PE}} - < \mathrm{RI_{PE}} >}{\mathrm{I}}, \quad (1)$$

que representa el porcentaje de $\mathrm{I_E}$ debida solamente a $\mathrm{I_{PE}}$[1]. Para tomar en cuenta los factores adicionales que pueden transformar una proteína en esencial se definirá la cantidad β. Para obtener este número generaremos redes aleatorias en las que se remueve la esencialidad de las proteínas y después se asignan aleatoriamente ($\mathrm{I_{PE}}$ - $\mathrm{RI_{PE}}$) enlaces, donde $\mathrm{RI_{PE}}$ se selecciona aleatoriamente de la distribución que generamos al calcular α. Después se declaran como esenciales los nodos que estén unidos por estas interacciones, la cantidad de nodos marcados de esta forma generalmente será menor al número de nodos que se tienen originalmente. A continuación se agregan aleatoriamente los nodos restantes, lo cual emula las proteínas que son esenciales por factores ajenos. Este proceso puede incluir nuevos enlaces esenciales en la red, con lo cual se puede definir β como el promedio de la fracción de nodos esenciales agregados en cada repetición.

De este modo se obtuvieron los valores de α y β a través de redes simuladas (tabla 4). Asumiendo que la probabilidad de que una proteína sea esencial se puede calcular como,

$$P_E = 1 - (1-\beta)(1-\alpha)^k, \quad (2)$$

la cual se puede reescribir como

$$ln(1-P_E) = k ln(1-\alpha) + ln(1-\beta), \quad (3)$$

donde $ln(1-P_E)$ varía en forma lineal con k. Como se muestra en la Figura 3, se calcularon los valores de $ln(1-P_E)$ observados para cada red excluyendo los nodos de alto grado, debido a que constituyen muestras de muy bajo tamaño. A partir de la regresión lineal de dichos valores se calcularon los valores de α y β a partir de las ordenadas de origen y de las pendientes obtenidas (script *figura_2_he.py* del Apéndice). Se puede apreciar que la regresión lineal mostró mejores ajustes para las redes LIT y REG que para Y2H y AP-MS, con coeficientes de correlación R^2 de $0,889$, $0,926$, $0,329$ y $0,402$, respectivamente (ver tabla 5).

Como se puede observar en las tablas 4 y 5, con excepción de la red AP-MS, los valores de α y/o β de las distintas redes difirieron entre los distintos métodos. Dado que REG

[1] Suponiendo que siempre $\mathrm{I_{PE}} > \mathrm{RI_{PE}}$.

	α	β	R^2
Y2H	0.021	0.161	0.37
AP-MS	0.041	0.222	0.40
LIT	0.072	0.280	0.89
REG	0.049	0.058	0.93

Tabla 5. Valores de α y β obtenidos mediante las regresiones lineales que se muestran en las figuras 3, junto con los respectivos parámetros de bondad R^2.

	No. total de pares	No. de pares mismo tipo	No. esperado (simulación)	No. esperado (regresión)
Y2H	522	349	230	190
AP-MS	11612	5965	3330	3562
LIT	707	375	191	198
REG	10793	6195	3270	3315

Tabla 6. El número total de pares se refiere al número de pares de proteínas no adyacentes que tienen tres o más vecinos en común. Se les clasifican como del mismo tipo si ambas proteínas en el par son esenciales o bien si ambas no lo son.

hasta el momento constituye la red más confiable de acuerdo a la exhaustividad de su relevamiento, a priori nuestros resultados indicarían que el método de cálculo de proporción de enlaces proteína-proteína esenciales y de proteínas esenciales no involucradas en enlaces esenciales difiere entre lo observado por regresión y lo esperado por azar, lo cual no se ajustaría la hipótesis de He y Zhang (2006).

