# Trabajo Computacional 2

Favio Di Ciocco, Daniel Pinto, Diego Espejo, Alejandro Barton

19 de octubre de 2018

## 1. Tres Visiones de la regla de Centralidad-Letalidad en proteinas

- Se entiende por "Regla de Centralidad-Letalidad" la idea de que en una red existen nodos que tienen una mayor importancia para la red que otros y que de ser removidos, alteran drásticamente la estructura de la red, desconectando pares de nodos o aumentando mucho la longitud del camino entre estos, llegando incluso a descomponer la componente gigante. Este fenómeno se puede observar en la red de proteínas de la levadura, en donde la remoción de estos nodos centrales afecta de manera terminal funciones biológicas vitales, de ahí que se le diga letal.

1. Es sobre esta idea que comienza a trabajar H. Jeong en su paper de 2001: "Lethality and Centrality in Proteins Networks"[1] respecto a la centralidad de ciertas proteinas en las redes de levadura. Él logra observar que la remoción de ciertas proteínas en la levadura con un alto conexionado en la red, provocan en la levadura una incapacidad de metabolizar nutrientes y reproducirse, al mismo tiempo que producen un rápido incremento en el diámetro de la red. Es esta correlación la que lo lleva a pensar que lo que ocurre es que existen ciertas proteínas esenciales para el desarrollo de la levadura y que además, los nodos más conectados de la red -*hubs*- tienen mayores chances de resultar ser proteinas esenciales que de no serlo.

   Es a este punto que él reporta que la probabilidad de que un nodo altamente conexionado sea una proteína esencial es tres veces mayor que la probabilidad de que un nodo con pocos enlaces lo sea.

2. Por otra parte, X.He propone una visión diferente al respecto, asignando la esencialidad a los enlaces entre proteinas y no a las proteínas en sí. X.He explica en su paper de 2006:"Why do Hubs tend to be essential in Protein Networks?"[2] que lo que en verdad sucede es que los hubs al ser nodos de alta conectividad tienen mayores chances de pertenecer a un enlace esencial, o lo que es lo mismo, a una interacción entre proteínas esencial.

3. Finalmente H.Zotenko propone una tercer visión a este dilema en su paper de 2008: "Why do Hubs in the Yeast Protein Interaction Network tend to be essential: Reexamining the Connection between the Network Topology and Essentiality."[3] Es en este Paper que Zotenko desestima las dos explicaciones anteriores respecto a la esencialidad de las proteínas o los enlaces, y le atribuye un caracter de centralidad a ciertos módulos biológicos complejos que se forman de la interacción de grupos de proteínas. Ella propone que remover proteínas de estos complejos es lo que produce la letalidad en la levadura. Para sostener esta proposición ella simula computacionalmente los efectos sobre la red de la remoción de proteínas esenciales y no esenciales según el criterio de H.Jeong y también hace un análisis similar para la remoción de enlaces, demostrando que la remoción de los llamados enlaces esenciales o las llamadas proteínas esenciales, en comparación con otros enlaces o proteínas no esenciales provocan efectos similares en la red.

   Es esto lo que la lleva a considerar la idea de que lo que esté sucediendo es que haya ciertos conjuntos de proteínas relacionadas que tengan importancia como grupo y que al ser removida una sola de las proteinas del grupo, el efecto sobre la red sea letal. Estos grupos son los llamados Módulos Biológicos Complejos.

## 2. Objetivos

A continuación replicaremos algunos de los análisis hechos en los papers de Zotenko, Jeong y He para estudiar cuatro relevamientos de las interacciones de proteínas entre levaduras.

