Centralidad vs letalidad: análisis de redes protéicas

Andino C. , Asplanato L. , Murchison, F.

Departamento de Física, UBA

Resumen

Se estudiaron las redes de interacción proteína - proteína (*PPI*) correspondientes a la levadura *Saccharomyces cerevisiae* con el objetivo de encontrar la relación entre la importancia de una proteína para el organismo (su *esencialidad*) y las propiedades topológicas de la misma en la red de interacción. Para eso contamos con las redes PPI: *Y2H*, *AP-MS*, *LIT*, y *LIT-REGULY* que brindan los ejes de interacción obtenidos mediante interacciones binarias, co-pertenencia a compuestos y relevadas de literatura, respectivamente.

1. Introducción teórica

Las herramientas de análisis de redes complejas son de amplio uso en el ámbito de la biología como en neurobiología [4], co-expresión génica [3] y biología celular [5] entre otros. Uno de los sistemas más analizados es el de la levadura *Saccharomyces cerevisiae* que posee un interactoma altamente desarrollado [8]. En función de la información de interacciones, es interesante analizar si existe una relación entre la importancia biológica de los nodos e interacciones de una red y la estructura de esta. Mediante análisis de la topología de las redes, se busca estudiar la existencia de indicadores que denoten importancia biológica.

Uno de los primeros modelos para este análisis se propuso en un artículo publicado por Jeong et. al[2], donde se estudiaron redes de proteínas partiendo de la hipótesis de que los hubs son esenciales porque mantienen la conectividad de la red. Es decir, plantear a la centralidad como medida de importancia biológica; de aquí surge el concepto de la *centralidad-letalidad*. Si un nodo es esencial, sacarlo de la red afecta de forma terminal a la reproducción o vida de la célula.

He et. al [6] proponen una mirada alternativa al trabajo de Jeong; se introduce la idea de que la relación entre topología e importancia biológica dentro de interacción de proteínas no es sólo inherente a las mismas, sino de la existencia de interacciones o enlaces esenciales, los cuales se distribuyen aleatoriamente en toda la red y, por lo tanto, la regla de *centralidad-letalidad*es una consecuencia de esto. Un hub es esencial porque tiene mayores probabilidades que un nodo de menor grado de ser parte de una interacción esencial.

Finalmente, en un artículo en el que se refutan ambos artículos anteriores, Zotenko et. al [7], demuestra que el indicador de importancia biológica no es ni una medida de conectividad

local ni una propiedad global de la red. Zotenko demuestra que existen conjuntos de proteínas y enlaces, llamados complejos de proteínas, que se manifiestan en una escala mesoscópica.

2. Resultados y análisis

Características de las redes analizadas

Durante el trabajo se estudiaron redes de distinta proveniencia para tener un amplio espectro en el análisis. La red de interacciones binarias, *Y2H*, proviene de las interacciones *yeast-two-hybrid* del genoma completo de Ito et al.[1], que provee interacciones de alta confianza puesto que se encontraron al menos tres veces experimentalmente. *LIT-REGULY*, por otro lado, se basó en usar 30000 resúmenes de la literatura para establecer interacciones reportadas en experimentos a baja escala. Finalmente, *LIT* y *AP-MS* fueron relevadas de la literatura y mediante experimentos de co-pertenencia proteica en complejos respectivamente. De esta manera, las redes seleccionadas para el análisis representan diversas interacciones con distintos relevamientos. Sus propiedades se detallan en la tabla 1 a continuación.

	Número de nodos	**Número total de enlaces**	**Grado medio**	**Coeficiente de clustering medio**
AP-MS	1622	9070	11.184	0.555
Y2H	2018	2930	2.904	0.046
LIT	1536	2925	3.809	0.292
LIT-REGULY	3307	11858	7.171	0.261

Tabla 1: *Propiedades estructurales de cada red analizada*

Como se espera, se tiene que la red *REGULY* presenta el máximo número de nodos y ejes, pudiendo deberse a una sobre-estimación en las interacciones al basarse únicamente en la adquisición de interacciones a partir de resúmenes de publicaciones. Asimismo, también presenta el menor coeficiente de clustering global, sugiriendo que hay poca clausura entre los nodos. La red de co-pertenencia proteíca (*AP-MS*) presenta el mayor coeficiente de clustering de todas las redes a analizar, coincidiendo con el hecho que posee el mayor número de ejes luego de la de literatura Reguly. Esto se explica al considerar que esta red proviene de pensar que, al remover un complejo protéico, cada proteína del mismo se encuentra interactuando con el resto.

Dado que buscamos relacionar comportamientos biológicos con estas propiedades inherentes a las redes, realizamos una comparación de las mismas calculando el número de ejes de cada una presente en las otras. Los resultados se muestran en la tabla 2.

