Relación entre topología y esencialidad en redes de interacciones entre proteínas de levadura

Lucio García, Lucas Longo, María Luz Vercesi

Redes Complejas
17 de octubre de 2018

1. Introducción

Al analizar redes de proteínas, se buscan encontrar relaciones entre las características biológicas de las mismas y las topológicas. Una proteína es esencial si es indispensable para la supervivencia del organismo. Se busca saber si es posible explicar la importancia biológica, como la esencialidad, a partir de la ubicación en la red. Existen distintas medidas para cuantificar la importancia de un nodo en la red. Algunas de ellas son locales y otras son medidas que tienen en cuenta la vecindad del nodo. Una de las medidas locales es el grado del nodo, y se ha propuesto en la literatura que la esencialidad de una proteína se debe a su función como *hub*, es decir, esencialidad debida a grados altos. A esto se llama la regla de centralidad-letalidad.

Una de las primeras hipótesis sobre esta relación fue de Jeong et al. [1], quien mostró que las proteínas altamente conectadas tienen una probabilidad tres veces mayor de ser esenciales que las proteínas con grado menor. La probabilidad de que la eliminación de una proteína resulte letal está correlacionada con su grado, y una posible explicación es que los *hubs*, nodos con grado mayor que el promedio, mantienen la conectividad entre proteínas de menor grado.

Esto presupone que la esencialidad es una característica individual de cada proteína. He et al. [2] propuso que la esencialidad también puede atribuirse a interacciones entre proteínas, y que por lo tanto, un nodo involucrado en uno de estos enlaces es esencial. De esta manera se manifiesta la esencialidad en *hubs* porque éstos tienen mayor probabilidad de involucrarse en un enlace esencial, por tener mayor grado.

Finalmente, Zotenko et al. [3] rechazó estas hipótesis, y propuso que la esencialidad de una proteína se debe a que la misma forma parte de un Módulo Biológico Complejo Esencial, un grupo de proteínas densamente conectadas que comparten una función biológica.

En este trabajo se buscó replicar los resultados obtenidos por estos autores, utilizando cuatro redes de proteínas: Y2H, AP-MS, LIT y LIT-Reguly. Para ello, se comenzó por caracterizar las redes mediante sus propiedades estructurales y el cálculo del overlap presente entre ellas. También, se analizó la vulnerabilidad de las redes ante la remoción de nodos, mediante distintas medidas de centralidad. Además, se trató de encontrar la correlación entre proteinas esenciales y centralidad de grado para ver si se cumple la regla centralidad-letalidad en ellas. Para analizar el modelo de esencialidad debido a enlaces esenciales, se buscó la probabilidad con la que una proteina resulta esencial debido a estos enlaces (α) y la probabilidad de ser esencial mediante otras razones (β) mediante dos métodos: ajustando una regresión lineal a la probabilidad que un nodo sea esencial(P_E) y realizando 1000 simulaciones de recableo de la red manteniendo su distribución de grado. Por último, se calcularon los pares de proteinas mediante el modelo de Zotenko y se hizo un análisis de asortatividad para los pares de nodos a primeros vecinos. Para todos los procedimientos se quitaron los auto-enlaces de las redes.

2. Características de las redes analizadas

Lo primero que se realizó fue un análisis estructural de las cuatro redes que se consideraron. En el Cuadro 1 (Tabla 1 Zotenko) se pueden observar los tamaños de las redes y algunas otras propiedades básicas que permiten saber cómo se comportan las proteínas estudiadas.

Se puede notar que una de las redes de la literatura (LIT-Reguly) es mucho mas grande que las demás (tiene más nodos), y además también es la que tiene más cantidad de enlaces. Sin embargo, los coeficientes de clustering de ambas son similares. Por el contrario, el coeficiente de clustering de la red Y2H es mucho menor. Esto es así ya que, por construcción, esta red se crea a partir de interacciones uno-a-uno por contacto físico, y por lo tanto es posible que se pierdan muchas interacciones.

