Letalidad y centralidad en redes de proteínas

CICCHINI Tomás, SZISCHIK Candela, VIDAL María Sol

Resumen

En este trabajo se estudia la regla de *centralidad-letalidad* para cuatro redes de interacción de proteínas correspondientes a la levadura. Dicha regla establece que las proteínas altamente conectadas con otras tienden a ser esenciales para el funcionamiento de un microorganismo. Se estudiará la validez de esta regla para las redes consideradas. Para analizar la causa subyacente de dicha regla, se analizará la relación entre algunas propiedades topológicas de una proteína en una red de interacción y su esencialidad para la vida de la levadura.

Introducción

Las proteínas se identifican tradicionalmente como los ladrillos fundamentales de las células y microorganismos. A su vez, el rol de las mismas se puede expandir a un elemento de una red de interacción entre proteínas, que puede tener asociada alguna funcionalidad contextual o celular. En una red de interacción entre proteínas, estas últimas son los nodos y las interacciones, los enlaces.

En general, a partir del análisis de la topología de una red de interacciones entre proteínas se busca inferir propiedades biológicas de cada proteína, como por ejemplo su funcionalidad biológica, su nivel de importancia para la vida del microorganismo involucrado, entre otras. Una forma de ver la esencialidad de una dada proteína experimentalmente es analizar las consecuencias fenomenotípicas de remover un único gen y observar cuánto es afectado por la posición en la red de la proteína asociada al gen. La importancia o esencialidad de un nodo se puede medir por la magnitud de los cambios causados por la remoción del nodo en la función de la red o en su aptitud biológica. Por ejemplo, el estudio de la supresión de un gen de todo el genoma muestra que una pequeña fracción de genes en el genoma son indispensables para la supervivencia o reproducción de un organismo: estos genes (o sus proteínas asociadas) se consideran **esenciales**.

Una pregunta interesante, que se tratará en este trabajo, es si la esencialidad de una proteína puede ser explicada en términos de su ubicación en una red de interacciones. Una de las primeras conexiones entre estas dos, en el contexto de redes de interacción de proteínas, es la llamada *regla de letalidad-centralidad*. Ésta fue observada por Jeong *et al.*[1], quien demostró que las proteínas con alto grado o *hubs* en una red de interacción tienen más chances de ser esenciales que las demás. La correlación entre conectividad y esencialidad fue confirmada por otros estudios, pero las razones de esta correlación no están claras. En particular, ¿es el grado u otra propiedad topológica más global la que tienen las proteínas esenciales en una red de interacción?

Jeong[1] sugirió que la razón por la que los *hubs* tienden a ser esenciales estaba relacionada con su rol central en mantener la conectividad global de la red. Lo que observó es que si se eliminaban *hubs* de la red,

en general, aumentaba el largo de los caminos entre nodos no adyacentes, es decir que el *hub* funcionaba como un atajo entre pares de nodos. Bajo la hipótesis de que la funcionalidad de un organismo depende de la conectividad entre sus partes, los *hubs* son predominantemente esenciales porque juegan un rol central para mantener esta conectividad.

He [2], en cambio, propuso que la regla de centralidad letalidad no se debía a una propiedad global como la conectividad, sino a propiedades locales. Él introdujo la idea de que las proteínas no eran esenciales sino que los enlaces lo eran. Las proteínas son esenciales por el hecho de estar involucradas en enlaces esenciales. Bajo esta hipótesis, los *hubs* son predominantemente esenciales ya que están involucrados en más interacciones que el resto, y por lo tanto tienen más chances de estar contenidos en un enlace esencial.

Zotenko[3] sugirió que la razón de la correlación conectividad-esencialidad no se encontraba ni a nivel global ni local, sino en una escala intermedia o mesoscópica. Zotenko y sus colegas propusieron que lo que era esencial para el ser vivo eran los complejos de proteínas, es decir, subgrupos de proteínas densamente conectados. Existen complejos esenciales y complejos no esenciales. Remover una proteína de un *cluster* esencial es letal para el organismo. Bajo esta hipótesis, los *hubs* tienen más chances de participar en complejos esenciales por su alta conectividad.

