Trabajo computacional 2: Centralidad y letalidad

Di Filippo Juan, Catoni Josefina, Yalovetzky Romina

Octubre 2018

1. Introducción

La motivación del presente trabajo es, principalmente, entender cuál es la correlación entre el grado de un nodo en una red de interacción de proteínas con el efecto fenotípico que ocasiona la eliminación de dicha proteína. Jeong et al. [1], propusieron que nodos de alto grado tienden a ser esenciales (su remoción produce efectos fenotípicos) ya que juegan un rol importante en mantener la conectividad global de la red, lo que se conoce como la regla de centralidad y letalidad.
Basados en los estudios de He et al. [2] y Zotenko et al. [3], se reprodujeron algunos resultados que muestran los enfoques adoptados por los mismos con respecto al papel que juegan los nodos esenciales en la red. Para poder realizar dichos análisis, se contó con cuatro redes de interacción de proteínas distintas. Como primera medida se estudiaron las propiedades estructurales de dichas redes y la relación entre el grado y la esencialidad biológica en las mismas. Luego, se procedió a realizar un análisis de la vulnerabilidad de las redes ante la remoción de nodos esenciales adoptando distintos criterios.Finalmente, se estudió como varía la probabilidad de una proteína de ser esencial dado su entorno.

2. Método y resultados

Trabajamos con las 4 redes de interacción de proteínas, dos de ellas obtenidas por experimentos (APMS y Y2H) y otras dos de la literatura. También dispusimos de una lista donde se enumeran las proteínas que se determinan eseciales en el sentido biológico.

2.1. Propiedades estructurales

Primero estudiamos las propiedades estructurales de las redes de proteínas a estudiar (Cuadro 1).

Red	Número de nodos	Número de enlaces	Grado medio	Coef clustering medio
Lit	1536	2925	3,81	0,29
LitReg	3307	11858	7,17	0,26
APMS	1622	9070	11,18	0,55
Y2H	2018	2930	2,90	0,05

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

Se puede ver diferencias significativas entre las redes para algunos de éstos parámetros. Esto se debe a que las redes se construyeron a partir de experimentos distintos. También los enlaces de la red APMS se corresponden con que ambas proteínas participan de los mismos complejos de multiproteínas mientras que los enlaces de la red Y2H se corresponden con contacto físico. Las redes de la literatura hacen una mezcla de ambos tipos de interacciones.

Con el fin de verificar sí las diferencias se deben a dicha razón lo que hicimos fue analizar el solapamiento de enlaces que se define como la fracción de interacciones en común entre dos redes (Cuadro 2).

Como se trata de redes de proteínas las redes son no dirigidas. Sin embargo, estudiamos sí se daba en algunos casos que se mostrara un enlace de la forma "AB" y también "BA". Notamos que eso solo sucede en las redes lit, litReg y Y2H pero para las mismas proteínas. Eso quiere que decir que se dan self loops. Tomamos el criterio de analizar también sí los enlaces de self loops eran compartidos entre redes. Entonces, calculamos la cantidad de enlaces compartidos considerando las interacciones de las proteínas con ellas mismas.

Se observa que la red Y2H es la que menos enlaces comparte con las demás redes. Esto es importante porque para algunos análisis que haremos más adelante es de esperar que se ven los efectos menos prominentemente en ésta red.

Lit	0,98	0,44	0,09
0,24	LitReg	0,21	0,04
0,14	0,28	APMS	0,03
0,09	0,16	0,09	Y2H

Cuadro 2: Cantidad de solapamiento de enlaces entre las redes estudiadas

Cada fila se corresponde con una red (se indica en la diagonal) y muestra la fracción de sus enlaces que se encuentran contenidos en cada una de las otras redes. Los elementos de cada fila son normalizados con la cantidad de enlaces total de la red de dicha fila. Por ejemplo, el 24 porciento de los enlaces de la red LitReg también están presentes en la red Lit.

