Trabajo Computacional Nº 2

Fernando Cornes - Juan Herrera Mateos - Ignacio Sticco
Grupo 3

Redes Complejas con Aplicaciones a Sistemas Biologicos
17 de octubre de 2018

1. Introducción

Las proteínas son cadenas complejas de aminoácidos (el orden y disposición de los mismos dependen del código genético de cada ente) unidos por un tipo de enlaces conocidos como "enlaces peptídicos". Los mismos cumplen diversos papeles dentro del funcionamiento de los seres vivos, y a su vez interactúan con otras biomoléculas con el fin de efectuar funciones de control, regulación, transporte, recepción, defensa, entre otras. Tradicionalmente se las identifica según la base de sus acciones individuales como catalizadores, moléculas de señalización, o bloques de construcción en células y microorganismos.

Debido a que las proteínas adquieren su función específica al interaccionar con otras proteínas, la mayoría de los procesos biológicos están regulados por las interacciones entre las mismas [1] [4]. Por tal razón, en los últimos años se ha invertido un gran esfuerzo en la identificación de las interacciones proteína-proteína (PPI) con el objetivo de conocer por completo el funcionamiento de los mecanismos celulares y permitir así el diseño de nuevos enfoques terapéuticos más efectivos [6].

Hoy en día en el estudio de tales interacciones se pueden encontrar dos tipos de enfoques principales: análisis experimentales y enfoques computacionales. El fundamento del enfoque basado en las redes de PPI o interactoma, como hemos visto a lo largo de la materia, es que sus características topológicas pueden revelar información biológica relevante. En este contexto, una estrategia recurrente consiste en la identificación de nodos relevantes, de acuerdo a distintos índices de centralidad de la red, con la intención de poder reconocer entidades biológicas significativas.

Una pregunta que podemos hacernos en este análisis es si las características biológicas de una proteína, en particular su esencialidad -o la necesidad de una dada proteína para la supervivencia-, puede explicarse en términos de su ubicación en el interactoma, es decir, si la centralidad topológica tiene alguna implicancia en la importancia biológica. Una de las primeras conexiones que se encontró fue la llamada regla de centralidad-letalidad, observada por Jeong *et. al.* [3], quienes demostraron que los nodos de alto grado en la red PPI de la levadura *Saccharomyces cerevisiae* contienen más proteínas esenciales de las que se espera por azar. Observaron que el grado de un nodo de las redes PPI de la levadura se correlaciona con el efecto fenotípico de su supresión y que un nodo de alto grado es tres veces más probable que sea esencial que un nodo de bajo grado. Ellos

sugirieron que el número excesivo de proteínas esenciales entre nodos de alto grado es atribuible al papel central que desempeñan los hubs (o nodos de muy alto grado) en la mediación de las interacciones entre proteínas menos conectadas. La eliminación de los hubs interrumpe la conectividad dentro de la red, medida por su diámetro o por el tamaño de su componente gigante (o mayor componente conectada), en mayor medida que la eliminación de un número equivalente de nodos al azar. Así, bajo el supuesto de que la función de un organismo depende de la conectividad entre varias partes de su interactoma, los hubs serían predominantemente esenciales porque juegan un papel central en el mantenimiento de la misma.

Por otro lado, en una investigación llevada a cabo por He *et. al.* [2] se concluyó que la mayoría de las proteínas son esenciales debido a su participación en una o más interacciones proteína-proteína esenciales que se distribuyen uniformemente al azar a lo largo de la red. Bajo esa hipótesis se propone que los hubs son predominantemente esenciales porque están involucrados en más interacciones (tienen alto grado) y por lo tanto, es más probable que se tengan más enlaces de tipo esencial.

En el trabajo de Zotenko et. al [7] se encontró que ninguna de las razones anteriores es determinante de la esencialidad: la mayoría de los hubs son esenciales debido a su participación en complejos biológicos esenciales, un grupo densamente conectado de proteínas con función biológica compartida altamente compuesto por proteínas esenciales.