	# nodos	# enlaces	Grado medio	Coeficiente de clustering promedio
Y2H	2018	2930	2.90	0.05
AP-MS	1622	9070	11.18	0.55
LIT	1536	2925	3.81	0.29
LIT-Reguly	3307	11858	7.17	0.26

Cuadro 1: Propiedades estructurales de las redes de proteínas estudiadas.

Otra manera de encontrar similitudes y diferencias entre las redes estudiadas es calculando el solapamiento entre los enlaces de las mismas. Nuevamente, las diferencias entre ellas se deben a las distintas técnicas de construcción de las redes. En Cuadro 2 (Tabla 2 Zotenko) se puede ver, por fila, la fracción de enlaces de una red contenida en las otras.

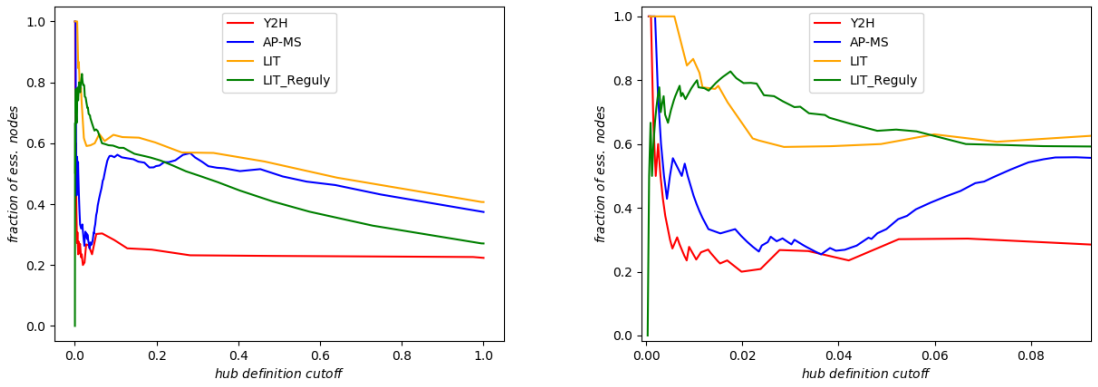
	Y2H	AP-MS	LIT	LIT-Reguly
Y2H	1	0.09	0.09	0.16
AP-MS	0.03	1	0.14	0.28
LIT	0.09	0.46	1	0.98
LIT-Reguly	0.04	0.22	0.25	1

Cuadro 2: Solapamiento entre los enlaces de las redes obsevadas. Cada fila corresponde a una red y muestra la fracción de enlaces de la misma contenidos en las otras. No se consideraron auto-enlaces.

A continuación se analizó la regla de centralidad-letalidad para cada una de las redes. Para eso, se estudió la fracción de nodos esenciales presentes en *hubs*.

Para comenzar, se tuvo que definir un umbral de grado, de forma tal de considerar como *hub* a todo nodo de grado mayor o igual al mismo. El procedimiento consistió en seleccionar un cierto grado de la red y considerar como *hubs* a todos los nodos con grado mayor o igual al elegido. Luego, se contabilizó la cantidad de enlaces esenciales en esos *hubs*. El procedimiento se repitió para cada grado presente en la red.

En la Figura 1 (Figura 1.a de Zotenko) se puede ver que la fracción de nodos esenciales aumenta al disminuir el umbral de cutoff, esto es, al aumentar el grado de los nodos considerados como *hubs*. En la red LIT-Reguly observamos una interrupción de la esencialidad para valores altos de k. Para tener un acercamiento más cuantitativo se calcularon los coeficientes de correlación τ de Kendall y ρ de Spearman los cuales se muestran en el Cuadro 3 (Figura 1.b de Zotenko).



(a) (b)

Figura 1: (a) Relación entre la fracción de nodos esenciales en *hubs* (nodos de grado mayor o igual a k) y el grado k para las distintas redes. (b) Ampliación en la zona de k altos.

	τ **de Kendall**	ρ **de Spearman**
Y2H	0.59 (4.2e-7)	0.72 (7.5 e-7)
AP-MS	-0.17(0.017)	-0.28 (0.007)
LIT	0.82(1.6e-10)	0.93 (1.73e-14)
LIT-Reguly	0.4 (1.0 e-6)	0.5 (9.9 e-6)

Cuadro 3: Coeficientes de correlación de Kendall y Spearman para cuantificar la relación entre el grado y la esencialidad.