2.2. Regla de centralidad letalidad

Se definen hubs como grupos de nodos de alto grado. Es decir que conformados por nodos que tienen al menos tal grado que se considera alto dado su contexto.Dado lo que comentó en la Introducción sobre la regla de centralidad y letalidad una pregunta es sí los hubs de alto grado (mayor o igual que un cierto parámetro) tienen una tendencia mayor a ser esenciales biológicamente. Es por eso que se analizó la fracción de nodos esenciales que existe en los distintos hubs de nuestras redes.

Lo que se hizo fue ordenar los nodos de las redes de mayor a menor grado y contar la cantidad de nodos que tenían al menos grado 1,2,...,k_{max}. Es decir que calculamos la cantidad de nodos en cada hub. Luego, a partir de la lista que se disponía de proteínas esenciales biológicamente se calculó la fracción de nodos esenciales de cada hub. El resultado se observa en la Fig.1. Se puede ver que para los hubs menos poblados, fracciones más cercanas a 0, que son formados por aquellos nodos que tienen al menos un grado bastante alto, la fracción de nodos esenciales colapsa para valores muy cercanos al 1. Luego, se produce un comportamiento de decrecimiento hasta colapsar en un valor de esenciales cuando la fracción de nodos es 1. Se puede ver que cada red alcanza un valor distinto. Esto tiene que ver con la cantidad de nodos esenciales que fueron mustreados en cada relevamiento de redes.

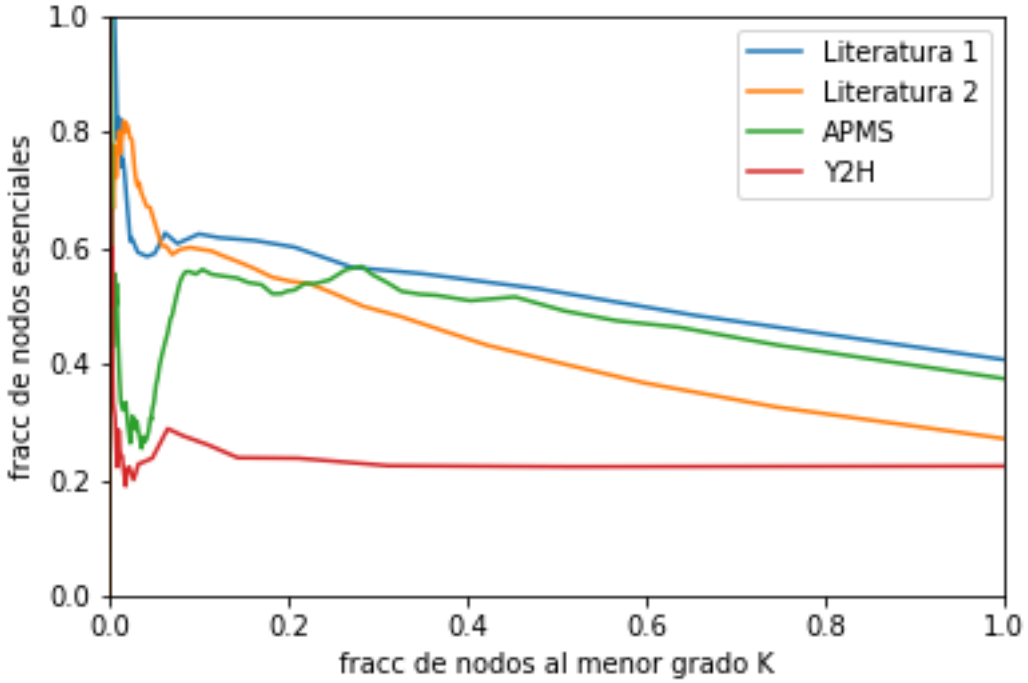

Figura 1: Relación entre grado y la esencialidad biológica en las redes estudiadas. Eje horizontal muestra la fracción total de nodos que forman hubs de determinado grado y el eje vertical la fracción de nodos esenciales biológicamente de dichos hubs

2.3. Vulnerabilidad

Luego, se realizó un análisis de la vulnerabilidad de la red ante la remoción de nodos a partir del estudio realizado por Zotenko et al. [3]; es decir, cuánto cambiaba el tamaño de la componente gigante de la red a medida que se iban quitando nodos, con el fin de verificar si la esencialidad podría asociarse directamente a la capacidad de mantener la conectividad de la red.