Es evidente que la razón subyacente de la regla centralidad-letalidad no es clara. Es por eso que en este trabajo se analizará primero la validez de esta regla para redes de interacciones de la levadura estudiadas, y se buscará evaluar distintas hipótesis sobre el porqué de este comportamiento.

Características de las redes analizadas

Como fuente de data de las interacciones entre las proteínas de la levadura se trabajó con cuatro redes distintas: Y2H, AP-MS,LIT y LIT-REGULY. Cada una de estas redes fue relevada con un método diferente. Traducir la información de los complejos de proteínas en interacciones -binarias- está determinado por la técnica experimental considerada y por el método de relevamiento dado, lo que significa definir qué es una interacción entre dos proteínas. Una de las técnicas consiste en utilizar un anticuerpo que capta una determinada proteína, permitiendo aislar un complejo de proteínas que la contenga. La red AP-MS fue relevada usando dicho método. Ésta red será llamada a lo largo del trabajo como red *Proteínas*.

Otra técnica experimental es el sistema de doble híbrido que analiza la interacción de a pares de proteínas. La red Y2H fue relevada usando este método. Ésta será referida como red *Binarias*.

Por último, las redes LIT y LIT-REGULY fueron construidas a partir de interacciones reportadas en la literatura. Éstas serán referidas a lo largo del trabajo como red *Literatura* y *Literatura Reg.*, respectivamente.

En términos generales, las redes *Proteínas* y *Binarias* provienen de experimentos a 'gran escala'. Mientras que las redes *Literatura* y *Literatura Reg.* provienen de experimentos a 'pequeña escala' (experimentos dirigidos centrados en pocas proteínas).

En primer lugar, se analizaron las características estructurales de las cuatro redes estudiadas. En la Tabla (1), se muestra el número de nodos, el número de enlaces, el grado medio y el coeficiente de *clustering* medio para cada red.

Red	Número de Nodos	Número de Enlaces	Grado medio	Clustering promedio
Proteinas	1622	9070	[11.183723797780518]	0.554636
Binarias	2018	2930	[2.9038652130822595]	0.046194
Literatura	1536	2925	[3.80859375]	0.292492
Literatura Reg.	3309	11859	[7.167724388032639]	0.260976

Figura 1: *Características estructurales de las redes de interacción de proteínas*

Se observa que el *clustering* promedio en la red *Binarias* es un orden de magnitud menor que el del resto de las redes. Esto se puede deber al método de relevamiento de dicha red de interacción, el cual reporta interacciones binarias entre proteínas y produce *clusters* poco densos.

A continuación se estudió el *overlap* entre las redes estudiadas, es decir, la fracción de enlaces compartidos entre dos redes. La Tabla (2) muestra lo obtenido. La misma no resulta simétrica ya que se normalizó el *overlap* por el número total de enlaces de cada red. Por ejemplo, la primera columna está normalizada por el número total de enlaces de la red *Proteínas*.

Proteinas	0.0887372	0.443761	0.212497
0.0286659	Binarias	0.0888889	0.0403913
0.143109	0.0887372	Literatura	0.241167
0.277839	0.163481	0.977778	Literatura Reg.

Figura 2: ***Overlap entre las redes analizadas.*** *Cada columna de la tabla corresponde a una red y muestra la fracción de sus enlaces que están contenidos en las otras redes testeadas. Por ejemplo el* $8,87\,\%$ *de los enlaces de la red Binarias está contenido en la red Proteínas.*

Dadas las diferencias en las técnicas experimentales usadas para construir estas redes y el hecho de que los enlaces en la red *Proteínas* corresponden a pertenencia a complejos multiproteicos, en la *red Binarias* a interacciones de contacto físico y en las redes *Literatura* y *Literatura Reg.* a una mezcla de estas dos cosas, no es sorprendente que las redes difieran significativamente en términos de la topología expresado en el grado medio, *clustering* promedio, *overlap*. La red más alejada es la *Binarias*. Por ejemplo, como se puede ver en la Tabla (2), los valores de la columna correspondiente a la red *Binarias* son, en general, de un orden menor que el resto de los valores. Esto se puede deber al método binario de relevamiento de interacciones entre proteínas. Esto hace que muchas de las interacciones presentes en las otras redes no se encuentren en la red *Binarias*.