## 3. Redes relevadas

Las redes relevadas a utilizar son las siguientes: LIT, AP-MS, Y2H y LIT Reguly. Las redes LIt y LIT Reguly fueron relevadas de literatura, la Y2H se obtuvo usando la técnica "Yeast to Hybrid", y la AP-MS se obtuvo a partir del método "Affinity Purification-Mass Spectrometry"

En la Tabla 1 se muestran los datos característicos de cada una de estas redes.

| | Número de Nodos | Número de enlaces | Grado Medio | Coeficiente de Clustering medio |
|---|---|---|---|---|
| LIT | 1536 | 2925 | 3.809 | 0.292 |
| AP-MS | 1622 | 9070 | 11.184 | 0.555 |
| Y2H | 2018 | 2930 | 2.904 | 0.046 |
| LIT Reguly | 3307 | 11858 | 7.171 | 0.261 |

Tabla 1: Cantidad de nodos, enlaces, grado medio y coeficiente de Clustering medio de cada una de las redes

Las redes se hallan ordenadas de menor a mayor según el número de nodos. Es a partir de esta tabla que se puede observar viendo la cantidad de nodos y de enlaces que el tamaño y la conectividad de cada red es variable. Por tanto, para poder establecer relaciones entre ellas es importante estudiar que tan semejantes son unas con otras. Esta semejanza se mide en función del overlapping de ambas redes, que se calculó a partir de la fracción de enlaces de una red que se repiten en la otra red. Estos valores se muestran en la siguiente tabla:

| | | | |
|---|---|---|---|
| LIT | 0.259 | 0.053 | 0.619 |
| 0.084 | AP-MS | 0.019 | 0.145 |
| 0.053 | 0.060 | Y2H | 0.090 |
| 0.153 | 0.111 | 0.022 | LIT Reguly |

Tabla 2: Valores de Overlapping entre las redes relevadas

Podemos ver en la segunda tabla que el par de redes más semejantes entre sí es el LIT-LIT Reguly, ésto es, comparten una mayor fracción de enlaces entre sí que todas las demás. Vemos también que ambas tienen un coeficiente de clústering medio parecido con lo cual nos parece razonable poder compararlas. A su vez, vemos que, respecto a los enlaces, la red LIT Reguly tiene aproximadamente 4 veces más enlaces que la LIT, con lo cual esperaríamos que el overlapping de la red LITR a la LIT no superara el 0,25 y efectivamente el valor que obtenemos es de 0.15.

Podemos observar también que la red que menos se asemeja a las demás es la Y2H, que cuenta con valores de overlapp debajo de 0,1, entendemos que ésto se da por el bajo coeficiente de clústering medio y por el bajo valor de grado medio de la red, lo cual nos dice lo poco conectada que está respecto de las demás.