El objetivo de nuestro trabajo es analizar la relación que existe entre la esencialidad de una dada proteína y determinadas propiedades topológicas de las mismas en el interactoma o red PPI. Esperamos reproducir los resultados de Zotenko en el sentido de llegar a la conclusión de que no son esenciales los nodos de alto grado o los enlaces en la red sino los complejos de proteínas.

2. Resultados y discusión

2.1. Características de las redes analizadas

Para limitar el impacto de posibles sesgos en los resultados reportados, analizamos redes de PPI en un dado tipo de levadura mediante datos relevados por cuatro métodos:

- Afinidad de purificación de proteínas seguido de espectrometría de masas (AP/MS) que identifica interacciones directas o indirectas.
- Análisis de doble-híbrido de la levadura de alta resolución (high-throughput yeast two-hibrid -Y2H-) que identifica interacciones directas o binarias.
- Datos reportados por [5] al analizar cerca de 30,000 trabajos en los que se reportan interacciones entre proteínas en experimentos en pequeña escala (LIT REGULY).
- Datos obtenidos de la literatura desde fuentes desconocidas (LIT).

En la Tabla 1 presentamos las principales propiedades estructurales de las redes consideradas: el número de nodos, el número de enlaces, etc. Como podemos observar se encuentran grandes diferencias entre los parámetros estructurales de las mismas debido a que los métodos de obtención son diferentes; como vimos en el trabajo práctico anterior algunos sobrestiman las interacciones y otros las subestiman. El valor más atípico para el coeficiente de agrupamiento *clustering coefficient* se encuentra en la red Y2H, lo cual indica que esta red está menos interconectada que las demás. De esta información puede verse que la red más esparza (poco densa) es la Y2H y mientras que la menos esparza es la AP-MS.

Tabla 1. Principales propiedades estructurales de las redes consideradas.

Red	**Número de nodos**	**Número de enlaces**	**Grado medio**	**Coeficiente de agrupamiento medio**
AP-MS	1622	9070	5.59	0.55
Y2H	2018	2930	1.45	0.05
LIT	1536	2925	1.90	0.29
LIT-REGULY	3307	11857	3.59	0.26

Por otro lado, se analizó la cantidad de enlaces en común entre dos redes cualesquiera (superposición). Los resultados obtenidos se presentan en la Tabla 2. Como puede observarse, el 98 % de los enlaces de la red LIT se encuentran en la red LIT-REGULY, mientras que solamente el 24 % de los enlaces de la LIT-REGULY se hallan en la red LIT ya que la primera red posee más enlaces y nodos que la última. Por otro lado, las redes Y2H y LIT tienen la misma fracción de enlaces en común (9 %).

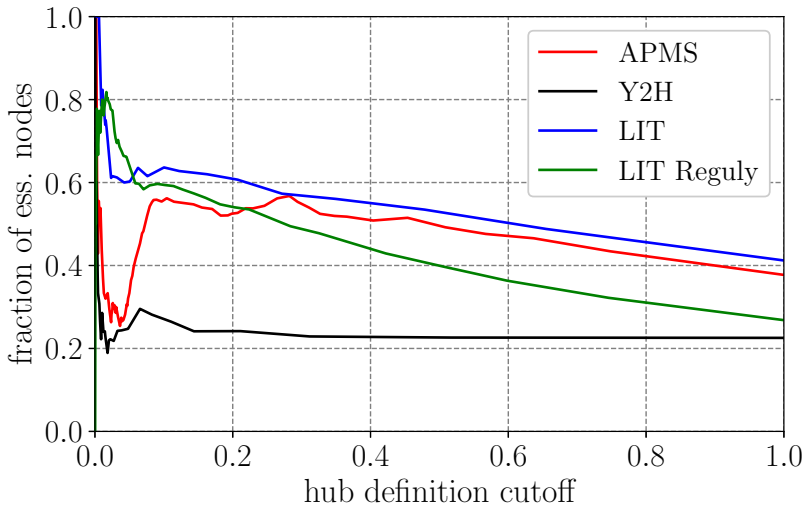
Tabla 2. Fracción de enlaces de una red que están contenidos en otra. Por ejemplo el 29 % de los enlaces de la red AP-MS se encuentran presentes en la Y2H.