Paso siguiente, procedimos a realizar un análisis que Zotenko *et al.* (2008) utilizaron para desafiar la hipótesis de enlaces proteína-proteína esenciales de He y Zhang (2006). Según estos últimos autores, si dos proteínas no interactúan entre sí la esencialidad de una sería independiente de la esencialidad de la otra, incluso cuando ambas proteínas comparten vecinos entre sí. Para contrastar esta afirmación primero se calculó en cada red la cantidad total de pares de nodos que no interaccionan entre sí pero comparten 3 o más vecionos directos. Luego se determinó la cantidad observada de pares del mismo tipo, de acuerdo a su carácter esencial o no esencial, y por último se calculó la cantidad de pares del mismo tipo esperada de acuerdo a los parámetros de α y β de He y Zhang (2006) obtenidos ya sea por el método de redes simuladas o por el método de regresión lineal (scripts *tabla_5_zotenko.py* y *tabla_5_zotenko_analisis.py* del Apéndice).

Como se puede observar en la tabla 6, en todos los casos se obtuvo un número esperado de pares de nodos del mismo tipo sustancialmente menor al observado, lo cual indicaría que la esencialidad de las proteínas no se distribuye al azar en una red de interacciones, refutando la hipótesis de He y Zhang (2006) en concordancia con los resultados de Zotenko *et al.* (2008) donde postulan que la esencialidad de las proteínas depende de su pertenencia a los llamados ECOBIMs (Essential Complex Biological Modules).

5 CONCLUSIONES

Con los métodos utilizados en el presente trabajo se pudo tener una idea de la heterogeneidad de las redes empezando por sus características generales que, de acuerdo con