De ahora en más, se analizará la componente gigante de cada red de interacciones de proteínas.

Se analizó si efectivamente la regla centralidad-letalidad está presente en las redes estudiadas. En otras palabras lo que se quiere ver es que efectivamente existe una correlación positiva entre el grado-conectividad- y la esencialidad de una dada proteína de la red de interacciones. Para probarlo se varió la definición de *hub*, es decir, el grado a partir del cual se considera a un nodo altamente conectado (*hub definition cutoff*), y se reportó para cada grado de corte la fracción del número de nodos esenciales que son

hubs sobre el número total de *hubs* a partir de una lista conocida de proteínas reportadas como esenciales para la vida de la levadura.

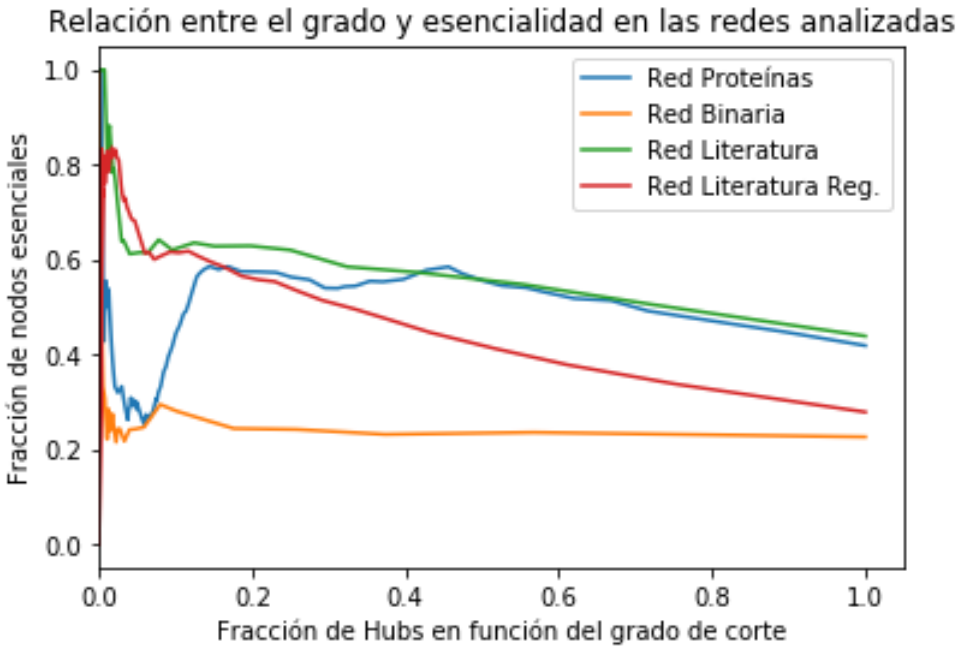

Figura 3: *Para cada red estudiada se muestra la fracción de nodos esenciales entre los hubs. El eje horizontal muestra la fracción de los nodos totales de la red que fueron designados hubs*

Lo que se observa en la Figura (3) es que efectivamente existe una correlación entre la conectividad y la esencialidad de una proteína en la red de interacciones. A medida que aumenta el grado de corte para definir a los *hubs*, la fracción de nodos esenciales entre los nodos más conectados aumenta. En otras palabras, cuando la fracción de *hubs* es cercana a cero (grado de corte muy alto, muy pocos *hubs*), la fracción de nodos esenciales entre ellos alcanza su máximo, llegando al valor límite cercano a uno, donde todos los *hubs* son esenciales, para la red *Literatura*. A medida que se disminuye el grado de corte, el número de *hubs* aumenta, pero la fracción de nodos esenciales entre *hubs* disminuye.