Se definieron distintos criterios para elegir en qué orden se irían sacando los nodos seleccionando distintos coeficientes de centralidad. Para cada coeficiente, la metodología fue la siguiente:
a) Nos quedamos con la componente gigante (CG) de la red y se descartó todo el resto.
b) Calculamos el coeficiente de cada nodo y se seleccionó aquel que tuviera el mayor valor de dicho parámetro.
c) Quitamos ese nodo de la red y registramos el tamaño de la nueva CG.
d) iteramos los pasos a)-c) hasta que el tamaño de la CG fuera menor a 10.

Los coeficientes de centralidad elegidos fueron: Centralidad de grado ('degree'), de intermediación ('shortest-path'), de vector propio ('eigenvectors'), de subgrafos ('subgraph') y de circulación ('current-flow'). Además, se agregó un criterio al azar ('random'), es decir, que se fueron quitando los nodos de manera completamente aleatoria.

Finalmente, a partir de los datos de esencialidad de los nodos, se quitaron todos los nodos esenciales de la red y se calculó el tamaño de la componente gigante resultante.

Se graficó el tamaño de la componente gigante (normalizado por el tamaño de la componente gigante original) en función de la fracción de nodos quitados (Nodos quitados/Tamaño de CG original). En la figura 2 se observan los resultados para las cuatro redes.

Se aprecia que para todas las redes, el efecto de remover los nodos esenciales se asemeja al de

quitarlos de manera aleatoria, mientras que quitarlos según su centralidad desarma la red mucho más rápido. En particular, El índice de intermediación fue el que más impacto en la conectividad general de la red, que se cuantifica a través del tamaño de la componente gigante.