Para las redes Y2H, LIT y LIT-Reguly se puede ver que los coeficientes muestran que hay una correlación positiva entre grado y esencialidad. La red AP-MS mostró un valor negativo. Sin embargo, su p-valor es alto en comparación con los de las demás, con lo cual no se puede decir que exista una correlación en esta red. En el resto se logró verificar la regla de centralidad-letalidad.

3. Análisis de vulnerabilidad

Para poder analizar la vulnerabilidad de las redes y la importancia de la esencialidad de las proteínas, se utilizaron distintas medidas de centralidad, entre las cuales se encuentran: centralidad de grado, centralidad de autovectores y centralidad de intermediación de caminos más corto (*DC*, *EC* y *SPBC*, por sus siglas en inglés).

La centralidad por grado corresponde simplemente a la cantidad de enlaces que posee un nodo con los demás. La centralidad de autovector mide la influencia de un nodo en la red. En otras palabras, si posee un valor alto, significa que está conectado a muchos nodos que a su vez están muy conectados. Por úlitmo, la intermediación de caminos más cortos es una medida que cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos.

Debido a que la centralidad de un nodo en la red puede definirse de diversas formas, se decidió hacer un análisis de la vulnerabilidad de la red ante la remoción de los nodos utilizando diversas medidas de centralidad, y de esta forma, tener en cuenta la influencia de los distintos aspectos de la topología de la red. A su vez, para comparar el rol de la esencialidad de las proteínas con estas medidas, se quitaron todos los nodos esenciales de las redes.

En la Figura 2 (Figura 3 de Zotenko), se hizo un análisis de la conectividad de la red. Para eso, se estudió cómo la remoción de nodos, para las distintas médidas de centralidad, afectan al tamaño de la componente gigante de la red.

Aquí se observa que, al igual que en el trabajo de Zotenko, remover nodos por orden de centralidad de grado tiene casi el mismo efecto que quitarlos mediante SPBC en todas las redes, menos en AP-MS. También se puede apreciar que la medida de centralidad de autovectores es menos representativa para la conectividad entre nodos que la de intermediación. Además, para el caso de la red AP-MS, se puede apreciar una diferencia entre la remoción de nodos mediante intermediación y los métodos con altos valores de centralidad. Teniendo en cuenta esto, y la poca diferencia entre el desmembramiento de la red para las figuras 2 *a, c* y *d*, entre DC y SPBC, se puede concluir que el método de medir la centralidad mediante intermediación de caminos más cortos resulta el más efectivo para determinar la importancia en las conexiones de la red.

Si se compara el tamaño de la componente gigante correspondiente a la fracción de nodos quitados, se observa que la remoción total de los nodos esenciales es menos disruptiva que los otros métodos. Si además, se excluye la red AP-MS, incluso dicho tamaño resulta ser similar al obtenido mediante la remoción aleatoria de nodos.

Luego de analizar la vulnerabilidad de cada red para las distintas medidas de centralidad se analizó si, efectivamente, existía alguna diferencia de conectividad entre nodos esenciales y no esenciales. Para ello, se midieron las fracciones de nodos restantes en la componente conectada más grande luego de quitar las proteínas esenciales en un caso, y las no esenciales de grado equivalente de forma aleatoria en el otro caso. Los errores en estos cálculos fueron omitidos debido a que resultaron despreciables.

En el Cuadro 4 (Tabla 3 de Zotenko), se puede observar que para las redes: AP-MS y LIT, no hay una diferencia significativa entre remover nodos esenciales y quitar de forma aleatoria los nodos no esenciales, dando a entender que en principio la centralidad de grado no es un buen determinante de la esencialidad de los nodos. Sin embargo, para las redes: Y2H y LIT-Reguly no ocurre lo mismo, e incluso parecería que las proteínas no esenciales poseen una mayor importancia en la conectividad de la red que las esenciales.