AP-MS	0.29	0.14	0.28
0.09	**Y2H**	0.09	0.16
0.44	0.09	**LIT**	0.98
0.21	0.04	0.24	**LIT-REGULY**

En la Figura 1 mostramos la relación entre los hubs y la esencialidad para todas las redes analizadas. En el eje de las abscisas exploramos distintos umbrales para la definición de hub. Es decir, si el valor en este eje es 0.2, debe interpretarse que el 20 % de los nodos de mayor grado fueron considerados hubs. El eje de las ordenadas muestra la fracción de nodos esenciales respecto una dada cantidad de nodos (considerados hubs). Cabe destacar que cuando la definición de hub es 1.0 significa que todos los nodos de la red son considerados hubs. Como se puede observar las cuatro redes convergen a valores diferentes en este punto. Eso significa que, a priori, la fracción de nodos esenciales es distinta para cada red (a pesar de tratarse de la misma especie *Saccharomyces*

cerevisiae).

Figura 1. *Fracción de nodos esenciales respecto de los nodos que fueron considerados hubs (eje y) vs la fracción de nodos de la red que fueron considerados hub (umbral).*



Hay que destacar que para todas las redes, el incremento de la fracción de nodos esenciales respecto del incremento del grado k, se interrumpe para valores altos de k (valores bajos en el eje x). Esto nos sirve para elegir un umbral de modo tal que el 10 % de los nodos sean considerados hubs.