Se puede apreciar que en contraste con las otras redes, la red *Binarias* exhibe solamente una correlación débil entre grado y esencialidad.

Aunque se haya visto que existe una conexión entre grado y esencialidad de una proteína de la redes estudiadas, la razón de esta correlación no es clara. El propósito de este trabajo de aquí en más va a ser entender la causa subyacente de la regla de centralidad-letalidad. El análisis se basará fuertemente en los trabajos de Jeong[1], He[2] y Zotenko[3].

Análisis de la vulnerabilidad

En esta sección se analizará la vulnerabilidad de las 4 redes estudiadas al remover nodos según un determinado criterio. Este criterio lo relacionaremos, por ejemplo, con medidas de centralidad.

Un índice de centralidad asigna un valor de centralidad a cada nodo en la red que cuantifica su prominencia/importancia topológica relativa dentro de la red. Esta noción de importancia dependerá de cómo se defina la centralidad y de qué parámetros se consideren. En otras palabras, dicha prominencia

topológica se puede definir de varias maneras, de forma tal que cada medida de centralidad enfatiza en diferentes aspectos de la topología de la red.

A grandes rasgos, se pueden dividir a los índices de centralidad en dos grandes grupos: locales y globales. En una medida local, el valor de centralidad de un nodo es mayormente influenciado por las características de su entorno local. En este trabajo se utilizaron dos índices de centralidad local: centralidad por grado y por autovector. En la de grado, el valor de centralidad de un nodo es igual a su grado. La medida dada por el autovector es aquella que considera como nodos centrales a aquellos nodos con vecinos centrales por lo que resulta un problema de autovectores.

Por otro lado, un índice de centralidad global se basa en el rol del nodo en mantener la conectividad entre otros pares de nodos de la red, es decir, el rol en preservar la conectividad global de la red. En este trabajo se utilizaron las medidas de centralidad global de camino más corto y de corriente de flujo. En el primer índice, la centralidad de un nodo es proporcional a la fracción de caminos más cortos que pasan por él, promediado sobre todos los pares de nodos de la red. La medida de corriente de flujo generaliza la de camino más corto incluyendo caminos adicionales, no sólo los más cortos.

Es interesante comparar la eficiencia de los *hubs* (nodos con alta medida de centralidad local) con la de nodos con altas medidas de centralidad global a la hora de conectar la red. Una forma de hacer esto es, para cada medida de centralidad, remover los nodos según un criterio de centralidad decreciente y, en paralelo monitorear la disminución del tamaño de la componente gigante. La Figura (4) muestra, para cada red de interacción, cómo afecta a la conectividad de la red la remoción de los nodos más centrales según distintas medidas. También se muestra el mismo efecto al remover nodos al azar y al eliminar todas las proteínas esenciales al mismo tiempo.

De la Figura (4) se observa que, en general, eliminar nodos según medidas de centralidad globales -camino más corto y corriente de flujo en este caso- es mucho más disruptivo que remover según medidas locales (grado y autovector). Es interesante ver que en la red *Binarias* la centralidad por grado resulta ser la más efectiva para desconectar la componente gigante. En las redes *Literatura* y *Literatura Reg.* la centralidad por grado se encuentra entre las más efectivas y en la red *Proteínas* no resulta tan eficiente.

A continuación, se examinó si el poder disruptivo de los *hubs* viene principalmente de los *hubs* esenciales. En la Figura (4) se puede ver que sacar todas las proteínas esenciales (los puntos en negro) está entre las dos formas menos eficientes de desarmar las cuatro redes: para la red *Proteínas* es la tercera menos eficiente, para la *Binarias* es la de menor eficiencia, para la red *Literatura* es igual de eficaz que remover nodos al azar, y para la *Literatura Reg.* es sólo más eficiente que eliminar al azar.

