Out[1]: Click here to toggle on/off the raw code.

1 Introducción a Redes Complejas en Biología de Sistemas

1.1 Trabajo computacional 2

1.1.1 Introducción

En 2001, Jeong et al. observaron que en diversas redes de proteínas de *Saccharomyces cerevisiae* los nodos con mayor número de conexiones tienen una alta tendencia a ser **esenciales**. Una mutación que provoque la pérdida de función de estas proteínas redunda en la muerte o incapacidad reproductiva del organismo. Desde entonces, este fenómeno ha sido descripto en muchos otros seres vivos, y los esfuerzos por explicar por qué tiene lugar han sido varios.

En una primera aproximación, el mismo Jeong postuló que los hubs (nodos de alto grado) son esenciales porque **mantienen la conectividad de la red**. He et al., por su parte, hipotetizaron en 2006 que lo verdaderamente esencial no son las entidades proteicas, sino **las interacciones entre ellas**. En este trabajo se postuló además que estas interacciones esenciales estaban uniformemente distribuidas en la red, y que así **la probabilidad de que dos proteínas no interconectadas sean esenciales era disjunta**.

En 2008, Zotenko et al. refutaron esta hipótesis al demostrar que en varias redes de distinto origen y en forma consistente, la probabilidad de que una proteína sea esencial depende de la esencialidad de proteínas no vecinas con las que comparte un número elevado de vecinos en común. Surgió así la idea de que **la esencialidad debe ser estudiada en un nivel**

organizacional superior, ya que **los nodos esenciales lo son porque su disrrupción provoca el malfuncionamiento de complejos proteicos que cumplen funciones biológicas esenciales**. De esta forma, la remoción individual de proteínas que no tienen conexión entre sí pero forman parte del mismo complejo puede alterar la misma función necesaria para la vida o la reproducción.

En el presente trabajo práctico se reproducirán los principales resultados y las principales conclusiones de los trabajos mencionados, utilizando cuatro redes de diverso origen descriptas en el apartado que sigue.

1.1.2 Características de las redes analizadas

Se utilizaron cuatro redes de interacción proteína-proteína de *Saccharomyces cerevisiae*:

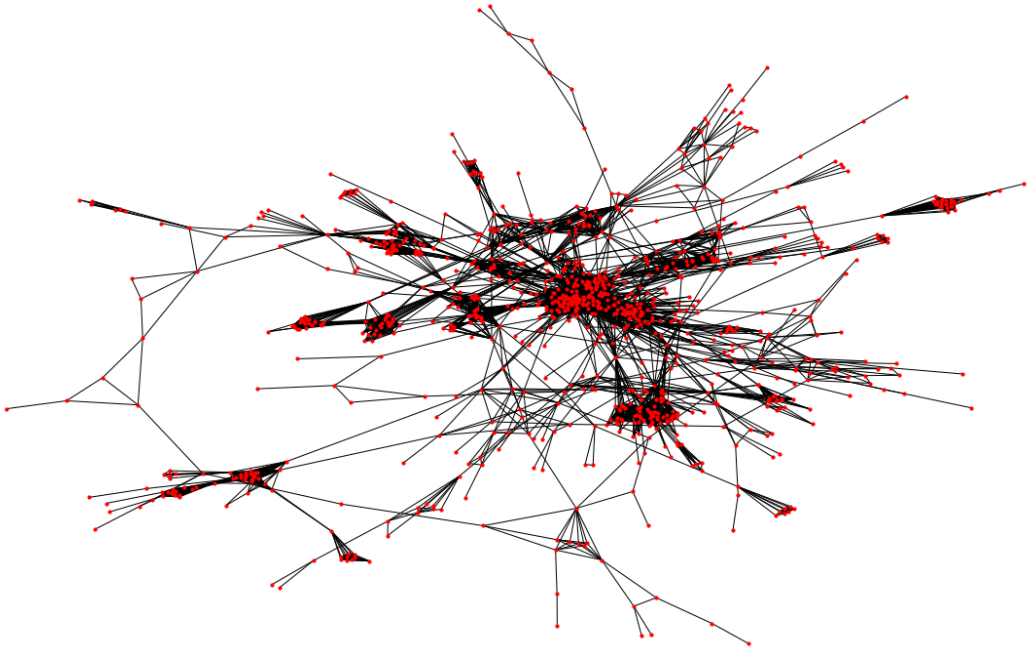
1. **AP_MS**: Constuída a partir de ensayos de coinmunoprecipitación. Utilizando anticuerpos específicos, se forman precipitados proteicos que se secuencian luego por espectrofotometría de masas. Las interacciones están sobreestimadas, ya que como la conectividad en cada complejo es desconocida, todas las proteínas coprecipitadas se reportan en forma de cliqué. Así, las interacciones reportadas correlacionan bien con la pertenencia a un mismo complejo proteico.
2. **Y2H**: Ensamblada a partir de ensayos de doble híbrido, por lo que reporta interacciones pareadas reales. Su desventaja principal es que no todas las interacciones posibles fueron testeadas, por lo que tiende a subestimar las interacciones reales y a estar enrriquecida en entidades de importancia científica previa.
3. **LITR**: Armada a partir de *text-mining* de trabajos científicos. Como la presencia consistente en publicaciones puede tener otras causas que la interacción física entre los nodos, este método tiende a sobreestimar la cantidad de interacciones.
4. **LIT**: Similar a LITR, pero producto de un curado manual posterior.