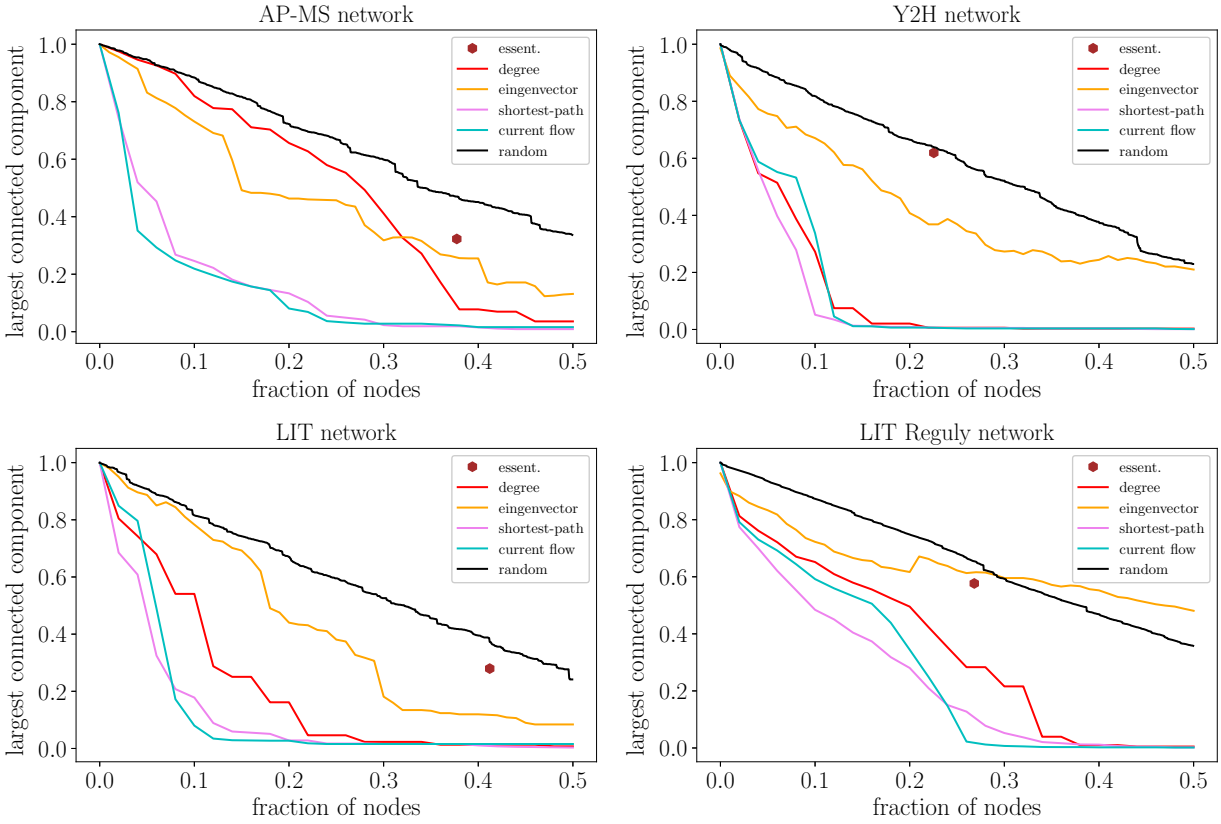
2.2. Análisis de vulnerabilidad

Como mencionamos anteriormente Jeong y sus colaboradores [3] sugirieron que la presencia excesiva de proteínas esenciales en los hubs, puede atribuirse al hecho de que los hubs cumplen un rol fundamental al mantener la conectividad general de la red. Para analizar cómo la remoción de nodos afecta la conectividad de la red, se analizó el tamaño de la componente gigante para diferentes fracciones de nodos removidos.

La Figura 2 muestra cómo la eliminación de los nodos más centrales, la eliminación de nodos al azar y la eliminación de todas las proteínas esenciales afecta la conectividad global de la red. En cuanto a la centralidad, se tuvieron en cuenta cuatro criterios diferentes: centralidad por grado, centralidad por autovector y dos medidas de centralidad por betweenness: shortest-path y current flow. Este análisis se hizo para las cuatro redes estudiadas. Las mediciones se hicieron eliminando nodos con mayor centralidad de forma no recursiva, por ejemplo: la eliminación del 20 % de los nodos más centrales (para cualquier criterio), se hizo a partir de la red original (no se hizo a partir de una red con nodos ya eliminados). El paso utilizado para efectuar el barrido fue 2 % en todos los casos.

La eliminación de nodos con alto Betweenness mostró ser más destructiva que la eliminación de nodos por grado o autovector en todos los casos. El hexágono marrón indica el tamaño de la componente gigante luego de eliminar todos los nodos esenciales. Es interesante notar que la remoción de todos los nodos esenciales no afecta a la conectividad tanto como la eliminación de nodos por cualquier criterio de centralidad. Notar que en el caso de la red Y2H, la eliminación de nodos esenciales afecta tanto como eliminar una fracción equivalente de nodos al azar.

Figura 2. *Fracción de nodos en la componente gigante en función de la fracción de nodos removidos. Las curvas: violeta, celeste, naranja y rojo muestran la vulnerabilidad a ataques contra las proteínas de mayor centralidad según: shortest path betweenness centrality, current flow betweenness centrality, eingenvector centrality y node degree centrality respectivamente. La curva negra representa la remoción de nodos al azar y el hexágono marrón la remoción de nodos esenciales.*



Los gráficos de la Figura 2 sugieren que la eliminación de nodos esenciales no tiene un impacto en la red mayor que el se podría tener con otras formas de ataque. Para validar esta idea, eliminamos n_R nodos no esenciales de las redes originales (donde n_R es la cantidad de nodos esenciales referida a cada red). Procuramos que la distribución de grado de estos nodos eliminados sea lo más parecida posible a la distribución de grado de los nodos esenciales de cada red. Los resultados se muestran en la Tabla 3. En la primer columna se muestra la fracción que representa la componente gigante luego de eliminar n_R nodos esenciales. En la segunda columna se muestra la fracción que

representa la componente gigante luego de eliminar n_R no esenciales con la misma distribución de grado que los n_R nodos esenciales.

Tabla 3. Impacto de la eliminación de proteínas esenciales en comparación con la eliminación de un número equivalente de proteínas no esenciales con la misma distribución de grado.