Como puede verse, la tabla 2 no es simétrica. Esto se debe al método de cómputo de solapamiento usado. Así como la primera columna muestra la intersección de la red *AP-MS* con las otras, normalizando por la cantidad de elementos en la misma, se repite para las otras tres redes.

AP-MS	0.089	0.444	0.213
0.029	Y2H	0.444	0.040
0.444	0.089	LIT	0.241
0.278	0.164	0.978	REGULY

Ojo! la tabla de Zotenko es de la fr

Tabla 2: *Fracción de nodos de cada red contenidos en las otras PPI's a analizar.*

Puede verse que la red *Y2H* es la que menor superposición tiene con el resto, mientras que *LIT*, con número similar de ejes (*interacciones reportadas*) posee un $\sim 40\,\%$ de solapamiento con *AP-MS* y *Y2H* y un $\sim 98\,\%$ con *LIT-REGULY*. Esto último tiene sentido al ser *REGULY* la red con mayor número de nodos y ejes, elevando las probabilidades de encontrar una interacción en esa. Cabe destacar que *AP-MS* sufre una excepción, puesto que tiene menos solapamiento que con *LIT*.

Centralidad vs Letalidad

El trabajo de Jeong et al. propone que la esencialidad de un nodo depende directamente con el grado del mismo y afirma que nodos de mayor grado son hasta tres veces más probables de ser esenciales que el resto. La hipótesis para este enfoque es que la eliminación que nodos de mayor grado remueve un gran número de interacciones de la red, rompiendo su conectividad y eliminando conexiones entre proteínas que sino no interactuarían.

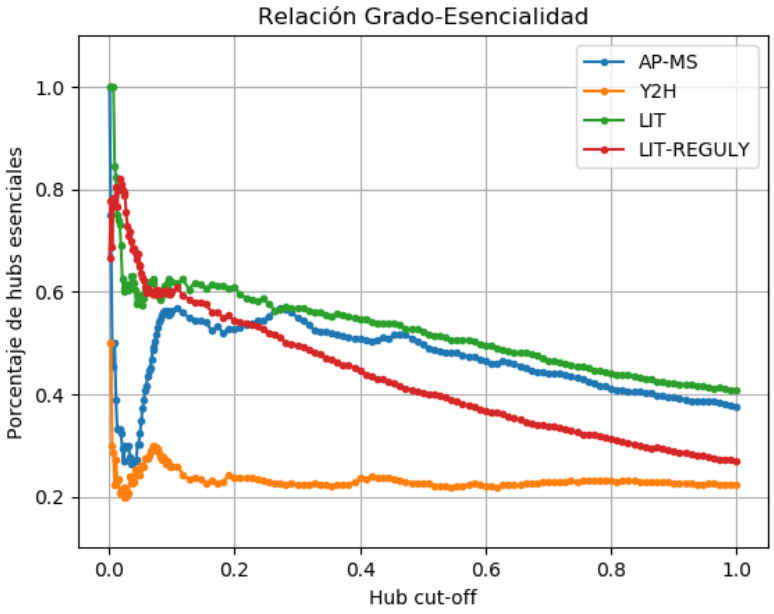


Figura 1: *Fracción de nodos esenciales ante distintos grados como límite de definición de un nodo hub.*

Para probar su validez dentro de las redes del análisis, se realizó el gráfico de la figura 1, empleando una base de datos de proteínas esenciales basada en un relevamiento de crecimiento ante deleción de genes. Se muestra el porcentaje de nodos esenciales dentro de los "*hubs*",

donde un hub se entiende como un nodo con suficiente grado, donde se amplia el umbral de grado para el cual un dado nodo entra en esa categoría.

Como puede apreciarse, la fracción de nodos esenciales dentro de los hubs considerados decrece rápidamente al ir disminuyendo el umbral de categorización. Esto muestra que en todas las redes los nodos de mayor grado tienden a ser esenciales, y este comportamiento decrece rápidamente con el grado. Se puede observar que la curva para la red *AP-MS* tiene un segundo máximo de porcentaje de esencialidad alrededor del 10 % de los nodos considerados como hubs. Esta misma red tiene el máximo grado medio de las cuatro analizadas; podría estar sucediendo que ambas familias de nodos, esenciales y no esenciales, posean un grado medio alto, por lo que al relajar la condición de hub, se tiene una población mixta de nodos, y a partir de cierto punto decae nuevamente la relación esencial-no esencial.

Análisis de vulnerabilidad

Los índices de centralidad, ya sean locales en el caso donde se considere un entorno de primeros vecinos para cada nodo, o globales, donde se consideran escalas mayores, dan un valor de importancia topológica de cada nodo.