Con estos resultados y teniendo en cuenta las correlaciones de Kendall y Spearman, obtenidas en la seccion anterior, se puede concluir que si bien, las proteínas esenciales están vinculadas a una centralidad de grado mayor que las proteinas no esenciales. éstas últimas, presentan medidas de centralidad de intermediación, más altas que las esenciales.

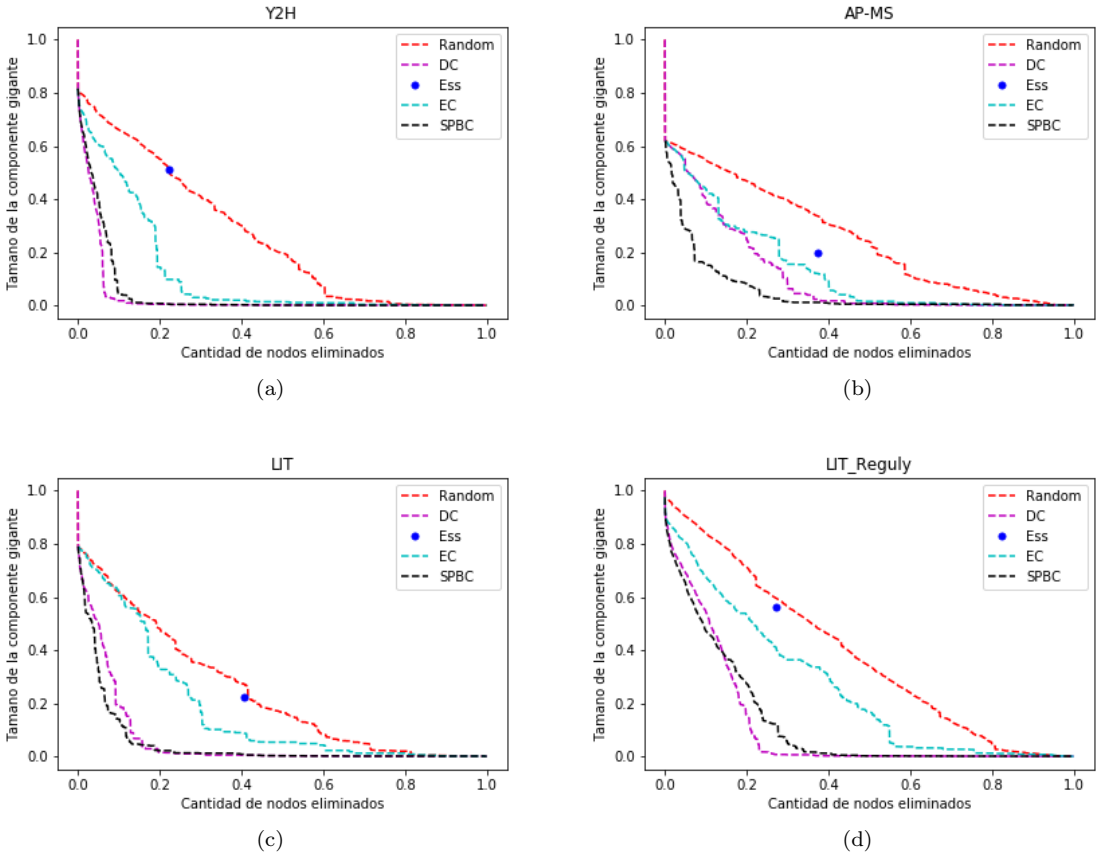


Figura 2: Vulnerabilidad de la red ante la remoción de proteinas, mediante distintos criterios de centralidad

Redes	Fracción restante Nodos esenciales	Fracción restante Nodos no esenciales
Y2H	0,51	0,12
AP-MS	0,20	0,21
LIT	0,22	0,28
LIT-Reguly	0,56	0,25

Cuadro 4: Comparación entre las fracciones de nodos restantes en las componentes gigantes ante la remoción de las proteinas esenciales y las no esenciales para las distintas redes.

4. Esencialidad: Interacciones esenciales y módulos biológicos

Se decidió evaluar el modelo de He[2] el cual intenta explicar la regla de centralidad-letalidad a través de la presencia de interacciones esenciales. Ambas proteínas involucradas en un enlace esencial son esenciales. Para ello, se tomó dicho modelo y se calcularon, para las distintas topologías de nuestras redes, los parámetros de probabilidad de que una interacción sea esencial, α, y probabilidad de que una proteína sea esencial por otras razones, β.