Se utilizó la lista de 1156 proteínas esenciales provista por el paper de He et al. A

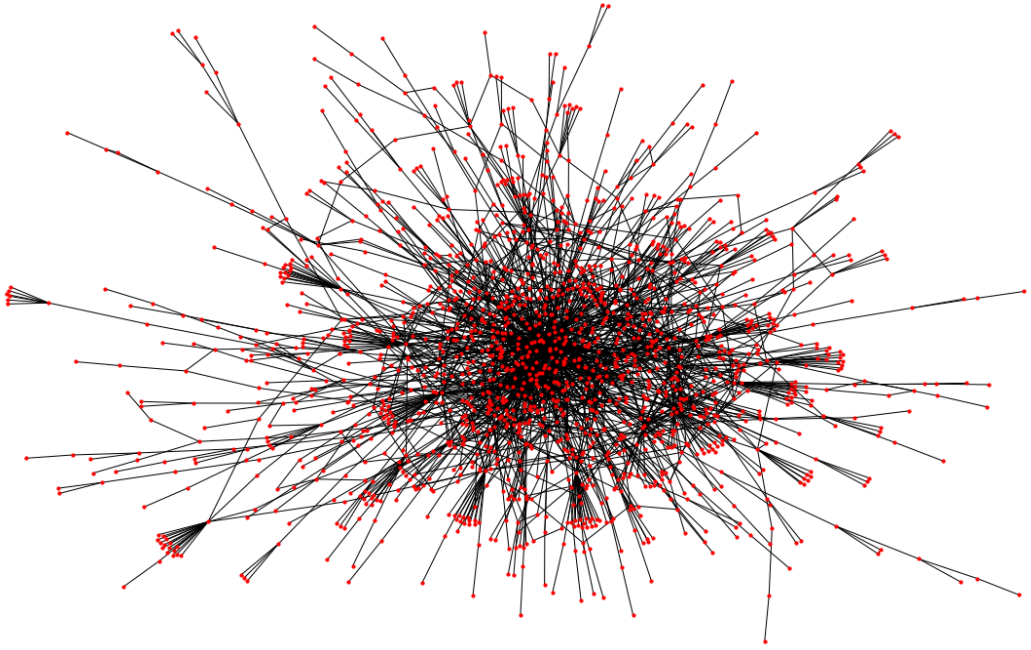
continuación se muestra un *layout* de la componente gigante de cada una de las cuatro redes mencionadas, seguida de una tabla con sus principales características. Notar que, como en la red AP_MS las unidades de construcción corresponden a complejos proteicos con un alto grado de conectividad, el coeficiente de *clustering* es muy alto.

```
Cantidad de proteinas esenciales: 1156
```

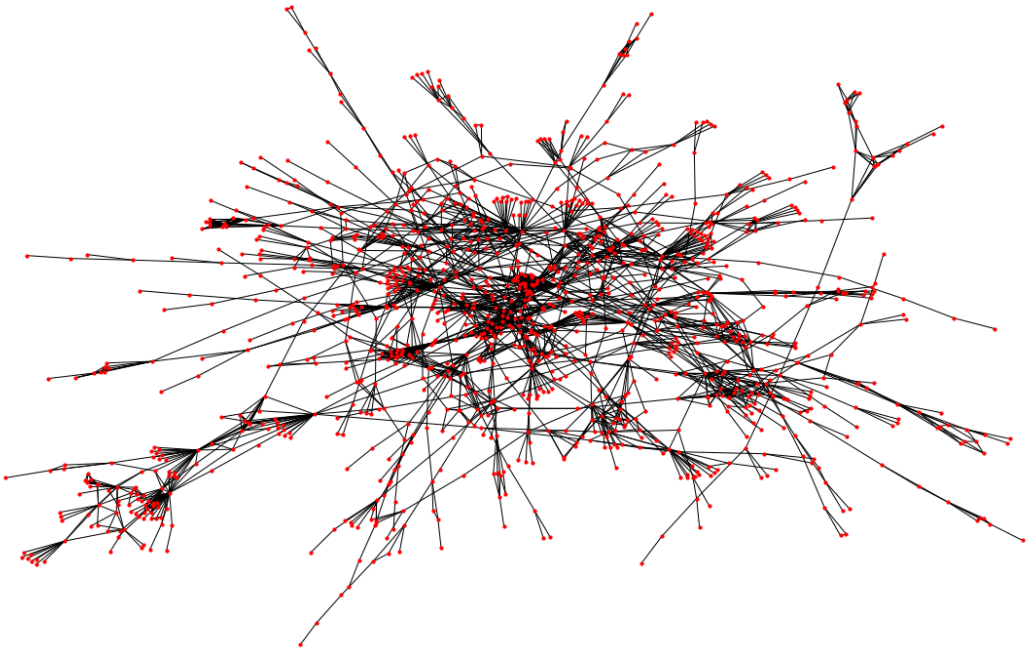
```
Nodos: 1622. Aristas: 9070. Esenciales: 607
```



```
Nodos: 2018. Aristas: 2930. Esenciales: 451
```