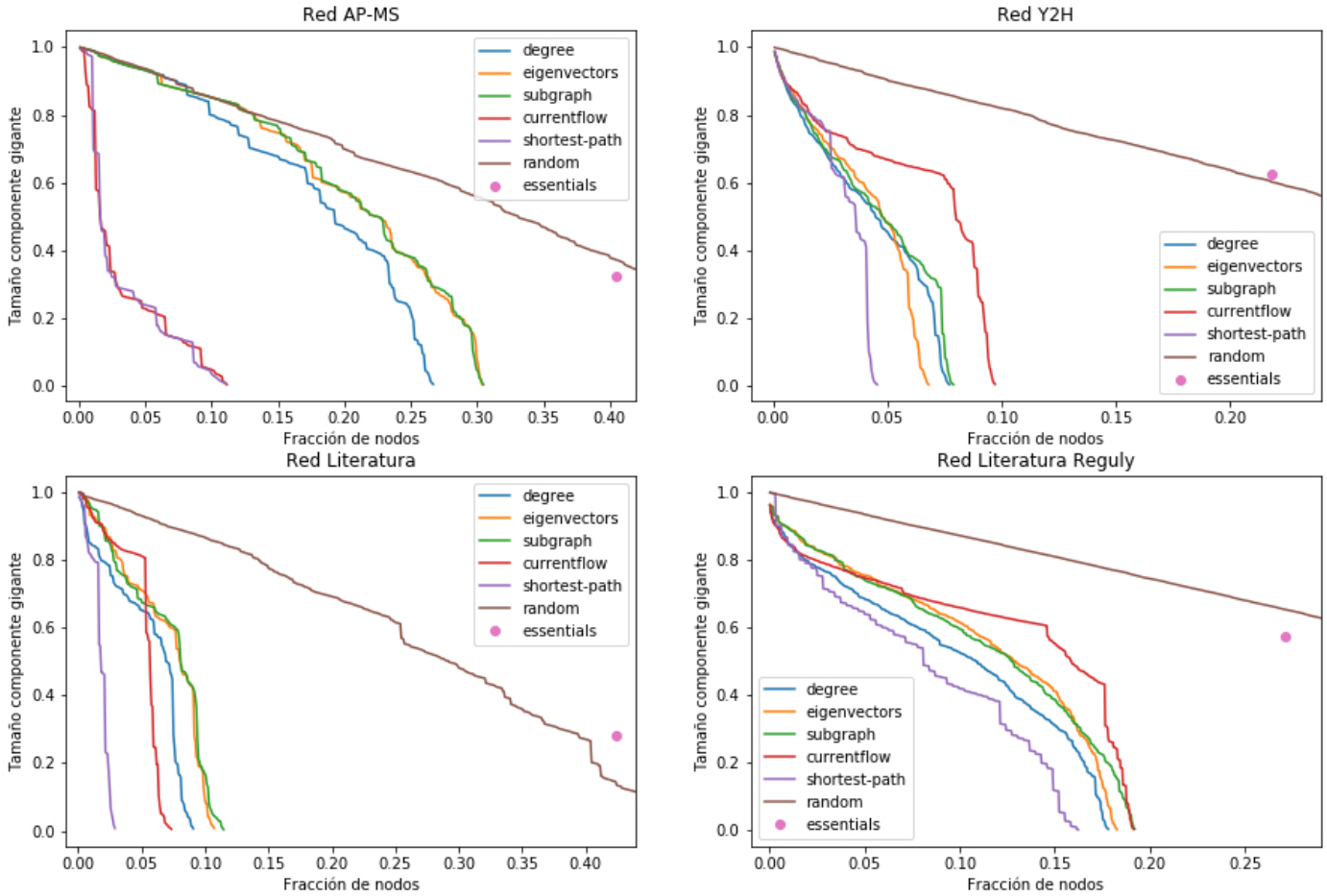

Figura 2: Vulnerabilidad de la componente gigante ante remoción de nodos con alta centralidad, esenciales o aleatorios. En el eje de abscisas se tienen la cantidad de nodos removidos normalizados por el tamaño de la CG original. En el eje de ordenadas se tiene el tamaño de la componente gigante resultante normalizado también por el tamaño de la CG original.

Se concluyó, entonces, que los hubs esenciales no son más importantes que los no esenciales en mantener la conectividad general de las redes.

Con el fin de darle aún más validez a lo concluido, se comparó la disrupción de la red al quitar todos los nodos esenciales de la componente gigante con la disrupción de quitar un número equivalente de nodos no esenciales seleccionados de manera aleatoria, siempre y cuando mantuvieran, bajo cierto criterio, la misma distribución de grado que los esenciales.

El criterio seleccionado fue el siguiente. Para poder mantener la misma distribución de grado al quitar nodos no esenciales, debemos verificar si existe algún grado para el cual haya nodos esenciales pero no haya no esenciales. Además, dentro de nuestra aleatoridad, queremos que todos los nodos no esenciales tengan una probabilidad no nula de ser "seleccionados"para ser removidos. Entonces, para cada red, se hizo un histograma con la distribución de grado para los nodos esenciales y no esenciales de la componente gigante (Ver figura 3).

A partir del histograma, se dividió el "vector"de grados en bins, de manera que todos los bins fueran lo más angostos posibles tal que se cumplieran las dos condiciones mencionadas previamente. Para ello, se tomó como bineo inicial el logarítmico, y se fueron ajustando los límites de manera que la cantidad de no esenciales por bin fuera mayor que de esenciales.

