Trabajo Práctico Computacional 02

Centralidad - Letalidad

Sasha Smolarchik, Gastón Bujía, Mariano Nicolini

17 de Octubre del 2018

sasha95@gmail.com, gastonbujia@gmail.com, mariano.nicolini.91@gmail.com

1. Introducción

A lo largo de muchos trabajos de investigación, se han confeccionado redes de interacción entre proteínas, en donde los nodos representan una proteína determinada y se establece un enlace entre ellas si existe alguna evidencia experimental de que interactúan físicamente entre ellas. En particular, en los trabajos de Jeong et. al.(2001), He et. al.(2006) y Zotenko et. al. (2008), se estudiaron redes de interacción de proteínas de la levadura *Saccharomyces cerevisiae* y se analizaron las consecuencias fenotípicas de la remoción de ciertas proteínas con un determinado criterio, en los organismos.

Se observó que existe una correlación entre la remoción de una proteína altamente conectada con otras y la correspondiente letalidad del organismo. Es decir, la eliminación de nodos con grados altos de la red estaba más asociado a la muerte o a la incapacidad de reproducción de la levadura que la remoción de nodos con grados bajos. A este fenómeno se lo denomina la regla de *centralidad-letalidad*, y a dichos nodos se los denomina esenciales. Conocer las causas de este fenómeno resulta importante para poder establecer cómo las propiedades topológicas de la red se relacionan con las diferentes propiedades fenotípicas del sistema que se está estudiando.

El objetivo de este trabajo es estudiar esta relación subyacente entre la topología de la red y la funcionalidad de cada proteína, poniendo a prueba las distintas hipótesis postuladas en los trabajos citados para un caso particular de redes de interacciones proteicas.

2. Características de las redes analizadas

Características generales

Con el fin de evitar sesgos experimentales, resulta de interés analizar y comparar diferentes redes de interacción entre las proteínas del mismo organismo. Estas fueron reportadas bajo diferentes criterios, metodologías y grados de certeza. Las redes que se analizaron en este trabajo fueron las redes *AP-MS*, *Y2H*, *LIT* y *LIT-REGULY*. Éstas dos últimas son redes construidas a partir de interacciones reportadas en la literatura. En la tabla 1 se resumen las propiedades estructurales más importantes de las redes con las que se trabajó.

	Clustering medio	Enlaces	Grado medio	Nodos
Red AP-MS	0.554636	9070	11.183724	1622
Red Y2H	0.046194	2930	2.903865	2018
Red Literatura	0.292492	2925	3.808594	1536
Red Reguly	0.261134	11858	7.171454	3307

Tabla 1: Propiedades estructurales de las redes de interacción analizadas.

En la tabla 2, se muestra la fracción de los enlaces de una determinada red que están presentes en las otras redes. Es importante notar que esta tabla no tiene por qué ser simétrica ya que, para cada red está normalizada respecto de su propia cantidad de nodos. Dadas las diferencias experimentales y metodológicas ya mencionadas en la construcción de cada red, no es de sorprender la significancia en las diferencias en cuanto a las propiedades estructurales y topológicas entre las redes.

	Red AP-MS	Red Y2H	Red Literatura	Red Reguly
Red AP-MS	1.000000	0.028666	0.143109	0.277839
Red Y2H	0.088737	1.000000	0.087201	0.156314
Red Literatura	0.443761	0.087350	1.000000	0.963932
Red Reguly	0.212515	0.038624	0.237772	1.000000

Tabla 2: Cada fila corresponde a una determinada red y los distintos valores que allí aparecen muestran la fracción de sus enlaces que aparecen en las otras redes.

Hubs

Una característica que resulta de interés, es la correlación entre los grados de cada nodo y su esencialidad. Una manera de estudiar esta correlación es preguntarse si los *hubs*, esto es, los nodos con un grado mayor a un cierto grado umbral, son más propensos a ser esenciales que cualquier nodo aleatorio de la red.

