Trabajo Práctico Computacional 02: Conceptos básicos

Emanuel Ferreyra, Bruno Kaufman, Ariel Salgado

17 de octubre de 2018

1. Introducción

Si vamos a hablar de letalidad y centralidad en redes de proteínas es menester mencionar el rol fundamental que estas tienen en la biología orgánica o celular. Además de constituir alrededor del 80 % del material viviente deshidratado de la célula (protoplasma), son encargadas de regularizar el organismo como parte constitutiva de las enzimas y de defenderlo cuando ocupan el rol de anticuerpos, entre otras funciones imprescindibles para la vida de una célula u organismo.

El análisis que se propone en la bibliogafía que estudiamos es considerar a las proteínas como nodos de una red o grafo, donde las conexiones entre ellos representan sus interacciones (PPI). Aún cuando las proteínas que se consideren son las mismas, debido a la multiplicidad de métodos de medición, en distintas redes pueden variar las PPI de una dada proteína. Sumado al aspecto estructural de la red, se ha podido identificar cierto número de proteínas esenciales. Que una determinada proteína sea esencial significa que anularlarla en el organismo resultaría letal para el mismo. De esta forma, la línea de trabajo que reproducimos aquí busca encontrar un paralelo entre la centralidad estructural de una dada proteína (en principio en términos del número de PPI que tiene), y su condición de esencial para el organismo. En particular, en nuestro análisis consideraremos cuatro redes de proteínas de levaduras: Y2H, APMS, LIT y LITR. La característica de esencial se obtiene a partir de una lista provista por He (2006). No todas las redes presentan las mismas proteínas, ni las mismas conexiones entre ellas.

Las redes PPI que consideramos presentan un número bajo de nodos altamente conectados (Hubs) y una cantidad mayoritaria de proteínas con grado bajo. Diversos estudios genómicos prueban de manera empírica que eliminar los nodos de mayor grado de la red tiende a ser más letal que eliminar un nodo poco conectado. Este fenómeno se denomina la "regla de centralidad-letalidad". Sin embargo, previo a los estudios de Zotenko et al. (2008), aun se creía que existía una relación fuerte entre el daño que producía la remoción de una dada proteína a la estructura de la red y su letalidad. Al igual que lo que reproduciremos aquí, Zotenko et al. mostraron que si bien las proteínas esenciales tienden a tener un grado alto en la red, en ningún caso su remoción daña especialmente la estructura de la red, comparado a la remoción de proteínas no letales de igual centralidad estructural.

En este trabajo reproduciremos entonces el análisis realizado por Zotenko et al., tanto los propios, como los que toman de otros trabajos.

2. Características de las redes analizadas

Comenzaremos analizando los atributos estructurales generales de las cuatro redes de proteínas de levaduras de las cuales disponemos: Y2H, AP-MS (APMS), LIT y LIT-REGULY (LITR).

En la tabla 1 se resumen las características principales de las cuatro redes en consideración: la cantidad de nodos, la cantidad de PPI, el grado medio y el coeficiente de clustering medio:

	Y2H	APMS	LIT	LITR
$\#V$	2018	1622	1536	3307
$\#E$	2705	9070	2844	11334
Grado Medio	2.680	1.118	3.703	6.854
Clustering Medio	0.024	0.619	0.346	0.124

Tabla 1: Resumen de magnitudes de las distintas redes consideradas

Podemos observar, que las redes LIT y LITR son más densas que las otras dos, aunque el grado medio de Y2H es alto considerable.

En la tabla 2 se muestra la superposición que hay entre las redes, es decir la proporción de interacciones que tienen en común. Cada fila corresponde con una sola red y se tabula la fracción de ejes que tiene contenidas en las otras redes analizadas.

	Y2H	APMS	LIT	LITR
Y2H	1.00	0.06	0.05	0.08
APMS	0.02	1.00	0.08	0.15
LIT	0.05	0.27	1.00	0.61
LITR	0.02	0.12	0.15	1.00

Tabla 2: Fracción de ejes compartidos entre las distintas redes. En la posición i,j se encuentra la fracción de ejes de la red i que pertenecen a la red j

Vemos que la red Y2H tiene un conjunto de PPI bien distinto de las otras. Por otro lado, APMS tiene bastantes interacciones que están en LITR. LIT esta principalmente contenida en LITR, con una buena parte también en APMS. Por último, LITR comparte esencialmente con LIT. No es de extrañar que LITR contenga bastante a las demás, debido a su enorme número de PPI.