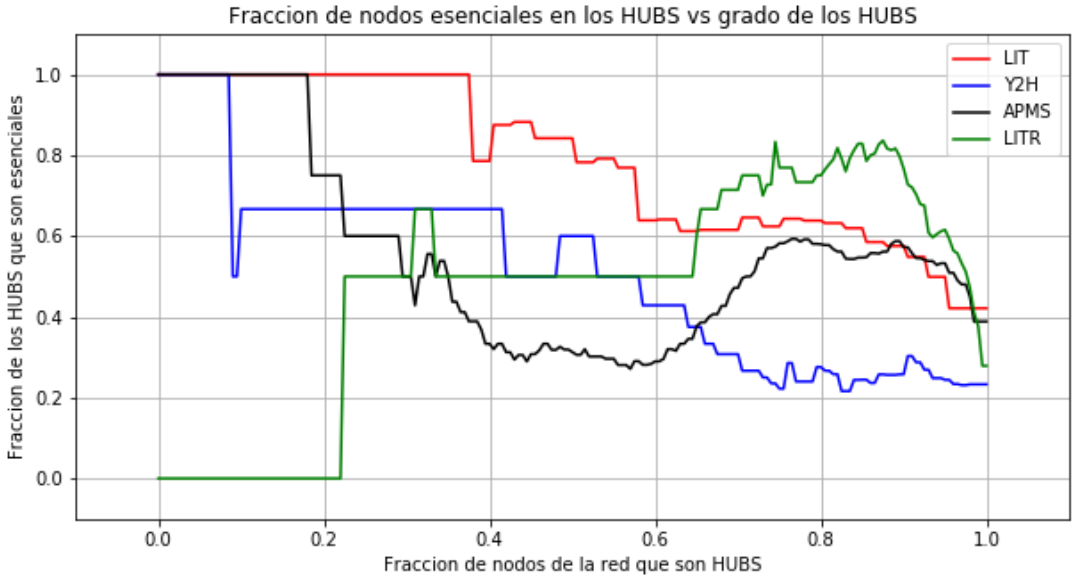


Figura 1: Fracción de nodos esenciales en función de la fracción de nodos de la red que son HUBS

De la figura 1 se puede observar que para las redes LIT, Y2H y la APMS los nodos de alto grado son en gran proporción esenciales, y que a medida que la definición de HUBS abarca una mayor cantidad de nodos, la cantidad de nodos esenciales se reduce a valores entre el 20 % y el 40 % . Esto último pasa también para la red LIT Reguly.

Por otra parte, para la red LIT Reguly suceden cosas que no son las esperadas, tanto para las fracciónes de nodos chicas como para las fracciones de nodos cercanas a uno.

Para las fracciones de nodos cercanas a cero sucede que la fracción de nodos esenciales es cero. Esto se debe porque hasta cerca del 0.2 de la red LITR, sólo un nodo cumple con ser un HUB, ya que la distancia en grados entre el nodo de máximo grado y el siguiente es mayor al 15 % . Entonces, como ese nodo no cumple con estar en la lista de esenciales, no hay forma de revertir este comportamiento. Esto se comprobó revisando manualmente la lista de nodos en la lista de proteínas esenciales de He y en la lista de nodos de la red LIT Reguly.

El otro comportamiento extraño está cuando se considera que la fracción de nodos de la red que son HUBS está entre 0.8 y 0.95. Para estos valores, la fracción de nodos esenciales en la red crece hasta valores cercanos a 0.8, y en el último tramo decae abruptamente a valores cercanos al 0.3.

Se revisó también la cantidad de nodos esenciales, que es 1120, y la cantidad de nodos en cada red que el programa detectaba como esenciales, y en ningún caso el programa detectaba más de 100 nodos esenciales, así que tampoco pareciera ser un problema de sobreestimación de nodos esenciales.

Por último para comentar, la razón de que las curvas de los gráficos se vean tan cuadradas en los valores iniciales de Fracciones de nodos de la red, está asociado a que los para estas fracciones del orden de 5 nodos son HUBS, entonces cualquier nodo extra que sea esencial hace un cambio abrupto en el gráfico.

## 4. Centralidad en las redes

A continuación, pasamos a analizar el impacto de remoción de nodos según diferentes criterios de centralidad. Los criterios de centralidad utilizados fueron 3:

- **Degree:** Centralidad $\propto$ grado del nodo
- **Betweenness:** La cantidad de caminos más cortos que pasan por el nodo
- **Eigenvector:** La centralidad de un nodo es el promedio de la centralidad de sus vecinos

La idea en general es asignar un valor de centralidad a cada nodo de la red, para luego ir sacando los nodos de mayor centralidad, observando cómo se altera la conectividad de la red. El parámetro que usamos para medir esta conectividad es el tamaño de la componente más grande, luego de remover los nodos más centrales, en relación al tamaño que tenía antes de removerlos. Así, a medida que removemos nodos de alta centralidad del grafo, estamos destruyendo las componentes conexas grandes, lo que se traduce, en la levadura, a ir destruyendo tanto la conectividad entre proteínas como los módulos que estas forman.