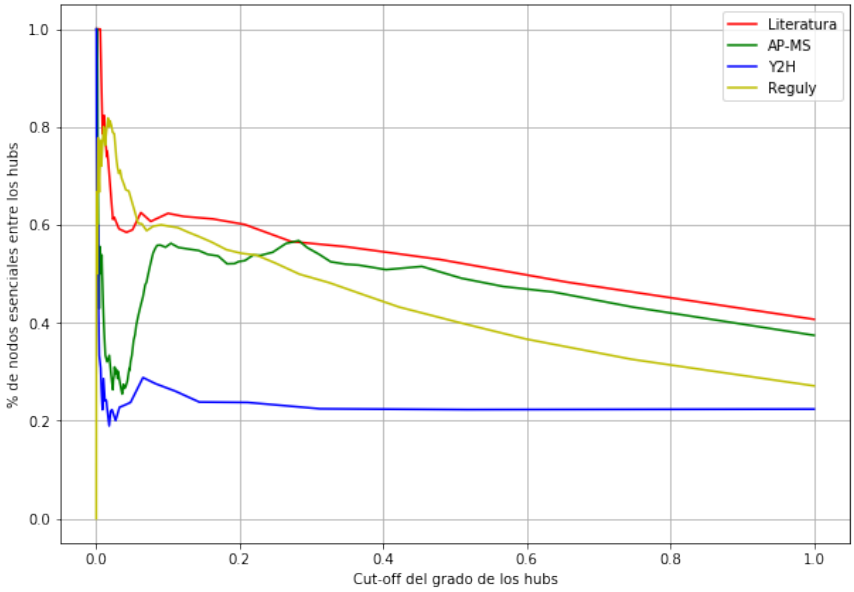


Figura 1: Se muestra la fracción de nodos esenciales entre los nodos de grado más alto, en función de la fracción de los nodos que fueron considerados como hubs, para cada red.

Con el objetivo de entender si existe esta correlación y elegir un umbral apropiado para cada red, resulta de utilidad observar el gráfico 1, donde se muestra la fracción de nodos esenciales que hay en el conjunto de nodos de grado mayor o igual a un cierto valor k en función del porcentaje de nodos que representan los mismos. El gráfico, parece sostener la esencialidad de los nodos de mayor grado como sostiene Jeoung et.al.(2001), donde la concentración de nodos esenciales es mucho mayor en los hubs para cortes muy altos que frente al resto de los nodos. Además, tal y como ocurre en el trabajo de Zotenko et. al., se puede ver que tomando al 20 % aproximadamente de los nodos con grados más altos sean considerados hubs resulta ser una elección razonable. Con excepción de la red *REGULY*, en donde claramente esto no se cumple, todas las redes interrumpen su crecimiento constante en la fracción de nodos esenciales respecto de los que son considerados hubs, alrededor de ese porcentaje. Estos resultados se condicen con la hipótesis planteada por Jeoung et. al. (2001), en la cual postula que los nodos con mayores interacciones (hubs) tienen mayor probabilidad de ser esenciales que los nodos con menos interacciones, donde postuló la ya mencionada regla de *Centralidad-Letalidad*.

3. Análisis de vulnerabilidad.

Aunque la medida de centralidad de grado nos permite entender el rol de manera local de un nodo (proteína) en la estructura de la red, es a veces muy informativo a gran escala como es el caso de redes *scale-free*. Sin embargo, no siempre nos dará suficiente información acerca del rol a gran escala de ese nodo en la red y para esto es necesario considerar diferentes criterios de centralidad con los cuales decidir qué tan importante es un determinado nodo en la topología de la red.