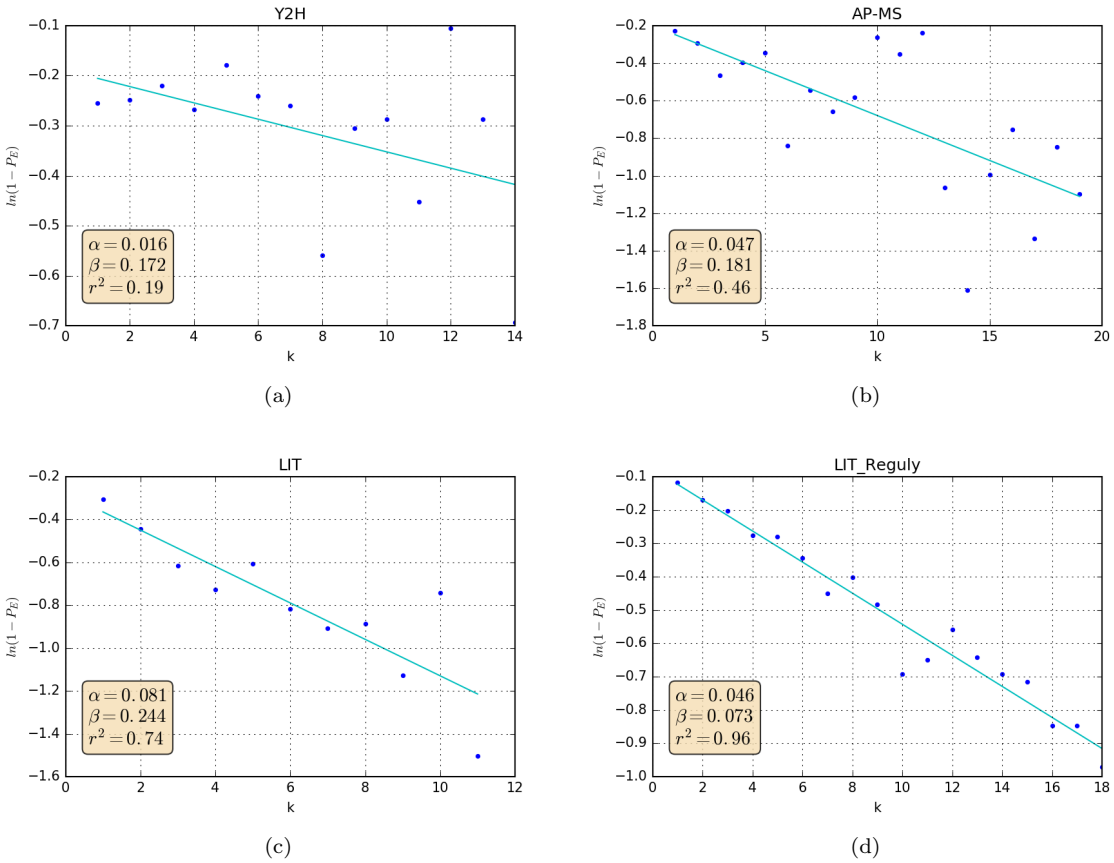
Esto fue realizado de dos maneras. La primera consistió en evaluar la probabilidad de que un nodo sea esencial, la cual, según el modelo de He, se encuentra dada por:

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

$$ln(1-P_E) = kln(1-\alpha) + ln(1-\beta) \tag{2}$$

Esto permitió estimar α y β a través de un ajuste lineal a los valores $ln(1-P_E)$ en función de la conectividad k. La probabilidad P_E fue calculada de manera frecuentista, considerando la fracción de nodos esenciales en la totalidad de nodos de grado k. Para grados altos, la cantidad de nodos no es fue suficiente para hacer un análisis estadístico y por lo tanto los ajustes se realizaron considerando nodos de bajo grado. En la Figura 3 (Figura 2b de He) se observan los ajustes realizados para cada una de las redes.