Lo que analizamos acá en principio tiene únicamente que ver con la topología de la red, pero, como veremos en la tabla 3, la centralidad Degree está correlacionada con sacar nodos esenciales, ya que remover nodos de grados altos produce prácticamente el mismo impacto en la componente gigante que remover nodos esenciales, pues los nodos esenciales suelen tener grado alto. Por eso es esperable que el impacto en la topología de la red se corresponda con un impacto en la supervivencia de la levadura.

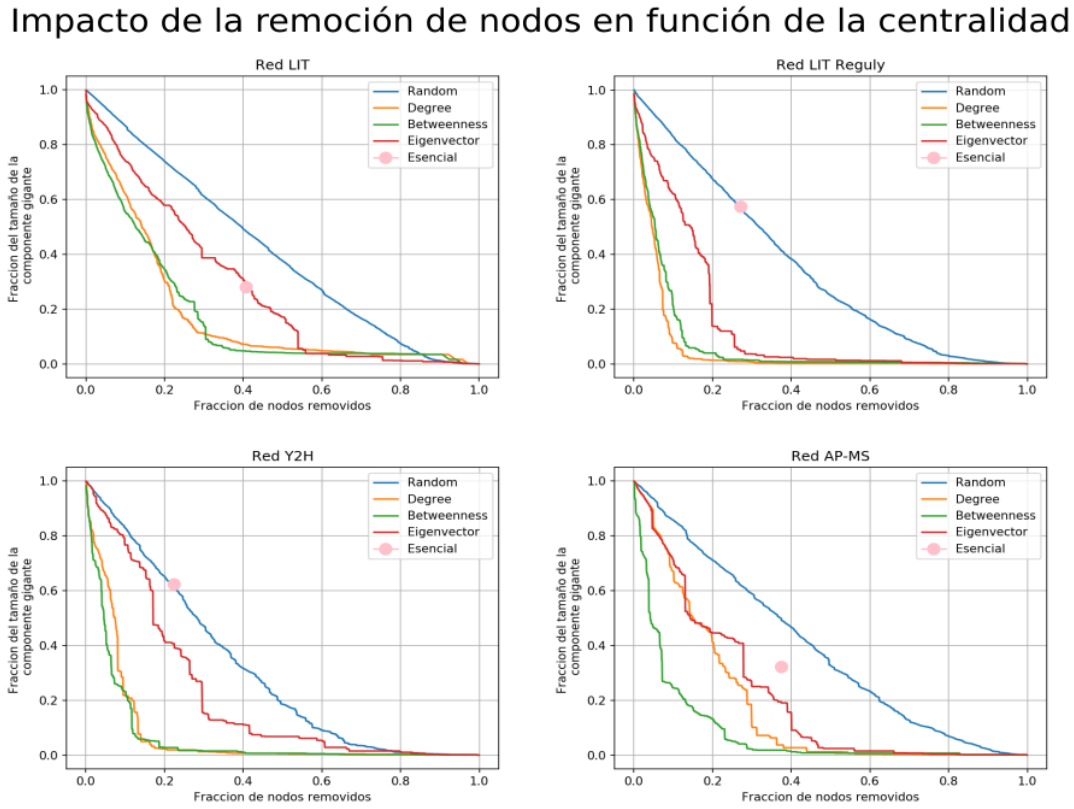


Figura 2: Medidas del impacto en la topología (conectividad) de la red según diferentes criterios de centralidad. El punto rosa es el tamaño relativo de la componente gigante luego de remover todos los nodos esenciales

Viendo la figura 2, se puede notar que, al haber removido una cantidad de nodos de mayor centralidad igual a la cantidad de nodos esenciales, el resultado en el impacto en la red es mucho mayor con la remoción de nodos centrales que con la remoción de nodos esenciales, para todos los criterios de centralidad.

De hecho, en dos de las redes, LIT Reguly e Y2H, el punto rosado en la figura 2 está sobre la curva de centralidad Random, lo que significa que incluso sacando nodos al azar en una misma cantidad que los nodos esenciales, la componente gigante tiene tamaños similares. Esto demuestra que, en algunos casos, los nodos esenciales no tienen nada que ver con la preservación de la conectividad de la red.