En segundo lugar, se comparó el poder disruptivo de remover todas las proteínas esenciales con el de remover de forma aleatoria un número equivalente de nodos no esenciales con la misma -o más cercana- distribución de grado que los esenciales. Para cuantificarlo se muestran en la Tabla (5) la fracción de nodos en la componente gigante luego de remover todas las proteínas esenciales en la primera columna, y en la segunda la fracción de nodos en la componente gigante al remover un set de proteínas no esenciales al azar pero manteniendo la distribución de grado de las esenciales. El criterio para esta selección fue elegir, cuando no existiese ya nodos del próximo grado igual al esencial a sacar el de grado menor más cercano. Esta elección fue pensada en función de la noción de que hay un número alto de nodos esenciales con grados altos, por lo que resulta más sencillo encontrar grados que sean menores.

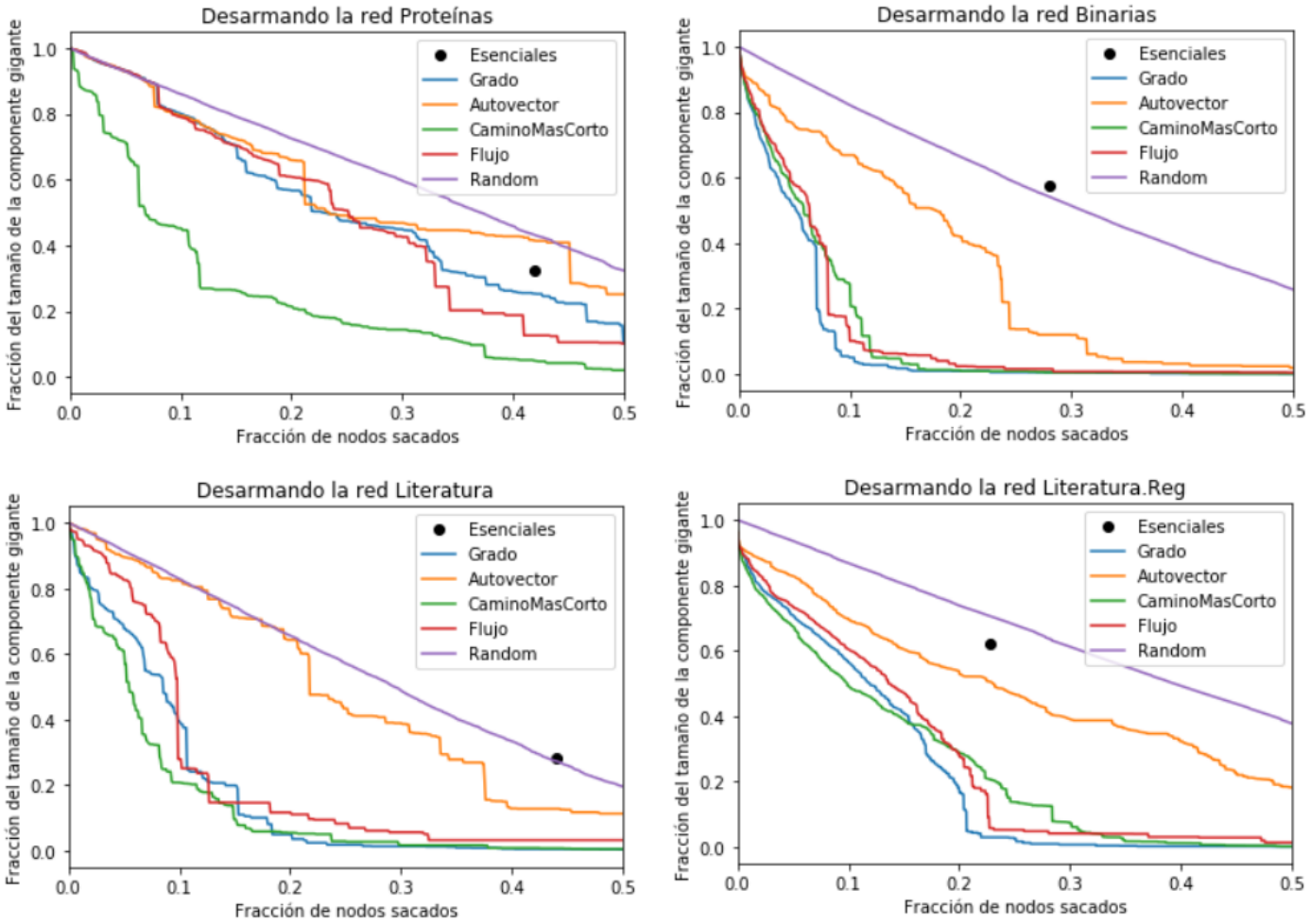

Figura 4: ***Vulnerabilidad a la remoción de las proteínas más centrales.*** *Para cada red, el impacto de remover nodos se cuantifica por la fracción de nodos en la componente gigante. Se muestra una curva por cada medida de centralidad que indica la fracción de nodos en la componente gigante como función de los nodos mas centrales removidos. También se exhibe el impacto de remover nodos en un orden al azar y el tamaño de la componente gigante cuando se remueven todos los nodos esenciales a la vez.*

Red	Esenciales	No esenciales
Proteinas	0.323	0.493±0.099
Binaria	0.624	0.622±0.022
Literatura	0.281	0.249±0.269
Literatura Reg	0.575	0.602±0.021

Figura 5: ***El impacto de remover un set de proteínas es cuantificado por la fracción de nodos en la componente gigante***. *Para cada red se muestra el efecto de remover las proteínas esenciales y el de eliminar un número equivalente de proteínas no esenciales de forma aleatoria con la misma distribución de grado. El error reportado en la columna* ***No esenciales*** *es el asociado a la desviación estándar de las iteraciones realizadas pesado por el valor medio calculado.*