Red	**Esencial**	**No esencial (Distribución aleatoria)**
AP-MS	0.322	0.344 ± 0.012
Y2H	0.619	0.616 ± 0.013
LIT	0.279	0.418 ± 0.003
LIT-REGULY	0.576	0.581 ± 0.005

Estos resultados muestran que para todas las redes analizadas, eliminar nodos no esenciales al azar afecta tanto a la conectividad de la red como eliminar todos los nodos esenciales (preservando lo más posible la distribución de grado). Ese decir, los hubs esenciales no son más importantes que los hubs no esenciales a la hora de mantener la conectividad global de la red. Por lo tanto, queda descartada la hipótesis de Jeong y colaboradores.

2.3. Esencialidad: módulos biológicos vs interacciones esenciales

He [2] y colaboradores propusieron una explicación para la regla de centralidad-letalidad en términos de las interacciones esenciales entre proteínas. Los autores sugieren que la esencialidad de las proteínas es explicada por un proceso aleatorio. Según este modelo la probabilidad de que una proteína de grado k sea esencial viene dada por:

$$P_E = 1 - (1 - \alpha)^k (1 - \beta) \tag{1}$$

Siendo α la probabilidad de que una interacción entre proteínas sea esencial. β la probabilidad de que una proteína sea esencial y k el grado (cantidad de vecinos). Despejando apropiadamente la ecuación 1 se obtiene la expresión 2 que es lineal en el grado k.

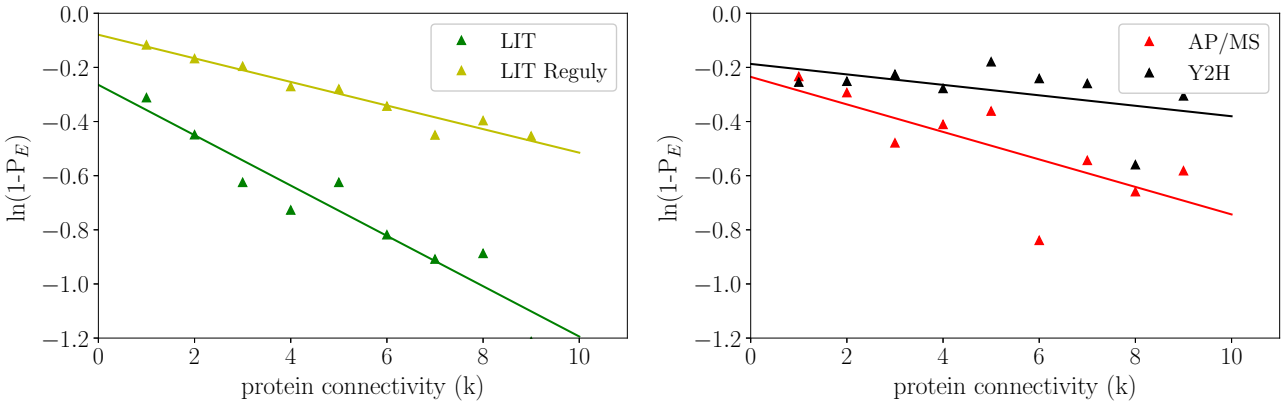
$$log(1 - P_E) = log(1 - \alpha)k + log(1 - \beta) \tag{2}$$

En la Figura 3 se muestra la data correspondiente a la ecuación 2 para cada una de las redes analizadas. Excluimos de este análisis a los nodos de grado mayor a 10. Para cada set de datos hicimos regresiones lineales para hallar los valores de α y β correspondientes a cada red.

Estos ajustes sugieren que la regla de centralidad-letalidad es una consecuencia de que los nodos con mayor grado están involucrados en una mayor cantidad de interacciones proteína-proteína que los nodos de bajo grado. Estos resultados se contraponen a la hipótesis de Jeon y colaboradores quienes proponen que la regla de centralidad-letalidad es producto de la importancia de los hubs con respecto al mantenimiento de la conectividad de la red. Si bien todo esto parece indicar que la

hipótesis de He y colaboradores es suficiente para explicar el fenómeno, más adelante la refutaremos y discutiremos acerca de la hipótesis más aceptada hasta el momento.