Nodos: 1536. Aristas: 2925. Esenciales: 625



Nodos: 3307. Aristas: 11858. Esenciales: 896

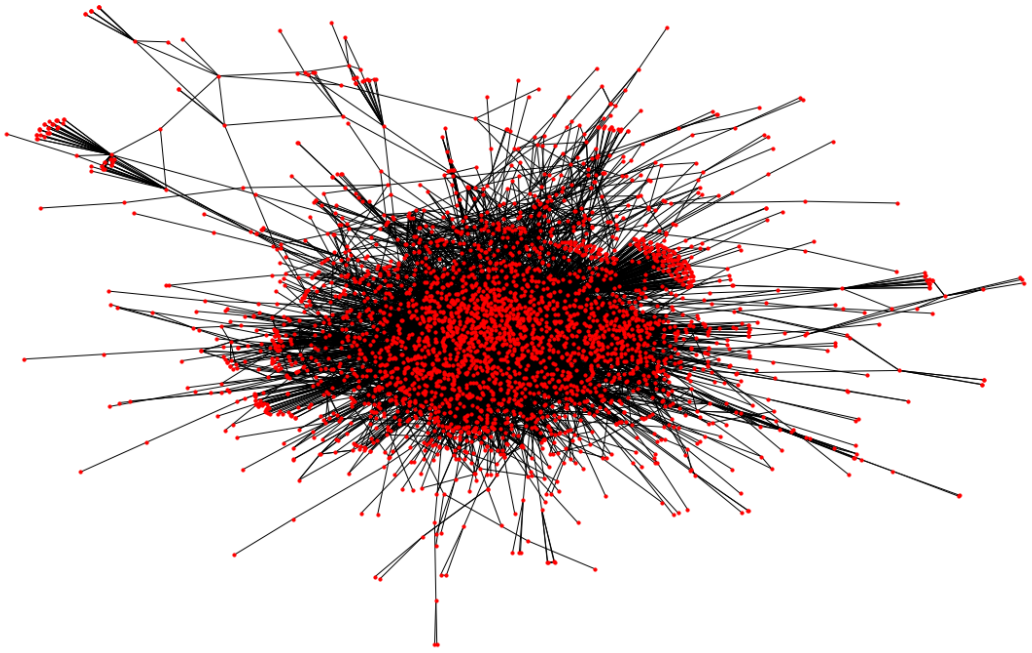


Tabla 1 de Zotenko et. al. (2008)

Out[9]:

	Nodos	Aristas	Gradomedio	Clustering
AP_MS	1622	9070	11.184	0.554636
Y2H	2018	2930	2.904	0.046194
LIT	1536	2925	3.809	0.292492
LITR	3307	11858	7.171	0.261134

Como las redes empleadas provienen de distinto origen y sus interacciones no representan lo mismo, se procedió a cuantificar el solapamiento de las interacciones presentes entre ellas

(de todos los pares de nodos posibles para cada red, se contaron aquellos presentes en cada una de las otras).

$$overlap = pe_ij/pe_i$$

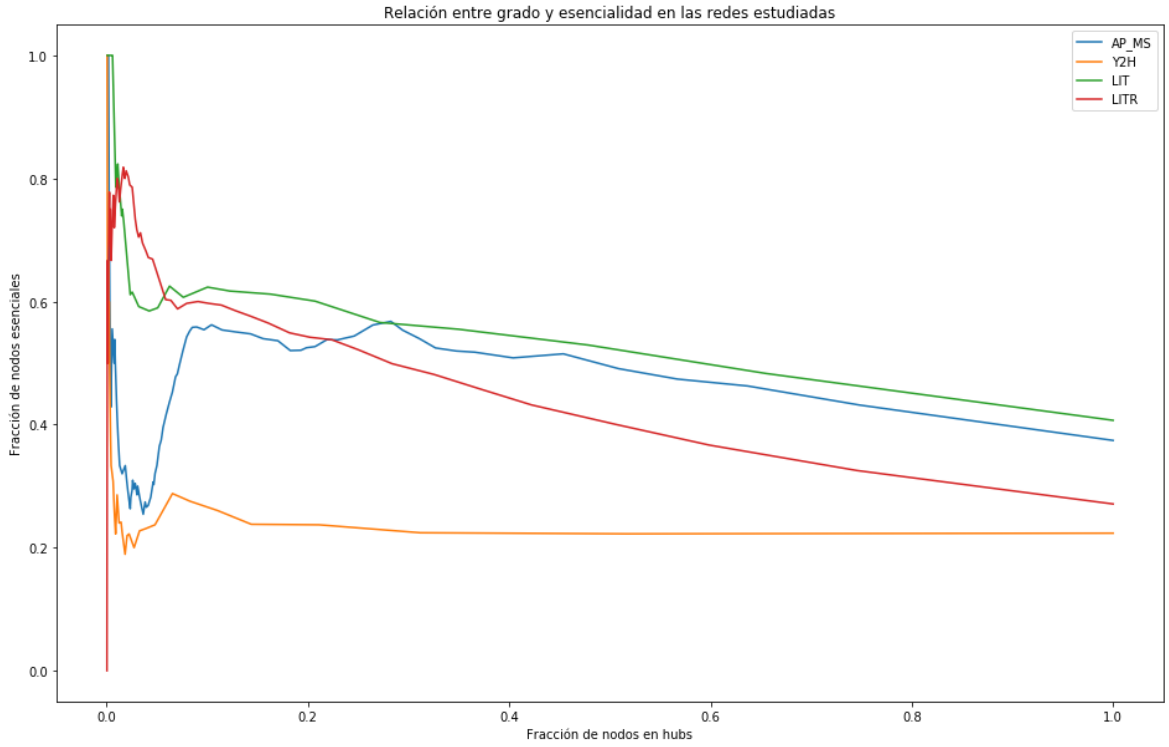
donde pe_i es la cantidad de interacciones de la red i presentes en la red j, y pe_i es la cantidad total de interacciones en la red j.