Como se observa en la Tabla (5), remover los nodos esenciales no es más disruptivo que remover un número equivalente de nodos no esenciales al azar manteniendo la distribución de grado. Por lo tanto, se concluye que aunque para las redes *Binarias*, *Literatura* y *Literatura Reg.*, remover *hubs* era uno de los métodos más eficientes para desconectar la componente gigante, esto no está relacionado con la esencialidad de esos nodos. Al contrario,las proteínas esenciales son indistinguibles de un número equivalente de nodos no esenciales *random* con la misma distribución de grado en términos del poder disruptivo de las mismas. En otras palabras, los *hubs* esenciales no son más importantes para mantener la conectividad de la red que los *hubs* no esenciales. Por lo tanto, esto muestra que la hipótesis de Jeong no resulta válida, pues se ve que la esencialidad no está relacionada con el rol de la proteína para mantener la conectividad global en la red.

Esencialidad: módulos biológicos vs interacciones esenciales

En este apartado se busca aportar a la discusión sobre si la esencialidad de las proteínas está dada por los enlaces en los cuales participan [2] o bien si está determinada por una estructura mesoscópica, asociada a la idea de complejos esenciales para una célula [3].

Con la voluntad de probar si son los enlaces los que definen la esencialidad de una proteína, se decidió seguir el camino propuesto por He[2]. Sean α la probabilidad de que una proteína pertenezca a un enlace esencial y β la probabilidad de que una proteína sea esencial por otra razón. Se tiene que, siendo ambas probabilidades independientes, $P_{NE} = (1-\beta)(1-\alpha)^k$ es la probabilidad de que una proteína con grado k no sea esencial. Por lo tanto,

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

es la probabilidad de que una proteína sea esencial dado su grado.