Para analizar la veracidad de la "regla de centralidad-letalidad" en las estas redes, utilizaremos, al igual que Zotenko et al., los resultados obtenidos de eliminar sistematicamente genes de supresión, teniendo en consideración genes que son esenciales para el crecimiento en ambientes ricos en glucosa.

Empezamos analizando cuál es la relación entre tener un grado alto y ser central. Para esto, consideramos todas las proteínas con grado mayor o igual a un cierto valor, y observamos qué fracción de las proteínas esenciales capturamos como "de grado alto". En la figura 1 vemos en el eje horizontal el valor de "cutoff" a, donde el criterio para considerar "alto" un

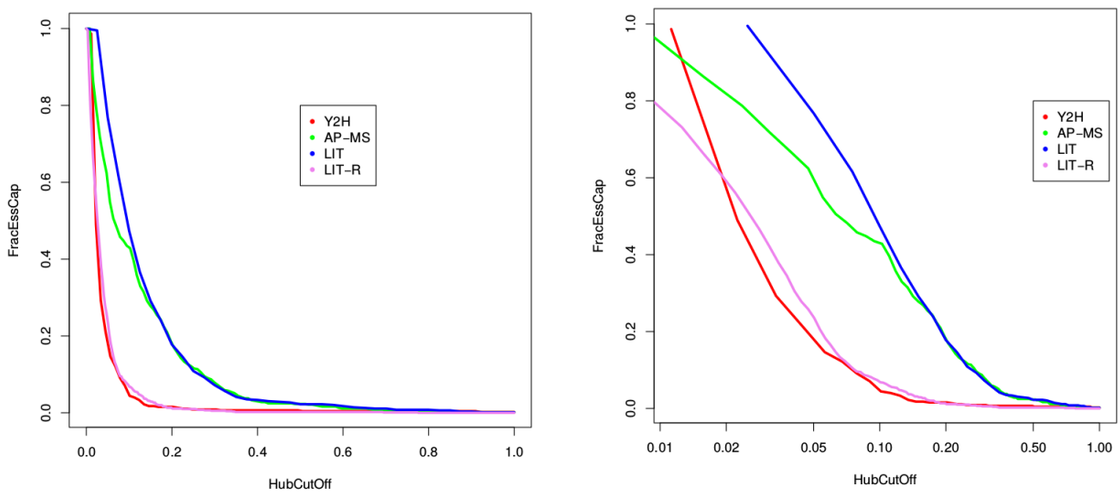

Figura 1: Proporción de nodos esenciales en escala linal y logarítmica

grado es que sea menor o igual que $(1-a)k_{max}$, siendo k_{max} el grado más alto de la red, y en el eje vertical la fracción de proteínas esenciales capturadas. Obviamente, al considerar a igual a cero capturaremos todas las proteínas de la red. Al aumentar a rápidamente decrece la fracción de esenciales, llegando casi a cero en $a = 0{,}1$ para Y2H y LITR. Vemos entonces que las proteínas esenciales tienden a encontrarse entre los grados altos en la red.

3. Análisis de vulnerabilidad

Con el objetivo de revisar el compromiso entre la esencialidad de las proteínas y su rol en la cohesividad de la estructura global, comparamos el efecto de extraer proteínas de la red siguiendo distintos criterios de centralidad de las proteínas, y lo comparamos con el efecto de remover todas las proteínas esenciales de la red. El resultado se puede observar en las figura 3. Los criterios para identificar la centralidad de una proteína son:

- Grado: el número de proteínas con las que interactúa una proteína.
- Betweenness: la fracción de las geodésicas (caminos más cortos) de la red que pasa por el nodo en cuestión.
- Eigenvalue: la importancia del nodo medida a través de su proyección sobre el autovector de mayor autovalor de la matriz de adjacencia de la red.
- Clustering Coef.: la fracción de pares de proteínas vecinas a una, que también interactuan entre sí.
- Random: la elección se hace puramente al azar. Promediamos 10 realizaciones de la extracción para hacerla menos ruidosa. La secuencia de proteínas a extraer es completamente al azar.