Out[10]:

	AP_MS	LIT	LITR	Y2H
AP_MS	1.00	0.57	0.88	0.45
LIT	0.60	1.00	0.99	0.48
LITR	0.43	0.46	1.00	0.40
Y2H	0.36	0.37	0.66	1.00

Las filas corresponden a las redes de referencia, y las columnas a las redes con las cuales se está comparando. Así, por ejemplo, el valor correspondiente al cruce entre la segunda fila y la tercera columna corresponde a la fracción de interacciones de LIT presentes además en LITR. Notar en este caso particular que como la primera es una versión revisada de la segunda, el solapamiento es casi total. El caso inverso, sin embargo (fila 3, columna 2) muestra un solapamiento drásticamente menor de tan sólo el 46%.

Se procedió entonces a estudiar la ley de centralidad-letalidad en cada una de las redes. Para esto, se definieron como hubs aquellas proteínas que presentan un grado igual o mayor a un valor k. Para cada número de k se calcularon la fracción del total de nodos que cumplen con la definición de hubs, y la fracción de nodos esenciales pertenecientes a estos hubs.



Así, valores del eje x cercanos a cero corresponden a una baja fracción de hubs y por ende a un valor de k alto. Se observa que los nodos de grado alto (izquierda del gráfico) corresponden en gran proporción a proteínas esenciales. La tendencia se observa para las cuatro redes, aunque es claramente más débil en Y2H. En el marco que presentan Zotenco et al, esto podría tener que ver con que, como en esta red las interacciones corresponden a contactos físicos reales las interacciones dentro de cada complejo son muchas menos y el grado de las proteínas esenciales tiende al ser más bajo.

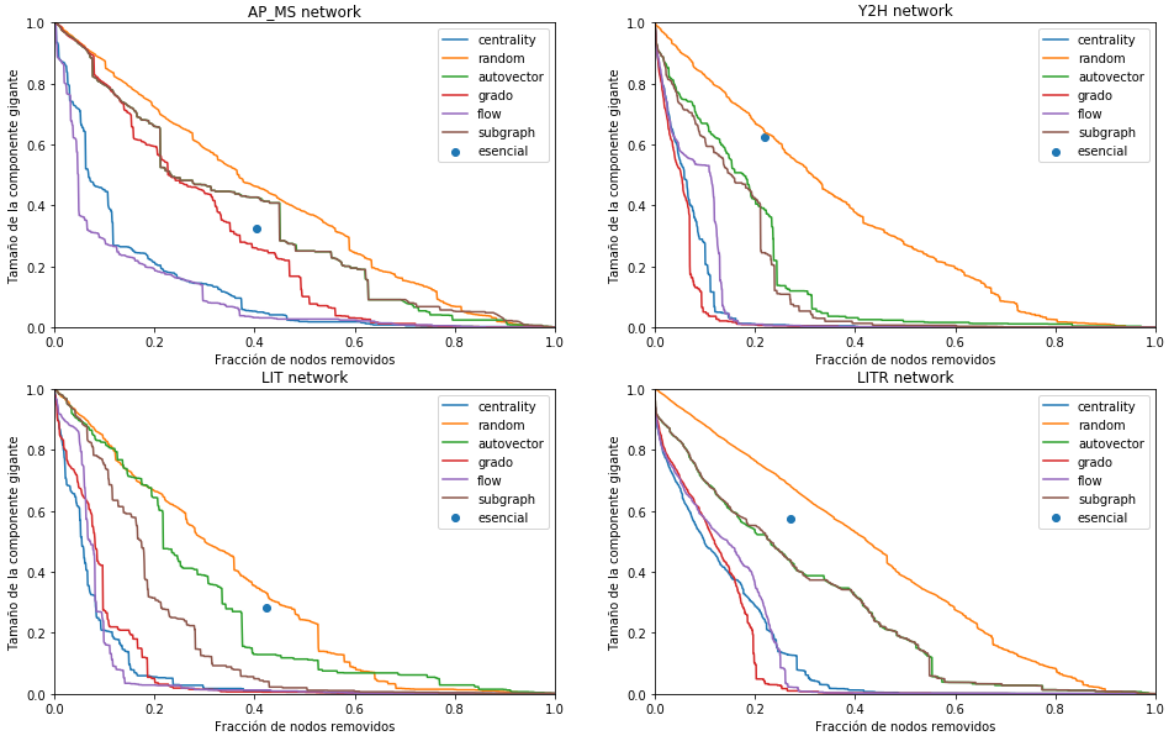
1.1.3 Análisis de vulnerabilidad

Los índices de centralidad asignan valores a los nodos de la red, como forma de cuantificar características topológicas. En este análisis trabajaremos con dos tipos de índices de

centralidad distintos. El primero corresponde a índices de centralidad local, donde el valor de centralidad está directamente relacionado con el grado de los nodos. El segundo hace referencia a la importancia del nodo en mantener la conectividad entre pares de nodos de la red, índice conocido como betweenness.

Para estudiar el papel de los hubs en las redes de interacción de proteínas estudiadas se compararon índices de centralidad local (degree centrality, eigenvector centrality y subgraph centrality), con indices de betweenness (shortest-path betweenness centrality (SPBC) y current-flow betweenness centrality (CFC)).

Así se comparó cuan vulnerables son las redes al desarmar por los distintos criterios, midiendo el tamaño de su componente principal al ir removiendo los nodos por valor de centralidad decreciente.