Se emplearon los índices de *grado* como parámetro local, calculando la fracción de nodos de la red a los cuales cada nodo está conectado, de *betweenness*, que contabiliza la fracción de caminos mínimos que lo contienen entre todos los otros pares de nodos de la red, *autovector*, se obtiene la centralidad de un nodo mediante la centralidad de sus vecinos en un proceso iterativo, y finalmente, empleando un modelo de caminata al azar para calcular el índice de *corriente* como índices globales.

Se realizaron dos análisis de interés relacionando la centralidad de un nodo con estas medidas de conectividad. Por un lado, se calcularon los coeficientes de conectividad para cada nodo en la componente gigante de la red - pequeñas redes secundarias no poseerán la información relevante y se puede asumir que no tendrán la mayor cantidad de nodos esenciales, puesto que están débilmente interconectadas, en comparación con la componente gigante- al comienzo. Con los valores ya adquiridos, se procedió a eliminar nodos en orden decreciente de cada parámetro topológico en cada red y se calculó el tamaño de la componente gigante al hacerlo. Así, se tiene una idea de la vulnerabilidad de la red ante la remoción de nodos de relevancia topológica bajo distintas consideraciones, como se ejemplifica en la figura 2.

Sin embargo, esto no contempla el hecho que luego de remover un nodo de la red la estructura de la misma cambia ya que se alteran los enlaces de los nodos restantes. Para eso, se realiza el mismo análisis de integridad de la red ante la remoción de nodos, pero la elección de los nodos a eliminar se basa en su valor de conectividad en la red, actualizados luego de cada remoción. Los resultados se muestran en la figura 3.

En ambos análisis se graficó el tamaño de la componente gigante luego de la remoción de todos los nodos clasificados como esenciales por la base de datos mencionada en 2 para tener como referencia, además de una remoción aleatoria de nodos.

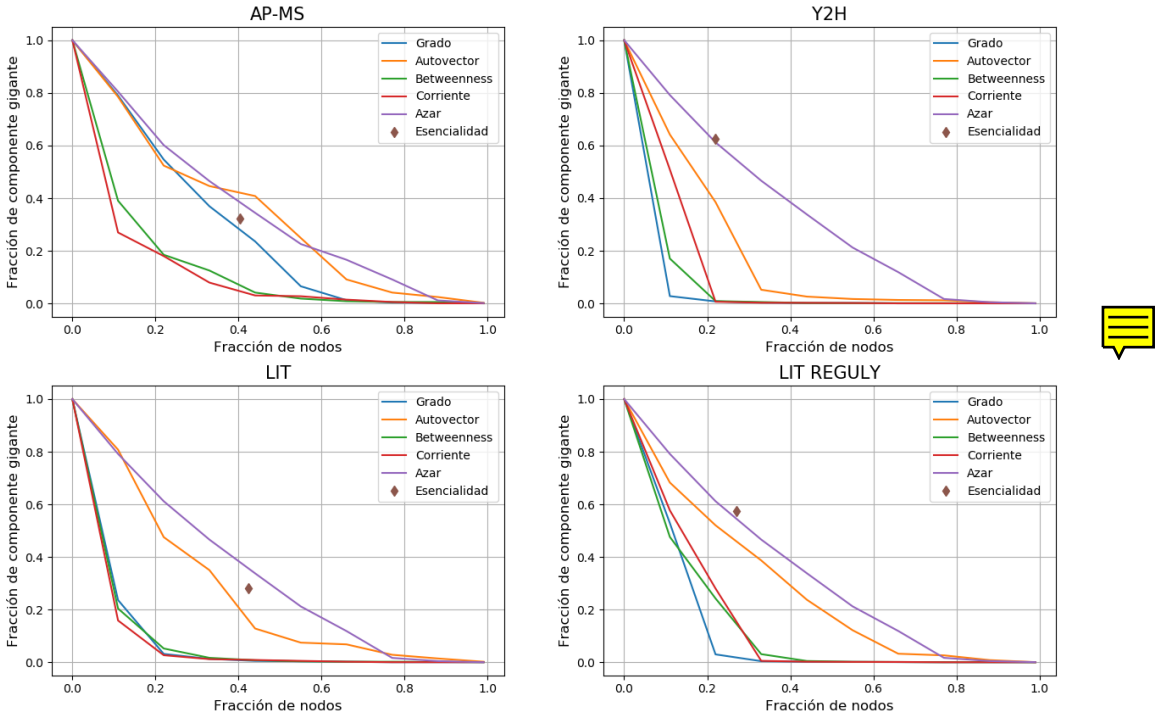


Figura 2: *Vulnerabilidad de las redes ante la remoción de proteínas de acuerdo a su índice de centralidad de grado, betweenness, corriente y autovectorial. Parámetros calculados sobre la red total y removidos en ese orden.*