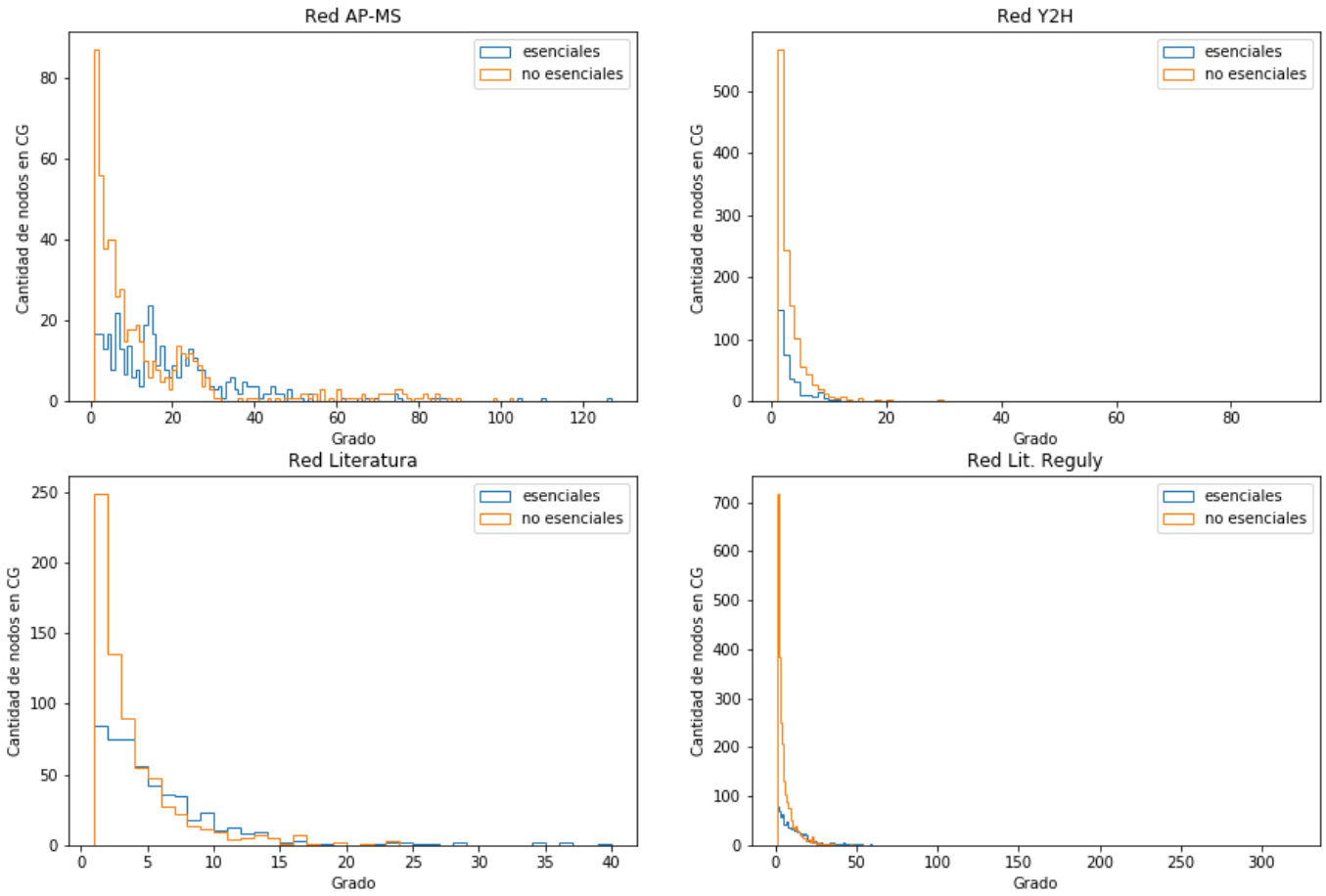

Figura 3: Distribución de grado de nodos esenciales y no esenciales en cada red. El histograma muestra la cantidad de nodos que tienen grado k en la componente gigante original.

Finalmente, se realizó la "selección"de nodos no esenciales que serían quitados. Para eso, se eligió de manera aleatoria la misma cantidad de nodos esenciales que hubiese por bin. Se seleccionaron 1000 combinaciones distintas de nodos para cada red. Para cada combinación se calculó el tamaño de la componente gigante final despues de remover los nodos. Normalizando este valor con el tamaño de la componente gigante original, se obtiene el impacto de quitar los nodos de la red. En el cuadro 3 se ilustran los resultados.

Una vez más, se observa que el impacto de remover nodos esenciales no es mayor al de remover no esenciales, lo que indica que la esencialidad no viene de la mano de capacidad de mantener conectada a la red.