Como era de esperar, las redes son más vulnerables al desarmado por índices de betweenness que a los índices de centralidad local, con excepción del índice de grado para las redes Y2H, LIT y LITR que las disrumpe de manera similar a los índices de betweenness. Esto podria deberse a que, como AP_MS esta más fuertemente conectada (mayor grado promedio) aunque se quiten los hubs, la componente esta conectada por varios caminos alternativos. Por otro lado, en las otras 3 redes, los hubs estan exhibiendo un elevado betweenness, por lo que los caminos importantes se dan a través de ellos (no hay tantos caminos alternativos).

Además, se observa que todos los índices son más eficientes a la hora de desarmar las redes que el criterio de desarmado por nodos random. También se graficó el tamaño de la componente principal de las redes al remover todos los nodos esenciales y se observa que siempre los criterios de betweenness son más eficientes para desmontar las redes a esa fracción de nodos dados o que la fracción de nodos requerida para lograr dicho tamaño de la componente es mucho menor para los índices de betweenness. En conclusión, los nodos esenciales no tienen por qué ser nodos importantes para la conectividad global de las redes.

Teniendo en cuenta lo anteriormente dicho se analizó si los hubs o nodos esenciales suelen ser más disruptivos para las redes que nodos del mismo grado que no son esenciales. Para esto se comparó el tamaño de la componente principal de las redes al remover todos los nodos esenciales con el tamaño de la componente principal de las redes al remover la misma cantidad de nodos no esenciales pero respetando la distribución de grado.Estos resultados se observan en la siguiente tabla.

Out[23]:

	Essential	Random nonesential mean	Random nonesential sd	P-valor
AP_MS	0.323705	0.421643	0.013821	6.883373e-13

	Essential	Random nonesential mean	Random nonesential sd	P-valor
Y2H	0.624165	0.625025	0.011724	4.707709e-01
LIT	0.281121	0.414303	0.003506	0.000000e+00
LITR	0.575062	0.581789	0.003850	4.031312e-02

Tanto Y2H como LITR valores muy similares de fracción de componente principal para las dos formas de desarmado. Lo que daría la pauta de que los nodos esenciales no nos más importantes que los no esenciales a la hora de mantener la conectividad global de las redes. Sin embargo, el las otras dos redes (AP_MS y LIT) pareciera ser más disruptivo desarmar por nodos esenciales que por no esenciales respetando las distribuciones de grado. Creemos que este fenómeno podría deberse a un error sistemático que se da cuando hay más nodos esenciales de un dado grado que no esenciales (fenómeno que ocurre sobre todo con los grados más elevados) y que por default al desarmar la red por nodos no esenciales se toman nodos de menor grado (en nuestra implementación). Con lo cual resulta razonable que la disrupción en dichas redes sea menor.

1.1.4 Esencialidad: Módulos biológicos vs. Interacciones Esenciales

Como se explicó en la introducción, Zotenko et al cuestiona la validez de la hipótesis sobre la importancia de las interacciones esenciales de He. A continuacion se estimaran los parámetros de He y se compararan con la metodología de Zotenko, que intenta cambiar el foco de esencialidad de las inteacciones a los complejos protéicos. Es preciso norar que He et al. aclaran que su modelo puede no funcionar bien en redes en las que las interacciones no representen contactos físicos reales, como es el caso de AP_MS. En este trabajo práctico se incluyó a esta red en el análisis de todos modos, a fin de reportar posibles anormalidades en los resultados.

Los parámetros de He, alfa y beta, representan la probabilidad de que un enlace sea esencial y que una proteina sea esencial independientemente de la topología, respectivamente. Entonces la probabilidad de que un nodo sea esencial es:

$$Pe = 1 - (1 - \alpha)^k . (1 - \beta)$$

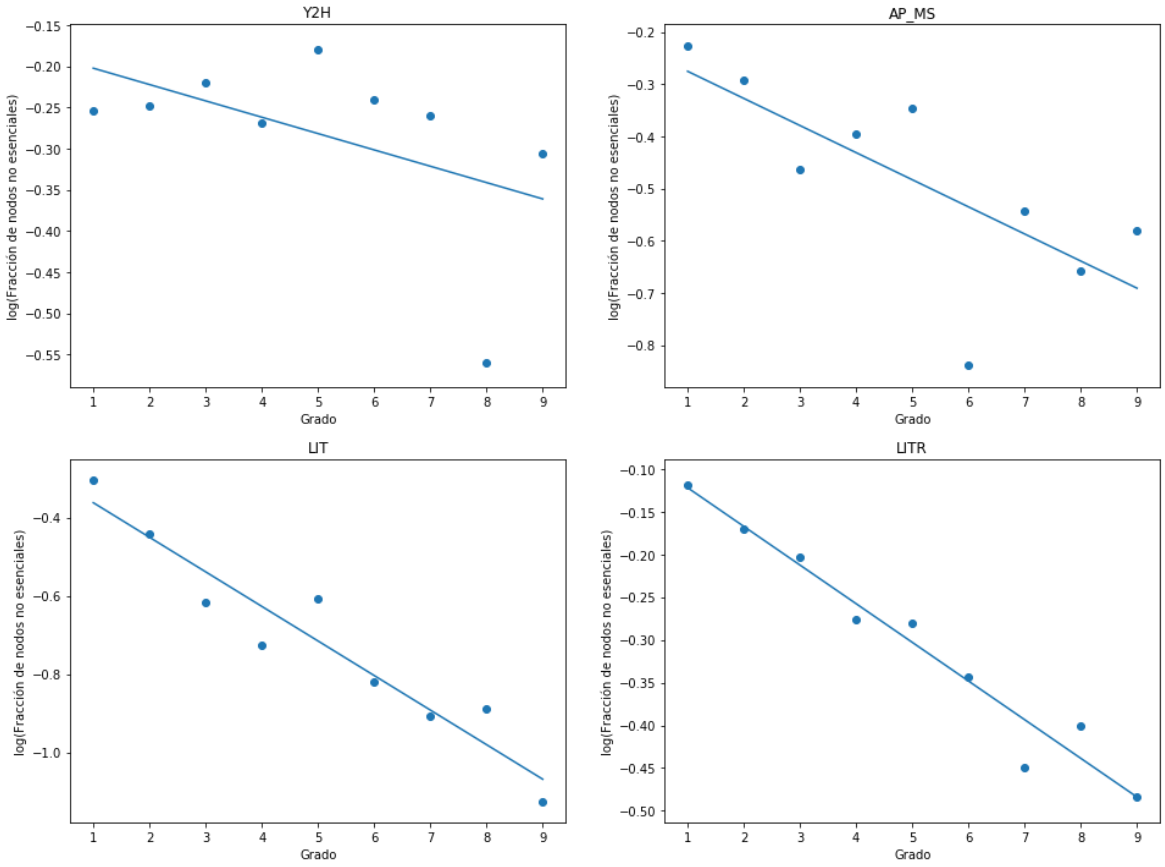
1.1.4.1 Cálculo de alfa y beta por regresión lineal