En todos los casos se ordenó la extracción de proteínas de la red en base a la centralidad que les correspondía en la red inicial. Luego de remover el nodo, se midió la fracción de proteínas en la componente más grande de la red. Dependiendo de cada red, el efecto de cada método varia:

- Y2H: Vemos que el método mas efectivo para destruir la componente más grande de la red es extraer las proteínas corresponde con ordenarlas por su grado. A continuación se sigue extraer siguiendo su betweenness, y bastante similar es extraer siguien el coeficiente de clustering o su componente de autovalor. Como es de esperarse, la extracción puramente al azar es la menos efectiva.
- APMS: En este caso el método más efectivo es seguir la betweenness, luego el grado, y la componente del autovalor. En esta red extraer al azar es comparable con extraer según el coeficiente de clustering.
- LIT: En esta red extraer siguiendo betweenness o grado tiene un éxito muy similar. Luego lo sigue el eigenvalue, y nuevamente el azar es más efectivo que seguir el componente de clustering durante una etapa.
- LITR: En línea con las anteriores, lo más efectivo es seguir el grado o la betweenness, luego eigenvalue y dependiendo la cantidad extraida, al azar o siguiendo el coeficiente de clustering.

En todos los casos, extraer todos los nodos esenciales de la red tiene un efecto igual o menos potente que el de extraer los nodos siguiendo la centralidad eigenvalue, estando en general entre esta última y la remoción al azar. A partir de este resultado podemos concluir que la característica de esencialidad no está fuertemente correlacionada con la cohesividad que genera la proteína en la red. Remover las proteínas esenciales no es un método eficiente para destruir la red.

Otro método para corroborar esta imagen consiste en comparar el efecto de extraer las proteínas esenciales, con el de extraer proteínas no esenciales con grados lo más similares posibles. Para esto, empleamos el siguiente algoritmo:

- Tomamos la distribución de grado de las proteínas esenciales, e identificamos el número de proteínas con cada grado que hay en la red.
- En caso de existir un número mayor o igual de proteínas no esenciales con el grado de interés, que proteínas esenciales con ese grado, elegimos una muestra al azar de proteínas no esenciales con ese grado.
- En caso de que no existan suficientes proteínas no esenciales con ese grado, tomamos todas las proteínas no esenciales disponibles, y adicionamos con las proteínas que se encuentren en un rango ± 1 del valor de interés, tomando siempre muestras al azar.
- En caso de que sigan siendo necesarias más, volvemos a ampliar el rango en 1.