El segundo método utilizado fue el propuesto por He[2]. Se realizaron 1000 simulaciones de la red y en cada una se generó un recableado de enlaces manteniendo la distribución de grado de los nodos. En cada iteración



(a) (b) (c) (d)

Figura 3: Relación entre la probabilidad (P_E) de que una proteína sea esencial y la conectividad de la proteína (k), y el ajuste lineal correspondiente. Cálculo de las probabilidades α y β (en porcentaje) a partir de la pendiente y la ordenada al origen. (a) $\alpha = 1{,}62 \pm 0{,}95$ y $\beta = 17{,}22 \pm 6{,}78$ (b) $\alpha = 4{,}68 \pm 1{,}21$ y $\beta = 18{,}10 \pm 11{,}87$ (c) $\alpha = 8{,}15 \pm 1{,}54$ y $\beta = 24{,}37 \pm 8{,}62$ (d) $\alpha = 4{,}56 \pm 0{,}24$ y $\beta = 7{,}33 \pm 2{,}52$

se contaron la cantidad de *IBEPs* (enlaces entre proteínas esenciales). Repitiendo el proceso, se obtuvo la distribución de *IBEPs* en función del grado como se muestra en la Figura 4. La línea vertical roja muestra el valor de *IBEPs* observados en la red real y que llamaremos E_{ess}.

El valor de α se calculó como la fracción de enlaces de la diferencia entre la cantidad de *IBEPs* observados en la red real E_{ess} y la media de la distribución de *IBEPs* m respecto a la cantidad de enlaces totales E de la red:

$$\alpha = \frac{E_{ess} - m}{E} \tag{3}$$

Para estimar β también se utilizó el método descripto por He[2], el cual consiste en realizar un etiquetado aleatorio de las N_{ess} proteínas esenciales de la red. Este método consitió en dos pasos. Primero se eliminó toda información de esencialidad de la red y luego se asignaron de forma aleatoria un número de enlaces de carácter esencial (*PPIs*) equivalentes a $E_{ess} - m$. Este proceso dió como resultado una cierta cantidad de proteínas esenciales a causa del factor enlace esencial que se denominó n_{PPI}. Luego, se marcaron las proteínas esenciales de forma aleatoria hasta alcanzar las N_{ess} proteínas esenciales de la red real. Se calculó β como la fracción de nodos adicionada:

$$\beta = \frac{N_{ess} - n_{PPI}}{N} \tag{4}$$

En el Cuadro 5 se resumieron los parámetros obtenidos por ambos métodos. Se puede ver que no hay una diferencia significativa y que ambos valores se superponen al tener en cuenta el error de los parámetros.