Linealizando la expresión en cuestión, pueden estimarse alpha y beta a partir de los valores obsevados mediante una regresión lineal:

$$ln(1 - Pe) = k . ln(1 - \alpha) + ln(1 - \beta)$$

Donde k es el grado de un nodo y Pe es la proporción de nodos esenciales de ese grado. Se pone un corte en grado 10 (al igual que en He. et al) ya que para mayores grados no se tiene suficiente cantidad de nodos para obtener una proporción confiable. Por ejemplo los valores empiezan a saltar a 1 o a 0.

Out[28]: `[<matplotlib.lines.Line2D at 0x1d4a50c0940>]`



1.1.4.2 Cálculo de alfa y beta por métodos de simulación

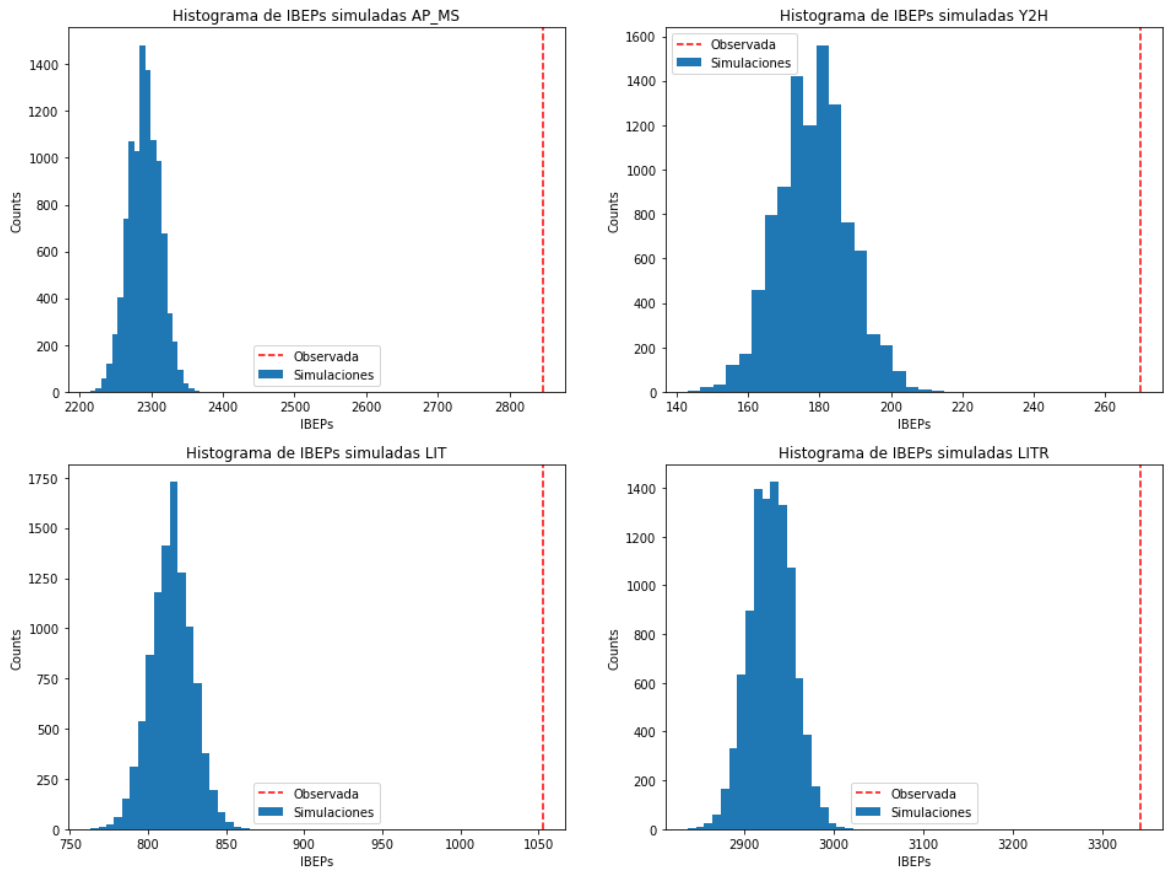
A parte de calcular alfa y beta ajustando los valores observados, se procedieron a hacer la sigientes simulaciones:

1) Para calcular alfa se simularon 10000 redes, partiendo de la red inicial. En cada una se recablearon 5000 aristas, manteniendo el grado de los nodos y se midió la cantidad de IBEPs (interacciones entre proteínas esenciales). Al igual que en He et al., en ningún caso la cantidad de IBEPs simulada fue superior a la de la red original. Por ello se tomo la hipótesis de que la diferencia corresponde a las interacciones esenciales de la red:

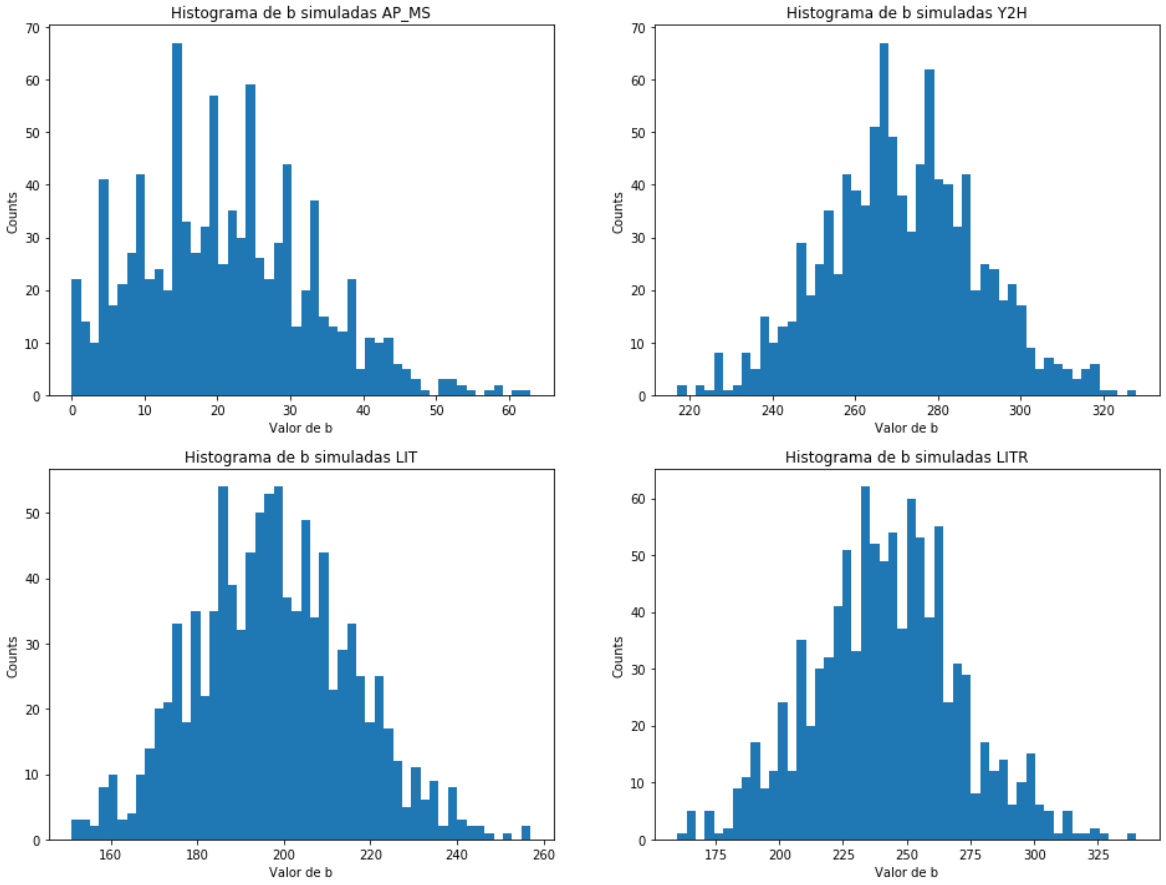
$$\alpha = IBEPs(org) - \mu(IBEPs(simulado))/\#aristas$$

2) Por otro lado, para estimar beta, se simularon 1000 redes, partiendo de la red inicial. En cada una, se asignaron n interacciones esenciales (siguiendo la distribución obtenida de las simulaciones de alfa) y a partir de esas asignaciones se definieron los nodos esenciales (que contenian esas aristas). Luego, para alcanzar la cantidad de nodos esenciales original, se asigaba un nodo como esencial al azar (de todos los nodos, no solo de los no esenciales). Finalmente se define beta como la cantidad de nodos que tuvieron que agregarse para llegar de los nodos esenciales asignados a partir de las interacciones esenciales a la cantidad original.

$$\beta = (nodos\_esenciales\_red\_org - nodos\_esenciales\_por\_inteaccion\_esencial)/\#nodos$$



Se observa en la figura superior la distribución de IBEPs obtenidas a partir de 10000 simulaciones para cada red. En concordancia con He et al. como en ningún caso la probabilidad de que el valor real de IBEPs (línea vertical roja) provenga de la distribución simulada es alta, se asume que la diferencia corresponde a las interacciones esenciales en la red.



Se observa en la figura de arriba la distribución en la cantidad de nodos esenciales estimados que no dependen de los enlaces esenciales (definida como b, que no es lo mismo que beta).