Teniendo en cuenta esta expresión, se analizaron las redes estudiadas de la siguiente manera. Se contabilizaron el total de proteínas esenciales con cierto grado k y se las dividió por el número total de proteínas con ese mismo grado. De esta manera se definió de forma frecuentista la probabilidad de ser esencial en función del grado para cada red. Con el objetivo de analizar la dependencia de dicha probabilidad con el grado, se trabajó con el $ln(1-P_E)$ para poder realizar un ajuste lineal de los datos obtenidos. Dicho análisis puede verse en la Figura (6).

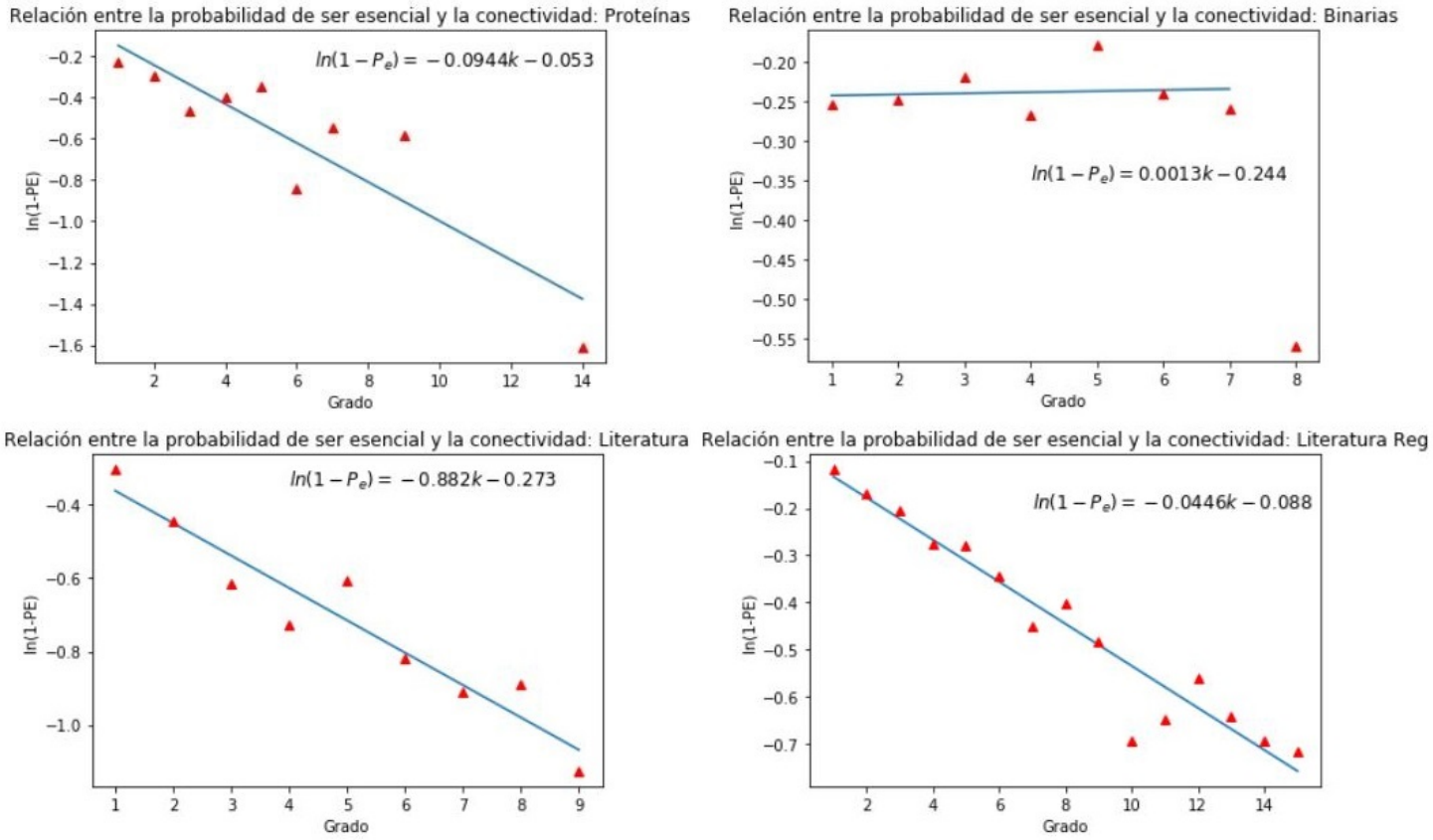

Figura 6: ***Relación entre la probabilidad de ser esencial en función del grado.*** *Sólo se tuvieron en cuenta aquellos grados para los cuales el número total de proteínas con dicho grado superaban un umbral de 30. De esta manera se tomaron puntos de relevancia estadística para cada una de las redes.*

Si bien la tendencia lineal de la dependencia entre $ln(1 - P_E)$ y k no es del todo apreciable para las cuatro redes, se utilizaron los parámetros obtenidos del ajuste para estimar α y β para cada red.

De las hipótesis del modelo de He se sigue que si dos proteínas no interactúan, entonces la esencialidad de una proteína del par no depende de la esencialidad de la otra proteína (se puede ver con P_E). Más aún, esta independencia se debería ver cuando el par de proteínas comparte vecinos. Para comprobar si esto es así efectivamente en las redes reales, se calculó el número total de pares de proteínas no adyacentes con tres o más vecinos en común (salvo para el caso de la red *Binarias* donde se pidió que compartieran más de uno) del mismo tipo, es decir, las dos proteínas del par esenciales o las dos no esenciales. Este número de pares del mismo tipo en la red real se comparó con el número esperado bajo el modelo de He con el ajuste lineal realizado anteriormente.