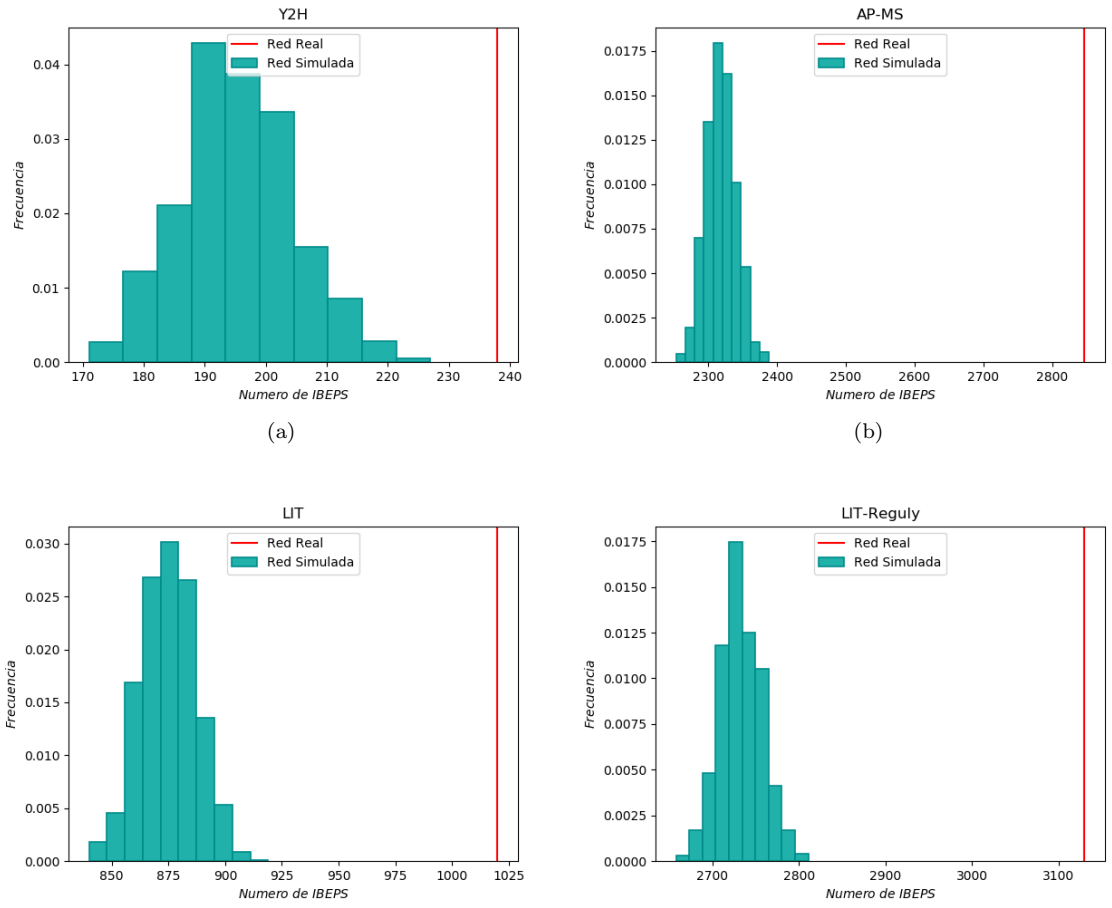


Figura 4: Frecuencia del número de enlaces entre proteínas esenciales para 1000 recableados de la red. La línea roja vertical es el valor de enlaces entre proteínas esenciales observados en la red real.

	Método 1		Método 2	
Red	$\alpha(\%)$	$\beta(\%)$	α (%)	$\beta(\%)$
Y2H	1.62 ± 0.95	17.2 ± 7	1.6 ± 0.3	18.6 ± 0.3
AP-MS	4.7 ± 1.2	18 ± 12	5.8 ± 0.2	3.1 ± 1
LIT	8.1 ± 2	24.3 ± 9	5.1 ± 0.4	25 ± 1
LIT-Reguly	4.6 ± 0.2	7.3 ± 2.5	3.5 ± 0.2	9.4 ± 0.9

Cuadro 5: Parámetros α y β del modelo de interacciones esenciales según dos métodos.

Calcular la probabilidad de que una proteína sea esencial mediante la ecuación (1) implica que se consideran independientes las probabilidades de distintas proteínas, es decir, que en el modelo se supone que si dos proteínas no interactúan, la esencialidad de una proteína, no depende de la esencialidad de la otra. Esto también se debe extender a pares de proteínas no adyacentes que comparten primeros vecinos.

Una forma de verificar si esta hipótesis se cumple, es contar en cada red la cantidad de pares de proteínas no adyacentes del mismo tipo (ambas esenciales o ambas no esenciales) que comparten vecinos, y compararlos con lo esperado según el modelo de He. En el Cuadro 6 (Tabla 5 de Zotenko) se pueden ver los resultados de estos cálculos.

	# nodos	# pares totales	# pares del mismo tipo	# pares esperados del mismo tipo
Y2H	2018	23073	15087 (65 %)	13693 ±113
AP-MS	1622	11613	5907 (51 %)	7482 ±84
LIT	1536	730	389 (53 %)	396 ±11
LIT-Reguly	3307	10777	6187 (57 %)	5660 ±64

Cuadro 6: Diferencia entre los pares observados y esperados donde ambas proteínas son del mismo tipo (esencial o no esencial). El número total de pares se refiere a los pares de proteínas no adyacentes con tres o más vecinos en común, exceptuando a la red Y2H en la cual se consideran con uno o más vecinos en común.