Con el fin de analizar el rol de los hubs y la vulnerabilidad de las redes respecto a la centralidad-esencialidad de sus nodos, resulta interesante, por un lado, analizar la manera en la que la eliminación de los nodos más centrales afecta en la desconexión de la red, y por otro lado, comparar como la eliminación de nodos esenciales contra la de nodos no esenciales influye sobre la conectividad de la red. Para el primer análisis, se utilizaron entonces diferentes criterios de centralidad locales, *degree centrality*, *eigenvector centrality* y *subgraph centrality* y por otro lado criterios de intermediación como el *shortest-path betweenness centrality* y *current flow betweenness centrality*. En la figura 2, se muestra el impacto de la remoción progresiva de los nodos más centrales, para cada red y según cada criterio de centralidad. Puede observarse que, en las redes *AP-MS* y *LIT*, la eliminación de los nodos más centrales bajo un criterio de intermediación es más efectivo en desconectar la red que la eliminación de nodos con centralidad local, como es de esperarse, mientras que en la red *Y2H* y en la red *REGULY*, esta regla no se cumple. En ésta última, puede apreciarse cómo el ordenamiento de los nodos de centralidad bajo *degree centrality* y *eigenvector centrality* es casi equivalente, viendo la superposición entre las curvas.

	Esenciales	Random no esenciales
APMS	0.323705	0.379465 ± 0.019438
Literatura	0.281121	0.417350 ± 0.004811
Y2H	0.624165	0.622151 ± 0.011863
Reguly	0.575062	0.540567 ± 0.004132

Tabla 3: La primer fila representa el tamaño relativo de la componente gigante luego de remover todos los nodos esenciales para cada red. La segunda columna, representa la distribución del tamaño relativo de la componente más grande luego de eliminar los nodos no esenciales de forma aleatoria, muestreado 1000 veces.

También resulta interesante observar que en general, la remoción aleatoria de nodos no esenciales, parece ser tan letal para la estructura de la red como la remoción de todos los nodos esenciales. Esto se puede observar mas claramente en la tabla 3, donde eliminamos de forma aleatoria nodos no esenciales como la cantidad total de nodos esenciales, y los muestreamos unas 1000 repeticiones para obtener una distribución de como varía el tamaño de la componente gigante en ambos casos. En la misma, no parece haber una diferencia demasiado significativa en ambos métodos.

Todo esto parece indicar que no hay una mayor importancia para los hubs esenciales en mantener la estructura de la red que para los hubs de nodos no esenciales. De esta manera, parece ser que el rol estructural de los hubs no

esta relacionado con la esencialidad de los nodos.