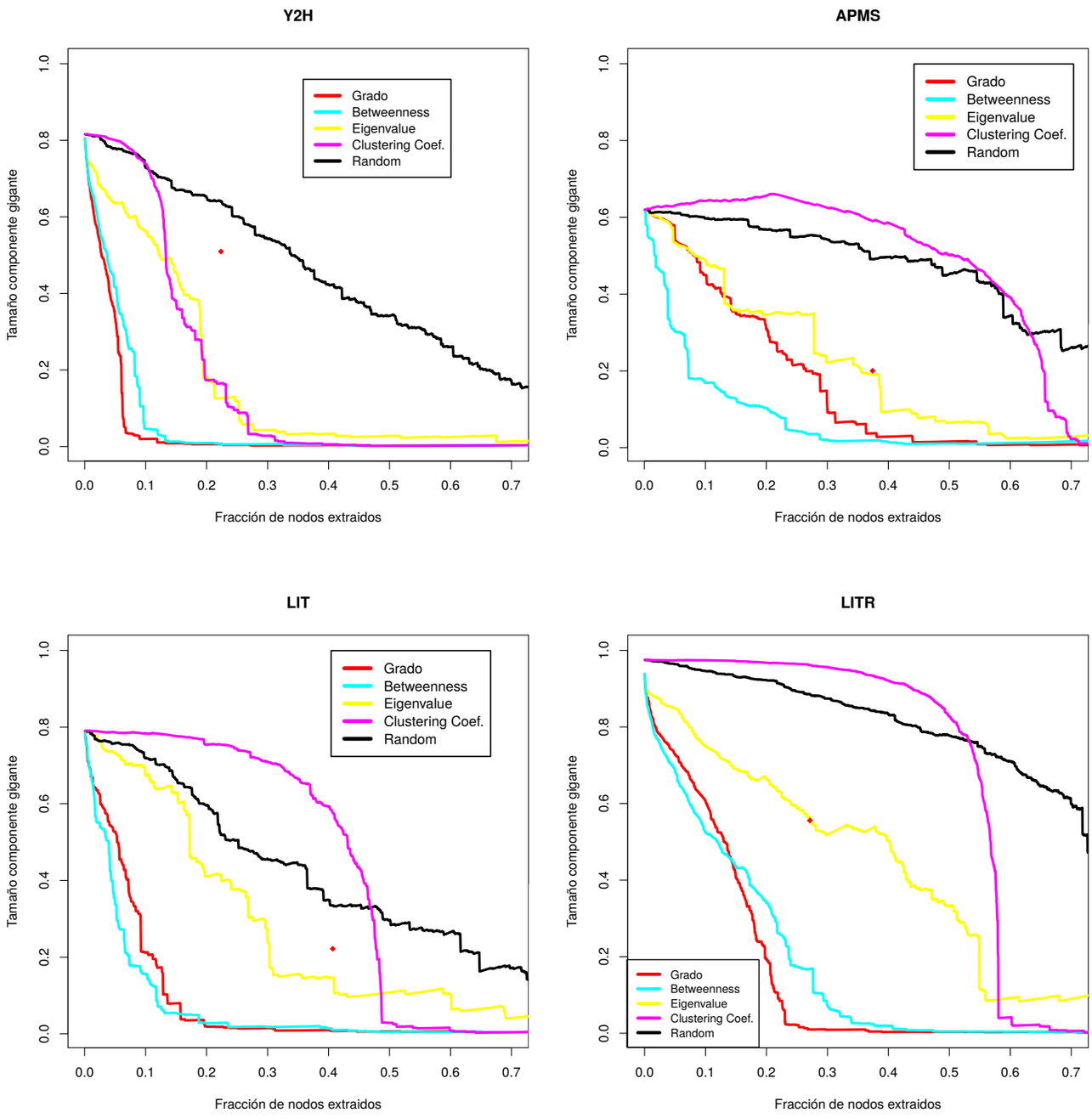

Figura 2: Fracción de nodos en la componente más grande para distintos metodos de remoción de proteínas de la red. El punto rojo indica el efecto de remover todas las proteínas esenciales.

- Repetimos este proceso hasta obtener todas las proteínas necesarias.

En la tabla 3 se puede observar el efecto de extraer las proteínas esenciales, comparado al de extraer no esenciales con el algoritmo anterior. Únicamente en el caso de la red LIT vemos una diferencia substancial entre ambos mecanismos, siendo más efectivo elegir las proteínas esenciales. Podemos atribuir esta diferencia al hecho de que en la red LIT los grados son bajos en general, siendo el más alto 40 para las proteínas esenciales y 24 para las no esenciales, lo que dificulta hacer similares las extracciones (en todos los otros casos, el grado más alto corresponde a una proteína no esencial). Siguiendo la línea que indican las otras tres redes, podemos decir que el éxito obtenido al destruir la red extrayendo proteínas esenciales no es muy distintos del que obtenemos al extraer proteínas no esenciales con el mismo grado. Por esto concluímos que desconectan la red únicamente debido a su grado, sin que la esencialidad aporte nada extra.

4. Esencialidad

En el artículo de He, se analiza la correlación entre la probabilidad de ser esencial y el grado de un nodo. He propone un modelo en el cual la probabilidad de que una proteína sea esencial ($P(E)$) se puede obtener a través de la probabilidad de que la proteína sea esencial (β) y la probabilidad de que ninguna de sus interacciones sea esencial ($1-\alpha$ por cada interacción). Esto resulta en la ecuación:

	Esenciales	No esenciales
Y2H	0.656	0.643 ± 0.011
APMS	0.320	0.347 ± 0.013
LIT	0.374	0.552 ± 0.005
LITR	0.763	0.720 ± 0.006

Tabla 3: Fracción ocupada por la componente más grande luego de extraer las proteínas esenciales, en comparación con extraer proteínas no esenciales con siguiendo el algoritmo explicado.

$$P(E|k) = 1 - (1-\beta)(1-\alpha)^k \tag{1}$$

Esta relación se ve lineal con si consideramos:

$$\ln(1 - P(E|k)) = ln(1-\alpha)k + \ln(1-\beta) \tag{2}$$

como se ve en la figura 2 del artículo. Recreamos dicho gráfico para las redes estudiadas, obteniendo los resultados vistos en las figuras 3 a 6.