La centralidad Betweenness que usamos es la *shortest-path betweenness*, que, al igual que en el paper de Zotenko, tiene un impacto muy similar a la centralidad Degree, como se puede obsevar en la figura 2. En todas las redes, estas dos centralidades son las que tienen mayor impacto. La centralidad eigenvector es la menos efectiva a la hora de destruir la red, lo que también concuerda con lo que sucede en el paper de Zotenko.

Esto quizás se deba a que la centralidad Betweenness mide directamente la capacidad de un nodo de

conectar otros dos nodos distintos, y entonces es una medida directa de lo importante que es un nodo para la conectividad de la red. En cambio, no hay una clara relación entre la centralidad Eigenvector y la conectividad de la red. Notemos que el hecho de que las curvas Betweenness y Eigenvector sean notablemente diferentes en la figura 2 nos dice que los nodos importantes para la conectividad no tienen que ser vecinos entre sí.

La curva Random sirve para contrastar con lo que se esperaría al sacar nodos al azar, y vemos que, como es esperable, tiene el menor impacto. Para calcular la curva, lo que se hizo fue sacar un nodo al azar, medir la componente gigante, luego volver a sacar un nodo al azar, y volver a medir la componente gigante, continuando así hasta sacar todos los nodos. Luego repetir el proceso 10 veces, y finalmente promediar.

| | Esenciales | No esenciales aleatorios |
|---|---|---|
| LIT | 0.374 | $0.328 \pm 0.003$ |
| AP-MS | 0.320 | $0.247 \pm 0.007$ |
| Y2H | 0.656 | $0.513 \pm 0.008$ |
| LIT Reguly | 0.767 | $0.578 \pm 0.004$ |

Tabla 3: Impacto de la remoción de proteínas esenciales en las redes en comparación de la remoción de proteínas no esenciales con la misma distribución de grado

Se cuantificó el impacto de la remoción de nodos en cada red a través de calcular la fracción de tamaño de la componente gigante. Se puede observar en la tabla 3 que para todas las redes la remoción de nodos no esenciales de manera aleatoria tiene un impacto mayor, ya que el tamaño de la componente gigante es menor en estos casos que en los casos de remoción de nodos esenciales.

Los valores de la columna de "No esenciales aleatorios"se calcularon realizando primero una lista de nodos que permitiera, en la medida de lo posible, respetar la distribución de grado de nodos esenciales. Es decir, que para un dado grado k, si en la remoción de nodos esenciales se quitaban 5 nodos de grado k, en la remoción de nodos no esenciales también se removieran 5 nodos de grado k. Esto no siempre era posible porque para grados cada vez mayores era más dificil encontrar una cantidad de nodos no esenciales igual a la cantidad de nodos esenciales. Es por esto que se recurrió a tomar prestados nodos de grados menores, lo más próximos posibles, quedando así la lista agrupando según el "grado.[a] los nodos, pero entendiendo que para ciertos grados, aparecián nodos de grados menores.

Una vez armada la lista, se hicieron 500 iteraciones en las cuales se tomaban grupos de nodos al azar respetando la cantidad de nodos que había que sacar. Se calculó el tamaño de la componente gigante, y luego de las 500 iteraciones se promedió para obtener el valor de la tabla, y se le tomó el valor de desviación estándar para obtener el error.