Figura 3. *Regresiones lineales de ln(1-Pe) y el grado de las proteínas para todas las redes estudiadas. De la pendiente del ajuste se calculó el valor de* α, mientras que la ordenada al origen sirvió para calcular β



En la Tabla 4 mostramos los resultados del ajuste lineal. La red Y2H mostró la menor bondad de ajuste $R^2 = 0.236$ mientras que la red Lit-Reguly resultó ser la de mejor bondad: $R^2 = 0.947$. Existe una diferencia notable entre la dispersión de las redes LIT y LIT-Reguly respecto a las redes AP-MS e Y2H. Esta diferencia puede deberse al hecho de que la data de literatura es más robusta por contar con mayor una mayor cantidad de datos recopilados y suponemos está mejor curada que las demás redes.

Tabla 4. Parámetros de ajuste lineal para obtener α y β de cada red

Red	**Pendiente**	**Ordenada al origen**	**R^2**	$\alpha(\%)$	$\beta(\%)$
AP-MS	-0.051 ± 0.010	-0.235 ± 0.100	0.536	5.0	20
LIT	-0.093 ± 0.011	-0.265 ± 0.064	0.904	8.8	23
LIT-Reguly	-0.044 ± 0.004	-0.079 ± 0.022	0.947	4.3	7.5
Y2H	-0.019 ± 0.013	-0.187 ± 0.074	0.236	1.8	17

Las suposiciones del modelo de He implican que si dos proteínas no interactúan, la esencialidad de una no depende de la esencialidad de la otra (solo existe correlación cuando si hay interacción entre ellas). Esta idea es válida aun cuando las proteínas comparten vecinos. Pusimos a prueba este 'corolario' del modelo de He, analizando a todos los vecinos no adyacentes con un número mayor o igual a v vecinos.

La Tabla 5 muestra los resultados de este análisis. El número total de pares del mismo tipo refiere a aquellos pares de nodos no adyacentes que comparten v o más vecinos y son ambos esenciales o no esenciales (tercera columna). El número esperado de pares del mismo tipo (cuarta columna) es el número total de pares 'esencial-esencial' y 'no esencial- no esencial' que se esperan

según el modelo de He et al. La siguiente expresión muestra el cálculo que hicimos para obtener este valor:

$$N_E = \sum_{ij} \left[P_E^i P_E^j + (1 - P_E^i)(1 - P_E^j) \right] \tag{3}$$

Donde N_E es el número esperado de pares del mismo tipo, P_E^j la probabilidad de que el nodo j sea esencial según la Ec. 1. La suma se hace sobre todos los pares de nodos que cumplen dos condiciones: no son adyacentes y tienen más de v vecinos en común. Cabe destacar que para cada red, se usaron los valores α y β calculados a partir de las regresiones lineales de la Figura 3.

Como se puede observar en la Tabla 5, en dos de las redes (Y2H y LIT REGULY) podemos rechazar la hipótesis de He ya que el número esperado de pares del mismo tipo es significativamente menor al número de pares del mismo tipo medidos en dichas redes. Por el contrario, no es posible realizar esta afirmación en las redes AP-MS y LIT.

Tabla 5. Número esperado de pares de nodos donde ambas proteínas son esenciales o no esenciales. Solamente se consideraron aquellos nodos no adyacentes con tres o más nodos en común para todas las redes, salvo la Y2H en la cual se consideraron uno o más vecinos en común.

Red	**Número total de pares**	**Número de pares del mismo tipo**	**Numero esperado de pares del mismo tipo**
AP-MS	11613	5924	7736 ± 3863
Y2H	23073	15019	14164 ± 156
LIT	730	393	402 ± 22
LIT-REGULY	10772	6158	5608 ± 250

Finalmente, se analizó cómo varía el número de pares de nodos no adyacentes esperados del mismo tipo (misma esencialidad) según el modelo de He con el número de vecinos en común considerados. En la Tabla 6 se presentan los resultados obtenidos. Como puede observarse, al aumentar el número de vecinos en común disminuye el número esperado de pares del mismo tipo en todas las redes. Este resultado es esperable ya que al aumentar la cantidad de vecinos en común del par de nodos, las condiciones con las que se elijen los mismos son cada mas restrictivas.