En la siguiente tabla pueden compararse los valores de alfa y beta obtenidos por ambos métodos:

Out[52]:

	Alfa (ajuste lineal)	Beta (ajuste lineal)	Alfa (simulación)	Beta (simulación)

	Alfa (ajuste lineal)	Beta (ajuste lineal)	Alfa (simulación)	Beta (simulación)
AP_MS	0.051+/-0.017	0.200+/-0.018	0.0613+/-0.0024	0.013+/-0.007
Y2H	0.020+/-0.013	0.167+/-0.015	0.0101+/-0.0011	0.134+/-0.009
LIT	0.084+/-0.009	0.239+/-0.021	0.0262+/-0.0015	0.129+/-0.012
LITR	0.0443+/-0.0035	0.074+/-0.007	0.0455+/-0.0027	0.073+/-0.009

De esta tabla pueden concluirse dos cosas: los valores para los coeficientes obtenidos por ambos métodos son comparables, y los valores de beta tienden a ser más grandes que los de alfa. Esto último (que se cumple en todos los casos salvo en las simulaciones de AP_MS, excluidas del análisis como fue mencionado más arriba) remite a que la probabilidad de que una proteína sea esencial por razones no relacionadas con la participación en una interacción esencial suelen ser mayores.

1.1.4.3 Comparación de pares del mismo tipo reales versus predichos por los modelos

Como se mencionó antes, estos modelos asumen que la probabilidad de dos proteínas no conectadas de ser esenciales es disjunta. La probabilidad de que un par de proteínas sean esenciales, entonces, puede calcularse como la probabilidad de una de ser esencial por la probabilidad de la otra.

$$Pe_{i,j} = Pe_i . Pe_j$$

donde Pei es la probabilidad de que la proteína i sea esencial, Pej es la probabilidad de que la proteína j sea esencial y Pei,j es la probabilidad de esencialidad conjunta del par i,j.

Como también se mencionó, según Zotenko et al. la esencialidad es un fenómeno que tiene lugar a otra escala: las proteínas esenciales son aquellas necesarias para el funcionamiento de complejos esenciales. Según este razonamiento, dos proteínas que pertenezcan al mismo

complejo deberían tener mayor probabilidad de ser esenciales independientemente de si están conectadas o no.

Teniendo esto en mente, se estudió el poder predictivo de los modelos propuestos por He para pares de proteínas no conectados que compartieran tres o más vecinos.

Se contaron los pares de nodos totales que cumplieran con esta condición en las cuatro redes bajo estudio, y se delimitó cuántos de ellos están compuestos por proteínas del mismo tipo (ambas esenciales o no esenciales). Luego, se calculó la probabilidad de esencialidad o no esencialidad de cada par utilizando los alfa y beta provenientes de cada uno de los enfoques descriptos más arriba. A partir de estas probabilidades se estimó el número de pares del mismo tipo esperados para cada red según ambas variantes del modelo de He, como la sumatoria de probabilidades de todos los pares de la red de estar compuestos por dos proteínas esenciales o dos proteínas no esenciales.

Sea n la cantidad de pares en la red que cumplen con las condiciones estipuladas, la cantidad de pares esperados se calcula como

$$Pares\_esperados = \sum_{i=1}^{n} Pe_i + Pne_i$$

donde Pei es la probabilidad del par i de contener dos proteínas esenciales, y Pnei es la probabilidad del par i de contener dos proteínas no esenciales.

Las incertezas reportadas surgen de la propagación de los errores propios de los alfa y beta utilizados en cada caso. Para la regresión lineal, los mismos surgen a partir de la pendiente y ordenada al origen estimados; para las simulaciones, a partir del desvío estándar de las distribuciones obtenidas (la normalidad fue testeada mediante la prueba de Shapiro-Wilks). Los resultados se muestran en la siguiente tabla.

Out[396]:

	Número total de pares	Número de pares del mismo tipo	Simulación	Ajuste lineal	P-valor simulación	P-valor ajuste lineal
AP_MS	11613	5907	7722+/-125	7772+/-954	6.454696e-48	2.533794e-02
Y2H	23073	15087	15904+/-279	14075+/-1192	1.730440e-03	1.981463e-01
LIT	730	389	413+/-6	399+/-12	1.141190e-04	1.917992e-01
LITR	10777	6187	5657+/-56	5633+/-69	7.530339e-21	7.637835e-16

Como puede observarse, en todos los casos (si bien AP_MS no debe ser tenido en cuenta para el análisis por los motivos explicados más arriba) alguno de los dos métodos reporta diferencias significativas entre los números esperado y real de pares del mismo tipo en cada red.

Esto prueba que, al menos en estas redes, la probabilidad de dos proteínas no conectadas de ser esenciales **no es disjunta**, sino que depende en principio de su cercania en la red. Para indagar en esto más exhaustivamente, podría repetirse el análisis utilizando todos aquellos pares de proteínas no conectadas que no pertenezcan al mismo complejo biológico (utilizando anotación ontológica, por ejemplo) independientemente de su conectividad en una red en particular. Si la hipótesis de Zotenko et al. es cierta, el modelo de He debería aplicar.

1.1.5 Bibliografía

- He, Xionglei, and Jianzhi Zhang. 2006. “Why Do Hubs Tend to Be Essential in Protein Networks?” PLoS Genetics 2 (6): e88.
- Jeong, H., S. P. Mason, A. L. Barabási, and Z. N. Oltvai. 2001. “Lethality and Centrality in Protein Networks.” Nature 411 (6833): 41–42.
- Zotenko, Elena, Julian Mestre, Dianne P. O’Leary, and Teresa M. Przytycka. 2008. “Why Do Hubs in the Yeast Protein Interaction Network Tend to Be Essential: Reexamining the Connection between the Network Topology and Essentiality.” PLoS Computational Biology 4 (8): e1000140.