## 5. Esencialidad: Módulos Biológicos vs. Interacciones Esenciales

X.He propone que dos proteínas que forman una interaccion-proteína-prteína(PPI) esencial deben ser esenciales y, por el contrario, interacciones entre proteínas esenciales (IBEPS) pueden o no ser esenciales debido a que la esencialidad de una proteína puede deberse a otros factores. Esto permite estimar el número de PPIs esenciales en una red dado que el número de IBEPs aumenta con el número de PPIs esenciales. Para analizarlo generamos una red de control de manera aleatoria recableando los enlaces y manteniendo el grado de cada nodo, repitiendo esto 10000 veces se obtiene la distribución del número de IBEPs en la nueva red y se observa que ninguno de los 10000 valores es mayor que el número de IBEPs de la red real. Esto sugiere fuertemente un exceso de IBEPs en la red real. Suponiendo que el exceso de IBEPs se debe totalmente a las PPIs esenciales, He define $\alpha$ como la probabilidad de que un nodo sea esencial debido a la estructura de la red. En su red observa un mayor porcentaje de nodos esenciales que el dado por alfa, entonces supone que hay otros factores que pueden hacer a un nodo esencial, definiendo de esta forma a $\beta$ como la probabilidad de que un nodo sea esencial debido a lo que llama otros factores.

Define entonces la probabilidad de un nodo de ser esencial dado su grado:

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

donde $\alpha$ y $\beta$ son las probabilidades antes nombradas y $k$ el grado del nodo, pudiendo calcularse para cada valor de grado de la red. Un detalle a tener en cuenta es que no utiliza nodos con grado mayor a 10 debido a su escasez, lo que los hace estadisticamente insuficientes. La ecuación anterior puede escribirse como

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

que es la que nosotros vamos a graficar para cada red.

Zotenko analiza lo propuesto por He: nota que de asumir el modelo de interaccion proteína esencial se deduce entonces que si dos proteinas no interactuan entre si, la esencialidad de una proteína del par no depende de la esencialidad de la otra preoteína. Más aún, esta independencia debería verse también cuando las preoteínas comparten vecinos interactuantes. Para probar si esto se da en la red real lo que hace es contar el número de pares de proteínas no-adyacentes con tres o más vecinos que sean ambos esenciales o no-esenciales(es decir, del mismo *tipo*) y compararlos con el número de pares esperados según el modelo de He.

Replicamos la figura 2B del papper de He y con ella obtenemos los valores de $\alpha$ y $\beta$ para cada una de nuestras redes. Luego repetimos la cuenta de pares hecha por Zotenko y evaluamos entonces el modelo de He para nuestras redes.

Para obtener la probabilidad de que un nodo sea esencial según su grado dividimos el número de nodos esenciales de la red entre el número de nodos totales para cada valor de grado, de esta forma graficamos todos los puntos obtenidos hasta grado aproximadamente diez en cada red, ajustamos esos datos por una función lineal del grado y obtenemos luego los valores de alfa y beta correspondientes. A continuación: las figuras correspondientes a las redes LIT, AP-MS, Y2H y LITR, debajo: tabla con los valores de $\alpha$ y $\beta$ según red.

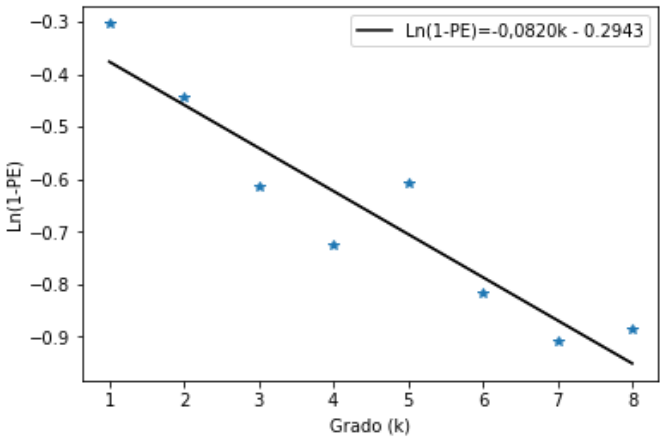


Figura 3: Relación entre la probabilidad de un nodo de ser esencial y su grado, red LIT