Red	Esenciales	No Esenciales
Lit	0,281	$0{,}423 \pm 0{,}006$
LitReg	0,575	$0{,}564 \pm 0{,}021$
APMS	0,324	$0{,}389 \pm 0{,}019$
Y2H	0,624	$0{,}687 \pm 0{,}010$

Cuadro 3: Impacto de remoción de todos los nodos esenciales de mi componente gigante en comparación la misma cantidad de no esenciales con la misma distribución de grado.El impacto se mide como el cociente entre la cantidad de nodos en la componente gigante resultante y la original.

3. Esencialidad

Otro de los estudios realizados en este trabajo fue analizar como varía la probabilidad de una proteína de ser esencial dado su entorno. En primer lugar, reproducimos uno de los resultados de He

et al. [2], que básicamente muestra que la probabilidad de un hub de ser esencial es alta debido a su conectividad. Por otro lado, siguiendo los pasos de Zotenko et al. [3], mostramos que las hipótesis asumidas para llegar al resultado anterior no son válidas, puesto que a partir de ellas se llega a resultados no correspondidos con lo observado en las redes reales.
Cabe destacar que en este apartado no se trabajará con la red AP-MS, pues como mencionan He et al. en su trabajo, el modelo no es apto para redes en las cuales los enlaces no representen interacciones directas entre proteinas sino que los mismos se dan entre proteínas que pertenecen a un mismo complejo proteico.

3.1. Interacciones esenciales

Dada una red, la probabilidad de que una proteína de la red sea esencial (α) y la probabilidad de que una interacción proteína-proteína sea esencial (β), la probabilidad de que una proteína de grado k sea esencial es:

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

A partir de esta expresión se encuentra fácilmente que:

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

Trabajando con las redes Y2H, Lit y LitReg, se calculó para cada una de ellas la probabilidad de un nodo de ser esencial en función del grado. Dicha probabilidad se calculó simplemente como la fracción de proteínas esenciales sobre el total, para cada grado presente en la red.
Luego, haciendo uso de que $\ln(1-P_E))$ y k estan relacionados linealmente (Ec. 1), procedimos a graficar esta dependencia y realizar el ajuste correspondiente para poder obtener los parámetros α y β. En la Fig. 4 se muestra el ajuste lineal realizado para cada red.

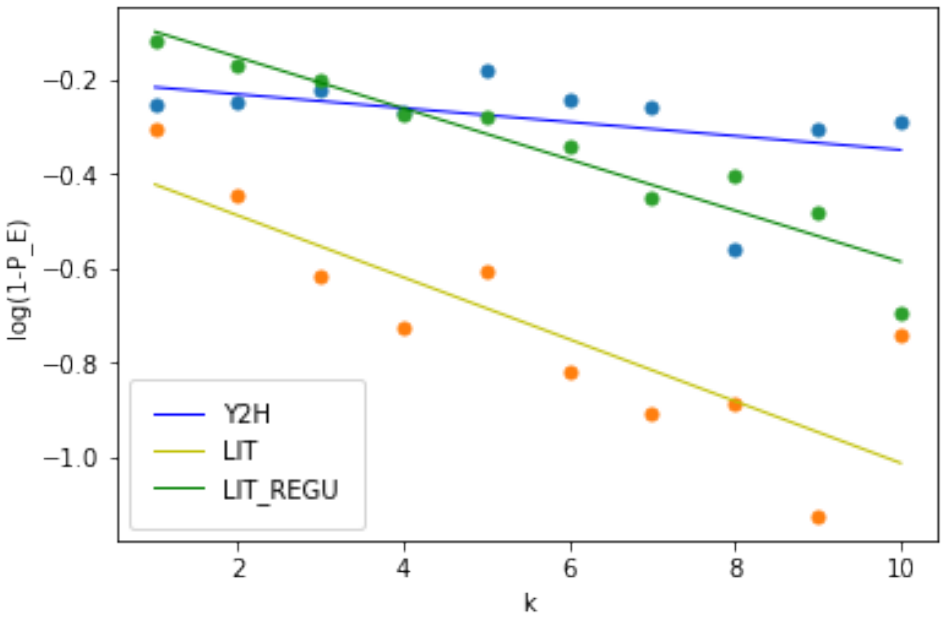

Figura 4: Relación entre la probabilidad de que una proteína sea esencial y su conectividad

Se observa que un nodo de alto grado es mas propenso a ser esencial que un nodo con baja conectividad. Esto se debería a que dicho nodo tiene mas interacciones y, por lo tanto, una probabilidad mayor de estar involucrada en una interacción esencial.