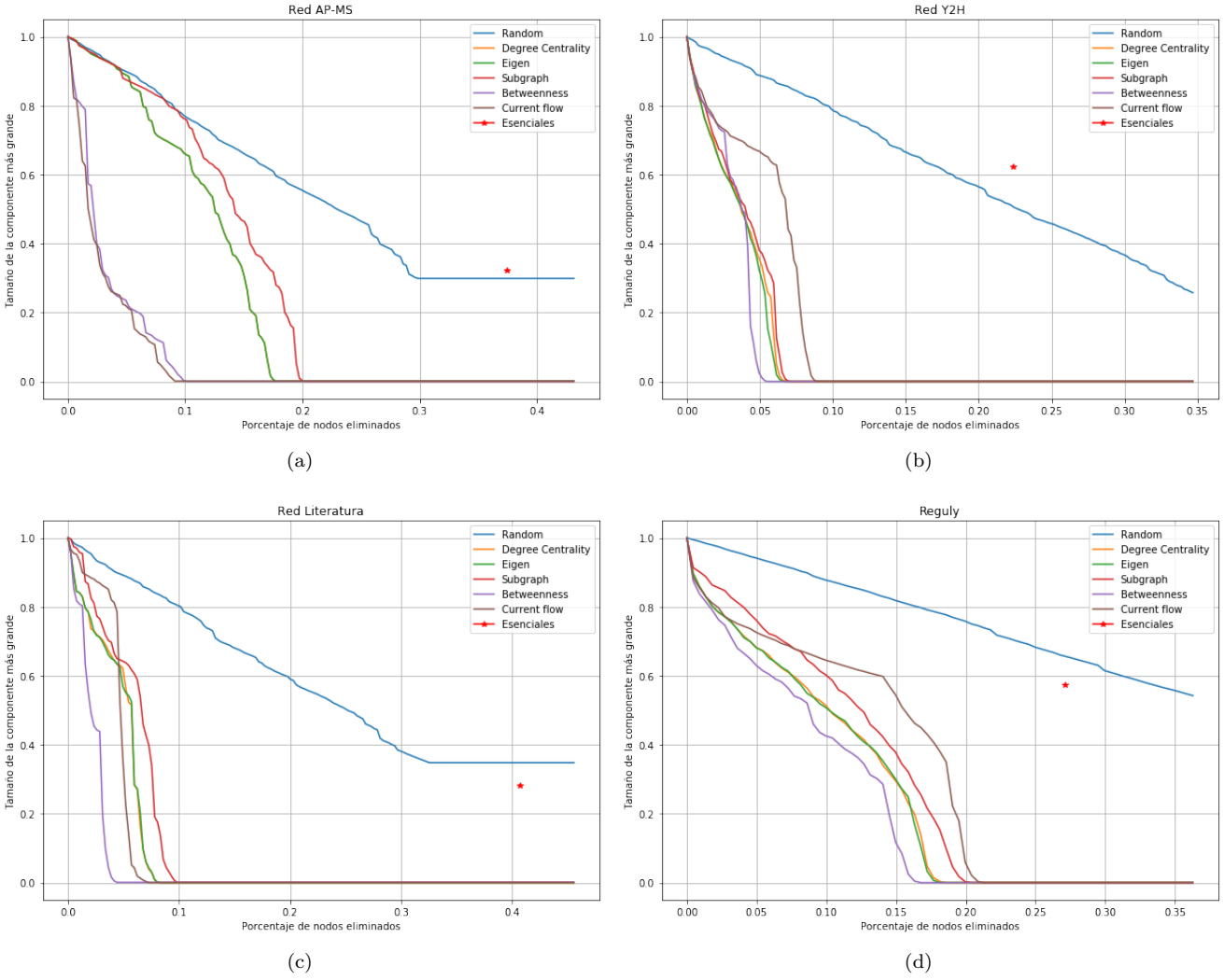


Figura 2: Se grafica el impacto sobre la componente gigante de la remoción progresiva de los nodos más centrales como el número de nodos en dicha componente, en función de la fracción de nodos eliminados. Cada curva corresponde a la remoción de nodos bajo un criterio de centralidad diferente, además de una curva que corresponde a la eliminación aleatoria de nodos.

4. Esencialidad: Módulos biológicos vs Interacciones esenciales.

Habiendo ya descartado la posibilidad de que el rol estructural de los hubs se corresponda a su esencialidad funcional en la red de proteínas, se pueden considerar las dos hipótesis presentadas en los trabajos de He et. al. y Zotenko et. al.

En particular, He et.al. propusieron la hipótesis de que la letalidad estaba asociada al grado k en la medida en que existen enlaces esenciales (además de nodos esenciales), y los hubs tienen simplemente más probabilidad de tener un enlace esencial. Pero cualquier nodo que tenga al menos un enlace esencial, es esencial también, por lo que los hubs tendrían más probabilidad de ser esenciales que las proteínas menos conectadas. Ellos supusieron que la probabilidad de que un nodo fuera esencial estaba determinada por sólo dos parámetros, a los que llamaron α y β. El parámetro α está definido como la probabilidad de que un enlace al azar sea esencial; β es la probabilidad de que un nodo al azar sea esencial. Por lo tanto, la probabilidad de que un nodo cualquiera de grado k sea esencial

está dada por:

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

Asumiendo la independencia de α y β, así como de la probabilidad de que dos enlaces cualesquiera sean esenciales. Y la probabilidad de que un nodo de grado k sea esencial es, entonces:

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

Tomando logaritmo natural de la Ec. 2 puede escribirse:

$$\ln(1 - P_E(k)) = \ln(1-\beta) + k\ln(1-\alpha) \tag{3}$$

Es decir, $\ln(1-P_E)$ es lineal con k. Para cuantificar el α y β de una red, puede considerarse que $P_E(k)$ es la proporción real de nodos esenciales para un dado grado, y así estimar α y β de un ajuste por cuadrados mínimos (esto podría compararse luego con los parámetros obtenidos de alguna otra manera para testear el modelo, que es lo que hicieron He et. al.). En la Fig. 3 se pueden ver estos ajustes lineales para cada una de las redes, y en la Tabla 4 los parámetros obtenidos para cada ajuste.