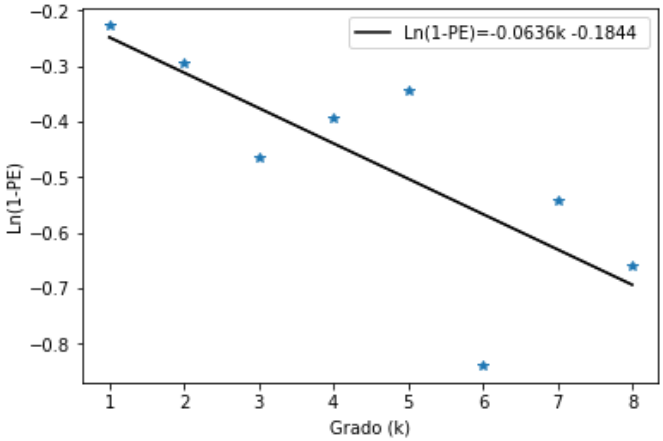


Figura 4: Relación entre la probabilidad de un nodo de ser esencial y su grado, red AP-MS

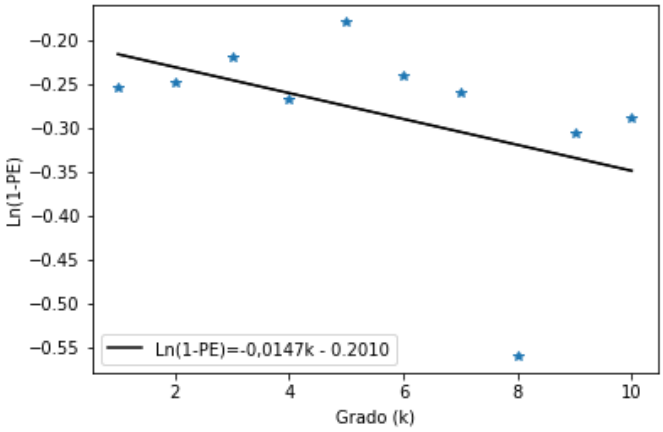


Figura 5: Relación entre la probabilidad de un nodo de ser esencial y su grado, red Y2H

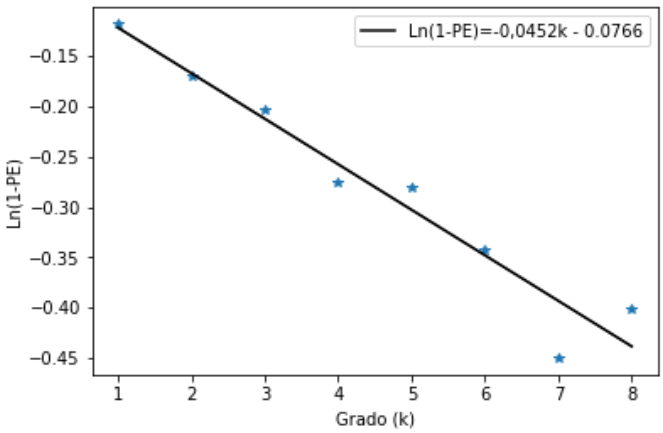


Figura 6: Relación entre la probabilidad de un nodo de ser esencial y su grado, red LIT Reguly

| Red | $\alpha$ | $\beta$ |
|---|---|---|
| LIT | 7,88 % | 25,49 % |
| AP-MS | 6,16 % | 16,84 % |
| Y2H | 1,46 % | 18,21 % |
| LITR | 4,42 % | 7,37 % |

Tabla 4: Valor de $\alpha$ y $\beta$ para cada red

Seguido a esto contamos la cantidad de pares de nodos no adyacentes con diferente cantidad de vecinos en común para cada red, también la cantidad de estos pares que son del mismo tipo y comparamos esta última cantidad con la esperada por el modelo dada la ecuación 1, los resultados los mostramos en la siguientes tablas:

| | Número total de Pares | Número de pares del mismo tipo | Número de pares del mismo tipo según el modelo |
|---|---|---|---|
| LIT | 10263 | 5985 | 5207 |
| AP-MS | 26180 | 13199 | 16092 |
| Y2H | 23079 | 15087 | 14377 |
| LIT Reguly | 220177 | 127596 | 120104 |