Se puede notar que en dos de las redes, Y2H y LIT-Reguly, la cantidad de pares del mismo tipo es mayor que lo esperado según el modelo. En la red LIT no se puede afirmar nada ya que los números son del mismo orden (el valor medido entra en el intervalo de confianza del esperado). La red AP-MS tiene menos pares del mismo tipo que los esperados según el modelo.

Zotenko utiliza sus resultados para refutar la hipótesis de independencia de probabilidad de ser esencial entre las proteínas. A partir de este resultado afirma que las proteínas esenciales tienden a interactuar con esenciales, y lo mismo para las no esenciales, y propone la existencia de módulos biológicos complejos esenciales, grupos de proteínas esenciales que interactuan entre sí y comparten alguna función biológica.

Ya que no se puede realizar una afirmación similar a partir de los resultados que se obtienen con estas redes, se puede calcular la asortatividad según esencialidad para saber si las proteínas se relacionan con otras del mismo tipo más que lo que se esperaría al azar. Es importante notar que esto no es equivalente a lo calculado por Zotenko, ya que al calcular asortatividad se consideran nodos adyacentes.

En el Cuadro **??** se pueden ver los coeficientes de asortatividad normalizados. Una red completamente asortativa presenta un coeficiente $C = 1$, una red azarosa $C = 0$ y una disortativa, $C = -1$. Se puede notar que en la red AP-MS y ambas redes de literatura hay asortatividad por esencialidad, ya que el coeficiente es un número $0 < C < 1$. Sin embargo, no se puede afirmar lo mismo para la red Y2H. Como ya se vio, el método de construcción de la red Y2H hace que algunos enlaces se pierdan por ser interacciones de contacto físico uno a uno, por lo que no todos los enlaces de los módulos biológicos aparecen en esta red.

	Y2H	AP-MS	LIT	LIT-Reguly
Asortatividad	0.07	0.35	0.31	0.21

Cuadro 7: Coeficientes de asortatividad por esencialidad normalizados para cada red.

5. Conclusiones

- Se pudo determinar la existencia de una correlación entre la centralidad de grado y la esencialidad de las proteinas para las redes Y2H, LIT y LIT-Reguly.
- Para el análisis de vulnerabilidad, se puede afirmar que SPBC, resulta ser la medida de centralidad más eficiente para clasificar los nodos y analizar la conectividad de las redes.
- En ninguna de las redes se observó una diferencia relevante en el tamaño de la componente gigante entre la remocion de las proteínas esenciales y la remoción de nodos aleatorios. Incluso, analizando particularmente la red AP-MS en el Cuadro 4 se puede apreciar que no hay diferencia entre la fracción de nodos en la componente gigante, quitando nodos esenciales o no esenciales.
- Una posible explicación para los valores de dicho cuadro es que tanto para: Y2H y LIT-Reguly, las proteínas no esenciales tengan medidas de centralidad de intermediación más altas que las esenciales. Mientras que para las redes: AP-MS y LIT, dichas medidas son similares.
- Se calcularon las probabilidades α y β según el modelo de He, considerando que las probabilidades de que dos proteínas que no interactúan sean esenciales son independientes entre sí. Sin embargo, comparando la cantidad de pares de proteínas del mismo tipo con vecinos en común, que son esperados utilizando el modelo de He con estos valores de α y β, se puede ver que para algunas de las redes(Y2H y LIT-Reguly) la cantidad observada es mayor.
- Para las redes en donde el modelo de He no se pudo rechazar(AP-MS y LIT), un posterior análisis de asortatividad por esencialidad junto con el resultado anterior sirvió para ver que las proteínas tienden a agruparse con otras del mismo tipo.

Referencias

[1] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai. ***Lethality and centrality in protein networks***, 2001.

[2] Xionglei He, Jianzhi Zhang. ***Why Do Hubs Tend to be Essential in Protein Newtworks?***, 2006.

[3] Elena Zotenko, Julian Mestre, Dianne P. O'Leary, Teresa M. Przytycka. ***Why Do Hubs in the Yeast Protein Interaction Network Tend to be Essential: Reexamining the Connection between the Network Topology and Essentiality***, 2008.