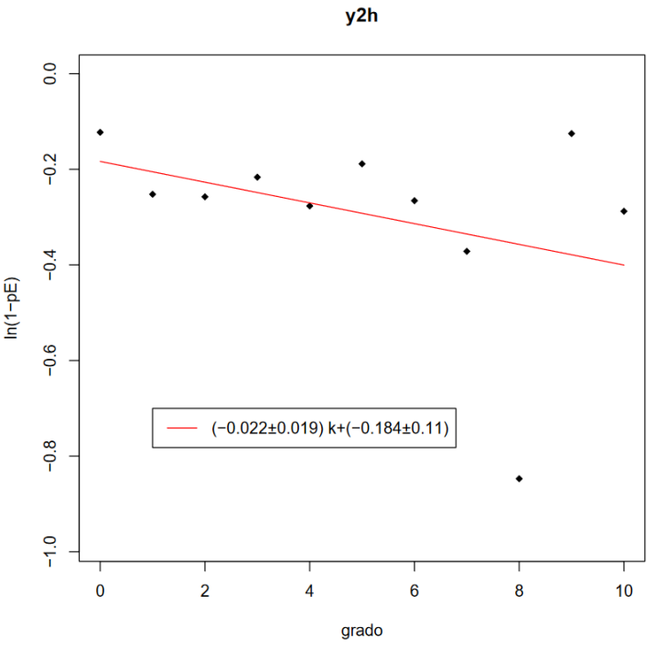

Figura 3: Gráfico mostrando la relación entre la probabilidad de que un nodo sea esencial, y su grado, para la red Y2H.

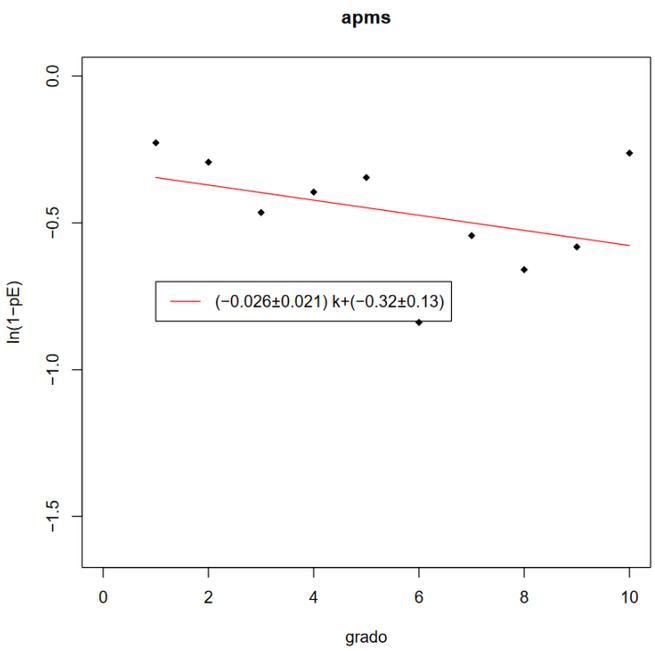

Figura 4: Gráfico análogo a la figura 3 para la red APMS.

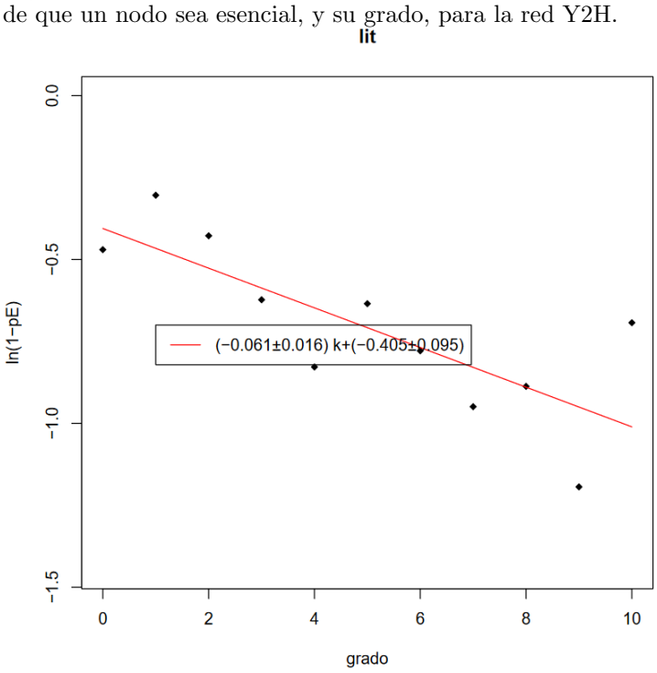

Figura 5: Gráfico análogo a la figura 3 para la red LIT.