Tabla 5: Diferencia entre lo obsrvado para cada una de las redes y lo esperado por el modelo de He considerando pares que compartan 1 o más vecinos

| | Número total de Pares | Número de pares del mismo tipo | Número de pares del mismo tipo según el modelo |
|---|---|---|---|
| LIT | 1858 | 1047 | 975 |
| AP-MS | 15467 | 7740 | 10449 |
| Y2H | 2258 | 1514 | 1327 |
| LIT Reguly | 43027 | 25898 | 22890 |

Tabla 6: Diferencia entre lo obsrvado para cada una de las redes y lo esperado por el modelo de He considerando pares que compartan 2 o más vecinos

| | Número total de Pares | Número de pares del mismo tipo | Número de pares del mismo tipo según el modelo |
|---|---|---|---|
| LIT | 730 | 389 | 394 |
| AP-MS | 11613 | 5907 | 8226 |
| Y2H | 522 | 352 | 295 |
| LIT Reguly | 10777 | 6187 | 5604 |

Tabla 7: Diferencia entre lo obsrvado para cada una de las redes y lo esperado por el modelo de He considerando pares que compartan 3 o más vecinos

| | Número total de Pares | Número de pares del mismo tipo | Número de pares del mismo tipo según el modelo |
|---|---|---|---|
| LIT | 383 | 195 | 214 |
| AP-MS | 9314 | 4793 | 6828 |
| Y2H | 185 | 121 | 101 |
| LIT Reguly | 5342 | 3065 | 2845 |

Tabla 8: Diferencia entre lo obsrvado para cada una de las redes y lo esperado por el modelo de He considerando pares que compartan 4 o más vecinos

Vemos como en las redes Y2H y LIT Reguly los valores del ajuste estan siempre por debajo del real, esto puede ser debido a que son las que, respecto de ambos valores $\alpha$ y $\beta$ de las demás, presentan menores valores, mientras que para la red LIT esto sucede tomando pares con 1 y 2 vecinos en común, pero al tomar más (3 y 4 vecinos en común) el ajuste si bien da por encima, no es mucho y podría considerarse dentro de los errores al obtener $\alpha$ y $\beta$ del ajuste, que para esta red son los mayores. Comparamos las redes LIT Reguly con AP-MS que son las que mayor grado medio, cantidad de enlaces y clústering medio tienen: si bien tomando pares que compartan 1 y 2 vecinos se obtienen valores muy superiores para la red LITR, a medida que aumentamos el número de vecinos compartidos la red AP0MS iguala y sobrepasa a la LIT Reguly, suponemos por el mayor grado medio y clústering que tiene la APMS, lo que hace que haya nodos de alto grado más conectados entre si, entonces con más posibilidades de conectarse entre pares del mismo tipo. Finalmente comparamos las redes LIT con Y2H dado que ambas tienen número de enlaces y grado medio similares, vemos que considerando las dos primeras tablas los valores obtenidos son similares y luego, en las dos siguientes, con la condición de compartir 3 y 4 o más vecinos, dada la baja conectividad de la red Y2H, el número de pares de esta última se reduce sustancialmente respecto de la otra.

Con todo esto no podemos concluir la veracidad o falsedad del modelo de He, al menos para todas las redes. Sí podríamos descartarlo, dada las diferencias entre el modelo y lo medido, considerando pares que compartan uno, dos,tres y cuatro vecinos, en las redes Y2H y LIT Reguly.

# 6. Bibliografía

[1]: https://www.nature.com/articles/35075138
[2]:https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020088
[3]:https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000140