Tabla 6. Número esperado de pares de nodos donde ambas proteínas son esenciales o no esenciales para distinto número de vecinos (v) en común.

Red	$v = 2$	$v = 3$	$v = 4$
AP-MS	9852	7736	6413
Y2H	1288	286	98
LIT	988	402	219
LIT-REGULY	22869	5608	2848

3. Conclusión

En este trabajo utilizamos cuatro interactomas para proteínas de la levadura *Saccharomyces cerevisiae* creados a partir de datos obtenidos por cuatro métodos diferentes, con el objetivo de poner a prueba lo aseverado por Zotenko y colaboradores [7], quienes refutaron las propuestas de Jeong [3] en el sentido de que los nodos de alto grado o hubs tienden a ser esenciales porque conservan la conectividad en la red, y por He [2] quien afirma que las proteínas esenciales participan en interacciones esenciales distribuidas uniformemente en la red y que los hubs tienden a ser esenciales simplemente porque están involucrados en un mayor número de interacciones.

En referencia a la hipótesis de Jeong, pudimos refutarla hallando que, para todas las redes analizadas, eliminar nodos no esenciales al azar afecta a la conectividad de la red (medida en términos de la cantidad de nodos de la componente gigante resultante) en la misma forma que eliminar todos los nodos esenciales. Es decir, los hubs esenciales no resultan más importantes que los no esenciales a la hora de mantener la conectividad global de la red. Además observamos que retirar proteínas esenciales es menos disruptivo que remover nodos de acuerdo a cualquiera de los índices de centralidad estudiados.

En cuanto a la hipótesis de He, encontramos para dos de las redes una diferencia estadística significativa entre los números de pares observados del mismo tipo y el esperado bajo el modelo, lo cual nos permitió descartar el modelo de interacciones esenciales en este caso; no pudimos hacer esta afirmación para las dos redes restantes.

Referencias

[1] Alvaro J González y Li Liao. "Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines". En: *BMC Bioinformatics* 11.1 (2010), pág. 537. DOI: 10.1186/1471-2105-11-537. URL: https://doi.org/10.1186/1471-2105-11-537.

[2] Xionglei He y Jianzhi Zhang. "Why Do Hubs Tend to Be Essential in Protein Networks?" En: *PLoS Genetics* 2.6 (2006), e88. DOI: 10.1371/journal.pgen.0020088. URL: https://doi.org/10.1371/journal.pgen.0020088.

[3] H. Jeong y col. "Lethality and centrality in protein networks". En: *Nature* 411.6833 (mayo de 2001), págs. 41-42. DOI: 10.1038/35075138. URL: https://doi.org/10.1038/35075138.

[4] Tony Pawson y Piers Nash. "Protein–protein interactions define specificity in signal transduction". En: *Genes Dev* 14.1 (2000), págs. 1027-1047. DOI: 10.1186/1471-2105-11-537. URL: https://doi.org/10.1186/1471-2105-11-537.

[5] Teresa Reguly y col. En: *Journal of Biology* 5.4 (2006), pág. 11. DOI: 10.1186/jbiol36. URL: https://doi.org/10.1186/jbiol36.

[6] Huiru Zheng, Haiying Wang y David H. Glass. "Integration of Genomic Data for Inferring Protein Complexes from Global Protein–Protein Interaction Networks". En: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.1 (feb. de 2008), págs. 5-16. DOI: 10.1109/tsmcb.2007.908912. URL: https://doi.org/10.1109/tsmcb.2007.908912.

[7] Elena Zotenko y col. "Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality". En: *PLoS Computational Biology* 4.8 (ago. de 2008). Ed. por Burkhard Rost, e1000140. DOI: 10.1371/journal.pcbi.1000140. URL: https://doi.org/10.1371/journal.pcbi.1000140.