	α	β	R^2	p-valor
Red AP-MS	2,46 %	19,3 %	0,26	0,13
Red Y2H	1,51 %	17,97 %	0,20	0,20
Red Literatura	6,17 %	30,3 %	0,69	0,0029
Red Reguly	5,36 %	3,7 %	0,92	0,000013

Tabla 4: A partir de los ajustes de la Fig. 3, para cada red se obtuvieron los parámetros α y β definidos por He et. al. (2006). También se muestran el R^2 y p-valor del ajuste lineal.

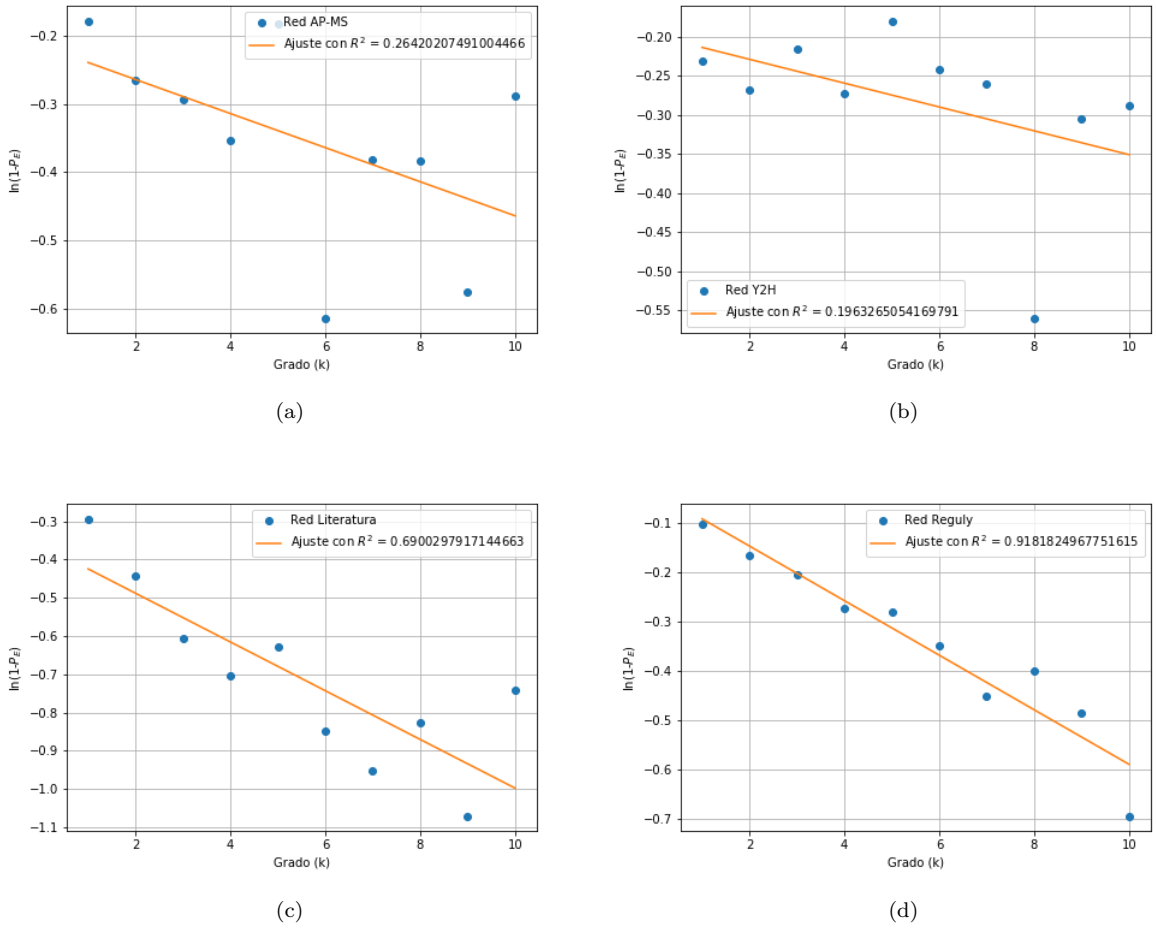
Como ya se mencionó, este modelo lleva implícita la independencia de la esencialidad de dos enlaces cualesquiera o dos nodos no adyacentes, y Zotenko et. al. (2008) pusieron esto en duda, partiendo de la base de que se sabía que existían sub-complejos de proteínas con funcionalidad similar, ricos en proteínas esenciales (o ricos en proteínas no esenciales). Para descartar la hipótesis de He et. al., se comparó la similitud entre nodos no adyacentes, pero mediados por un gran número de vecinos en común, para la red real y lo esperado por el modelo de enlaces esenciales. Se entiende que dos proteínas son del mismo tipo (similares) si ambas son esenciales o ambas son no-esenciales. En particular, según el modelo la probabilidad de dos nodos no adyacentes cualesquiera de ser esenciales es independiente, por lo que la probabilidad de que dos proteínas cualesquiera de grados k_1 y k_2 sean las dos esenciales o las dos no esenciales, es:

$$P_{sim}(k_1, k_2) = P_{NE}(k_1)P_{NE}(k_2) + P_E(k_1)P_E(k_2) \tag{4}$$

En la Tabla 5 se comparan los resultados del modelo con los de la red real. En particular, para casi todas las redes, el número de pares del mismo tipo es mayor que el número esperado, excepto en la red AP-MS.

	Nro. total de pares	Nro. de pares del mismo tipo	Nro. esperado del mismo tipo
Red AP-MS	11569	5875	6074
Red Y2H	522	352	291
Red Literatura	718	383	378
Red Reguly	10777	6187	5791

Tabla 5: La primera columna representa la cantidad de pares de nodos no adyacentes con 3 o más vecinos en común. La segunda es cuántos de estos pares son del mismo tipo (ambos esenciales o ambos no esenciales). La tercera es cuántos pares del mismo tipo se esperan por el modelo de ajuste lineal usando los parámetros de la Tabla 4.



(a) (b) (c) (d)

Figura 3: Gráficos de $\ln(1 - P_E)$ en función del grado k, con ajuste lineal según el modelo de He et. al. (2006). Los parámetros obtenidos están en la Tabla 4.

El hecho de que el número de pares del mismo tipo supere la expectativa indica que no se está teniendo en cuenta algún factor de dependencia de tipo entre nodos no adyacentes que comparten varios vecinos (en este caso, 3 o más). Es decir, el hecho de que dos nodos no adyacentes compartan varios vecinos parece estar incrementando la probabilidad de que sean del mismo tipo. Esto no se ve para la red AP-MS, y de hecho fue excluida en el trabajo de Zotenko et. al. La razón por la cual fue excluida es que el propio relevamiento de la red implica relaciones no-binarias, mientras que el modelo propuesto por He et. al. fue concebido pensando en redes de interacción binarias. Esto tiene que ver con que en la Ec. 2 la probabilidad de que un nodo sea esencial crece con el grado, y el relevamiento de la red AP-MS introduce un sesgo en el grado, puesto que no necesariamente cada proteína de cada clique de la red interactúa de manera física con todas las otras.

Por estos motivos, Zotenko et. al. introducen la noción de módulos biológicos complejos esenciales y no esenciales, dando cuenta de estructuras mesoscópicas altamente conectadas de la red, que pueden (ECOBIMs) o no (non-ECOBIMs) ser ricas en nodos y enlaces esenciales. En estas estructuras en particular (ECOBIMs), los nodos de mayor grado tienden a ser esenciales, más de lo que lo serían por azar en una estructura altamente conectada enriquecida en nodos esenciales.