Para recrear el número de pares del mismo tipo a partir del modelo de He, se generaron de manera iterativa listas de esencialidad, determinadas por los valores α y β, con las cuales se calcularon los pares del mismo tipo. Finalmente, este valor se comparó con la cantidad total de pares que efectivamente presentan las redes estudiadas. Los resultados pueden observarse en la Tabla (7).

Red	Número total de pares	Número de pares del mismo tipo	Esperados por el ajuste lineal
Proteinas	11569	5875	5125±0.04
Binaria	522	352	203±0.1
Literatura	718	383	223±0.1
Literatura Reg	10777	6187	3152±0.01

Figura 7: ***Número total de pares de proteínas no adyacentes con 3 o más vecinos en común.*** *Los nodos en el par son del mismo tipo si ambos son esenciales o no esenciales. Se reporta el número observado y esperado por el ajuste lineal de pares donde las dos proteínas son del mismo tipo. El error reportado es el asociado a la desviación estándar de las iteraciones realizadas pesado por el valor medio calculado.*

Puesto que el número estimado de enlaces es en todos los casos inferior al valor real para cada red, el hecho de asignarle la propiedad de esencialidad a los enlaces y a su vez, asignar cierta probabilidad a que una proteína sea esencial por otras razones, parece no reproducir los resultados obtenidos. En otras palabras, el modelo de He no captura la correlación observada que hay en la esencialidad de una proteína del par con la esencialidad de la otra. De esta manera, la idea de que hay mayor cantidad de pares del mismo tipo, puede llevar a pensar que la razón de la esencialidad se halla no en propiedades de enlaces, sino que ésta descansa en estructuras más grandes, como lo son los complejos proteicos esenciales, los cuales o son esenciales o no lo son. Esta visión de complejos esenciales/no esenciales sí refleja la correlación existe en la esencialidad de proteínas de un par.

Referencias

[1]. Jeong H.,(2001) *Lethality and centrality in protein network.* Nature vol 411.

[2]. He X.,Zhang J. (2006) *Why do hubs tend to be Essential in Protein Networks?*, PloS Genet 2(6).

[3]. Zotenko E., (2008) *Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality.* PloS Comput Biol 4(8): e1000140