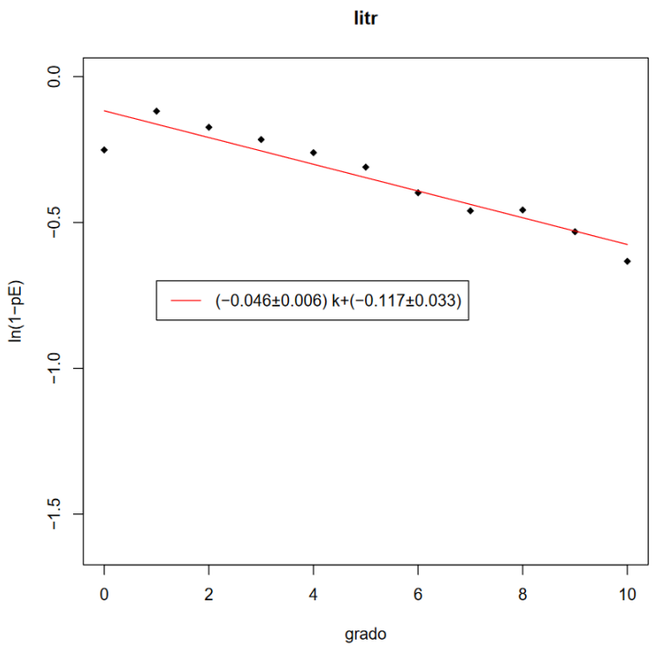

Figura 6: Gráfico análogo a la figura 3 para la red LITR.

Figura 7: Fracción de nodos en la componente más grande para distintos metodos de remoción de proteínas de la red. El punto rojo indica el efecto de remover todas las proteínas esenciales.

Se puede ver en la mayoría de los casos una correlación decreciente entre la ordenada y la abscisa, y en el caso de la red LITR se observa un buen ajuste con el modelo. Esta tendencia indicaría que la probabilidad de que un nodo sea esencial decae como

$$P(E) = 1 - Ce^{Ak} \tag{3}$$

con $\log(C)$ la abcisa de los gráficos 3 a 6, y A la pendiente. Siendo negativa la pendiente y positiva C, esta probabilidad se acerca a uno a medida que aumenta el grado. Estas cantidades se pueden ver resumidas en la tabla 4. Vemos que para las redes Y2H y APMS el valor de α obtenido tiene un error relativo considerable, haciendo dificil separarlo del cero. En las redes LIT y LITR este valor es claramente distinto de cero.

Luego obtenemos una tabla similar a la tabla 5 del artículo de Zotenko. En esta, se considera a los pares de nodos no adyacentes que comparten al menos tres vecinos. Estos nodos pueden ser esenciales o no esenciales. En la tabla 5 se muestra la cantidad total de estos pares y la cantidad cuyos dos nodos pertenecen al mismo tipo (esencial o no esencial). Además, se compara con la expectativa de los pares del mismo tipo, basada en la relación vista previamente entre la probabilidad de ser esencial y el grado del nodo. Calculamos el número esperado de pares esencial-esencial o no esencial - no esencial como

$$\sum_{pares\ ij} P(E|k_i)P(E|k_j) + (1 - P(E|k_i))(1 - P(E|k_j)) \tag{4}$$

Red	α	β
Y2H	-0.022±0.019	-0.168±0.092
APMS	-0.026±0.020	-0.274±0.094
LIT	-0.059±0.015	-0.333±0.063
LITR	-0.045±0.006	-0.110±0.029

Tabla 4: Valores de α y β obtenidos para las distintas redes.

Red	**Pares totales**	**Pares iguales**	**Pares iguales esperados**
Y2H	522	352	261.6164
APMS	11613	5907	5382.023
LIT	730	389	223.026
LITR	10777	6186	3763.264

Tabla 5: Caracterización de los pares de nodos no adyacentes compartiendo tres vecinos o más.

5. Conclusiones

A partir de los análisis realizados, podemos concluir que existe una relación entre el grado de una proteína y su condición de esencial, debido a que la mayoría de las proteínas esenciales tienden a tener grados altos. Sin embargo, no parece valer la relación inversa (que las proteínas con grados altos sean esenciales). El resto de los análisis muestra que las proteínas esenciales no ocupan un rol dominante en la estructura de la red, ni son especialmente centrales, en términos del daño que produce su remoción. El último resultado sugiere que las proteínas esenciales tienden a juntarse con otras, en vez de estar esparcidas por la red. No suelen estar muy lejos, y tienden a tener muchos vecinos en común. Esto agrega un ingrediente fundamental al modelo de He, ya que este no considera correlaciones de este tipo. La propuesta de Zotenko parece razonable: las proteínas consideradas esenciales se estudian más asiduamente, aumentando las chances de que se descubran las proteínas esenciales que están cercanas. Las esenciales aún desconocidas en cambio serían más dificiles de encontrar, requiriendo buscarlas al azar.