No se incluyeron datos con $k > 10$ puesto que esos nodos restantes representan menos del 5 % de los mismos y las muestras de sampleo son pequeñas. En el cuadro 4 se muestran los α y β obtenidos a partir de los coeficientes de los ajustes lineales.

RED	α	β
Y2H	$(0{,}55 \pm 0{,}01)\,\%$	$(21{,}9 \pm 1{,}0)\,\%$
Lit	$(6{,}36 \pm 0{,}08)\,\%$	$(30{,}0 \pm 2{,}1)\,\%$
LitReg	$(5{,}28 \pm 0{,}03)\,\%$	$(4{,}27 \pm 0{,}14)\,\%$

Cuadro 4: α y β obtenidos a partir de los coeficientes del ajuste lineal para cada red

3.2. Módulos biológicos

En el apartado anterior se trabajó con la suposición de que si dos proteinas no interactuan entre si entonces la esencialidad de una no afecta la probabilidad de que la otra sea esencial. En particular, esta suposición vale en el caso en que dos proteínas no interactuen directamente entre sí pero compartan vecinos. Para verificar la validez de esta hipótesis se procedió de la siguiente forma:

- Calculamos la cantidad de pares de proteínas no interactuantes con tres o más vecinos en común para la red LitReg y con 2 o más vecinos en común para las redes Y2H y Lit. Para encontrar los pares de proteínas que cumplen estos requisitos utilizamos la matriz de adyacencia A: Basta recorrer los elementos de A y A^2 y buscar los elementos que satisfacen simultaneamente $A_{ij} = 0$ y $A^2_{ij} \geq \#$ vecinos compartidos
- De la lista de pares totales, nos quedamos con un subconjunto correspondiente a pares esencial-esencial y no esencial-no esencial, lo que llamamos pares del mismo tipo.
- Dados los α y β calculados a partir de la regresión lineal (cuadro 4) calculamos la cantidad esperada de pares del mismo tipo (N_{exp}) utilizando la lista de pares totales:

$$N_{exp} = \sum_{pares(i,j)} P_E^i \times P_E^j + (1 - P_E^i) \times (1 - P_E^j)$$

En el cuadro 5 se pueden observar los resultados obtenidos, los cuales muestran una diferencia apreciable entre los pares del mismo tipo reales y los esperados bajo las suposiciones del modelo anterior.

RED	Pares Totales	Pares del mismo tipo	Pares del mismo tipo esperados
Y2H	2258	1514	1307 $\pm$15
Lit	1858	1047	961 $\pm$8
LitReg	10777	6187	5779 $\pm$8

Cuadro 5: Comparación entre el número de pares del mismo tipo presentes en la red real y el esperado a partir del calculado utilizando α y β (Ec 1)

Este resultado muestra que la probabilidad de que dos nodos no enlazados sean esenciales no es independiente, contradiciendo la hipótesis utilizada al calcular α y β en la subsección anterior. En particular, es importante destacar que la cantidad de pares esperados es siempre mas baja que la cantidad real. Esto muestra, en definitiva, que la escala con la que esta relacionada la esencialidad no es del todo local, es decir, la probabilidad de un nodo de ser esencial depende no solo de su conectividad, sino de en que comunidad de nodos se encuentre.

Referencias

[1] JEONG, Hawoong, et al. Lethality and centrality in protein networks. Nature, 2001, vol. 411, no 6833, p. 41.

[2] Why Do Hubs Tend to Be Essential in Protein Networks?; Xionglei He, Jianzhi Zhang; Plos genetics,2006.

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