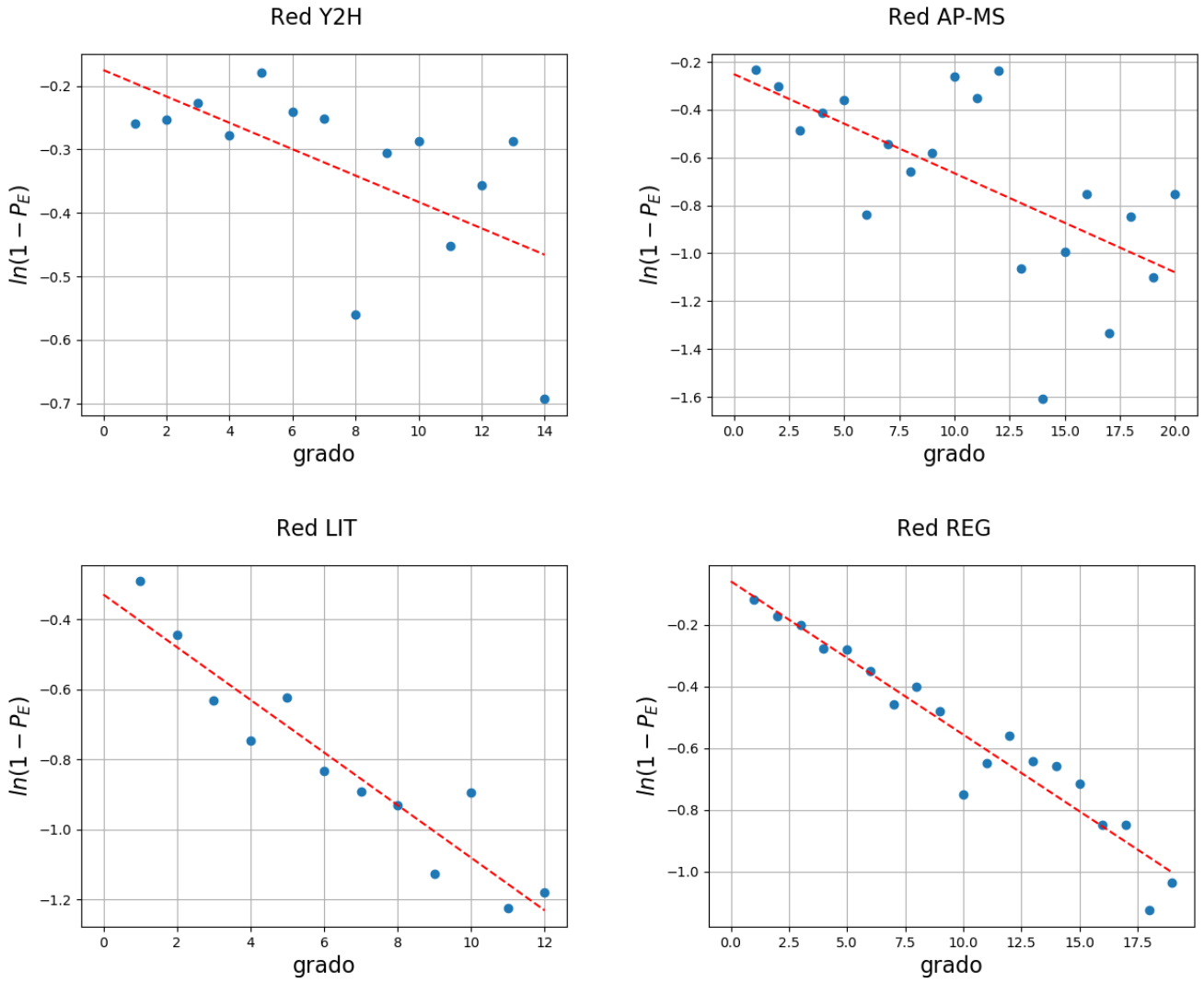

Figura 3. Regresión lineal de $ln(1 - P_E)$ en función de k.

el modo en que fueron relevadas, presentan diferencias esperables (e.g., un mayor grado medio para la red AP-MS, un mayor número de nodos y enlaces para la red REG de datos curados de la literatura relevados de 30 mil publicaciones).

Se observó que tres de las cuatro redes analizadas muestran un claro comportamiento acorde con la regla de centralidad-letalidad, con excepción de la red de interacciones binarias Y2H. Ésta última mostró además un coeficiente de clustering sustancialmente menor al del resto de las redes, y un muy bajo grado de solapamiento. En base a estas observaciones, inferimos que, en caso de contar con una base de datos de esencialidad de proteínas en una red dada, la medición de proporción de proteínas esenciales en función de fracción de hubs revela si la red bajo estudio se ajusta a la regla de centralidad-letalidad, aportando un indicador de la calidad del relevamiento.

Los análisis de vulnerabilidad en cambio, muestran un comportamiento similar para todas las redes analizadas y aportaron resultados que muestran que en redes de interacciones de proteínas de levaduras los nodos correspondientes a proteínas esenciales no cumplen un mayor rol estructural que las proteínas no esenciales. Por lo tanto nuestro análisis claramente se inclina en favor de la refutación a la interpretación de Jeong et al. (2001) de la regla de la centralidad-letalidad, que relaciona causalmente el rol estructural de una proteína en el mantenimiento de la conectividad de una red con su importancia funcional en el organismo.

El análisis de la esencialidad de las interacciones entre proteínas mostró que en las redes estudiadas en el presente trabajo práctico, en proteínas que corresponden a pares de nodos que no interactúan entre sí pero comparten vecinos, la esencialidad de una no es independiente de la otra de acurdo a la hipótesis de interacciones esenciales de He y Zhang (2006). Nuestros resultados por lo tanto se inclinarían en favor de la hipótesis de Zotenko *et al.* (2008) que postula la existencia complejos multiprotéicos cuyos componentes comparten una función biológica específica y se hallan enriquecidos en proteínas esenciales, mientras que otros complejos multiprotéicos de acuerdo a su funcionalidad carecerían por completo de proteínas esenciales.

APPENDIX A: HERRAMIENTAS

En la dirección https://github.com/andres1074/Redes-TP2 se pueden encontrar los códigos que han sido mencionados a lo largo del trabajo y que fueron utilizados para obtener las cantidades aquí presentadas.

Referencias

He X., Zhang J., 2006, PLOS Genetics, 2, 1

Jeong H., Mason S. P., Barabási A.-L., Oltvai Z. N., 2001, Nature, 411, 41 EP

Zotenko E., Mestre J., O'Leary D. P., Przytycka T. M., 2008, PLOS Computational Biology, 4, 1