Como primera observación, se nota que la remoción de los nodos con una única calificación inicial genera un decrecimiento más suave en el tamaño de la componente gigante. Además, en todas las redes, excepto la *AP-MS*, la remoción de nodos esenciales vulnerabilizó en menor medida que una igual remoción de nodos en función de otros parámetros topológicos, y de manera similar a la remoción de nodos al azar. En el caso de la remoción con recategorización de índices topológicos es evidente que la sustracción de nodos esenciales genera una disminución en la componente gigante similar a la remoción por azar. Además, cualquier otro parámetro de conectividad considerado, generó una ruptura casi total de la componente gigante de la red, en una fracción de nodos entre 5 % y 40 % menor. El hecho que haya tanta diferencia entre los dos métodos de remoción indica que la red cambia sus propiedades topológicas en gran manera al remover un nodo. En función de los gráficos presentados, la conectividad de grado como propiedad local y la autovectorial como global parecen ser los mejores criterios de centralidad para las redes, excepto en el caso de la red *AP-MS*, donde *betweenness* y grado muestran un efecto similar al azar durante la mayoría del proceso de remoción, y por lo tanto, los parámetros de interés serán el autovectorial y la corriente.

En términos generales, se puede acordar que la remoción de todos los nodos esenciales en una red es menos disruptiva que la remoción de igual número de nodos de acuerdo a cualquier parámetro topológico de conectividad, y dentro de ellas, el grado de un nodo suele ser una

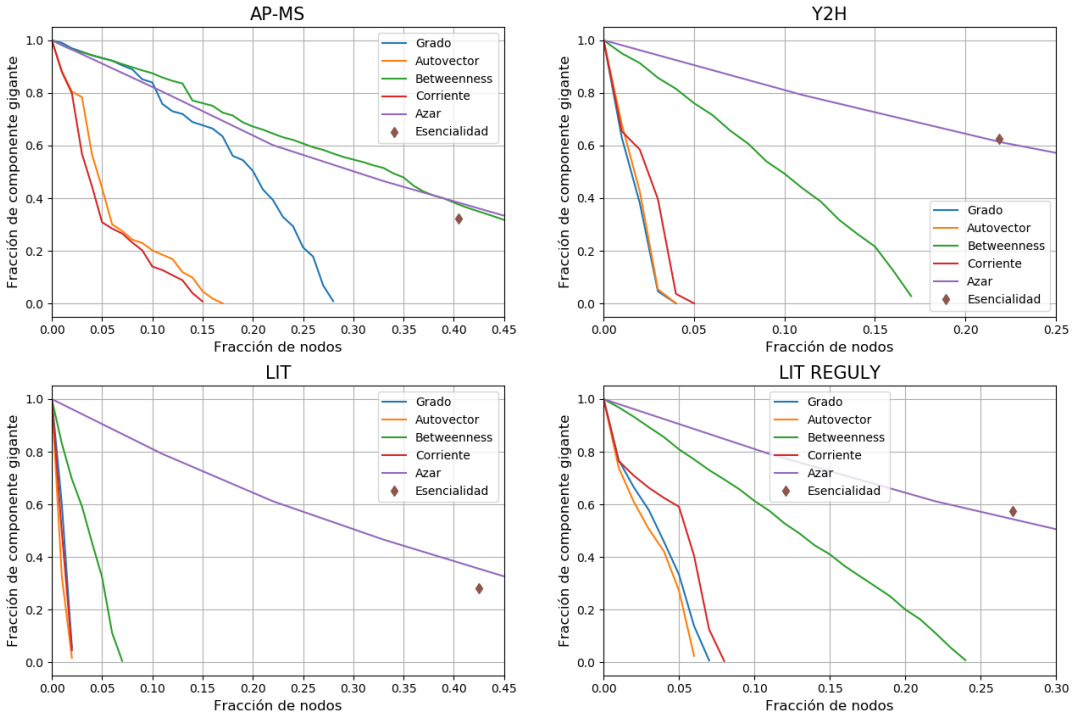


Figura 3: *Vulnerabilidad de las redes ante la remoción de proteínas de acuerdo a su índice de centralidad de grado, betweenness, corriente y autovectorial. Parámetros actualizados antes de cada remoción.*

buena elección.

Se realizó un análisis para comparar el efecto de la remoción de todos los nodos esenciales y la remoción del mismo número de nodos seleccionados al azar, pero que mantengan la misma distribución de grado.

Como ya vimos que la mayoría de los nodos esenciales están en los hubs (o equivalentemente, los nodos de mayor grado son esenciales en mayor proporción), sería imposible que los nodos a eliminar tengan exactamente la misma distribución de grado. La selección consistió en armar una lista de nodos esenciales y una de no esenciales ordenadas en orden decreciente de grado. De esta manera, comparando uno a uno, se eliminaron los mayores grados posibles y, cuando el grado de los esenciales era menor a los no esenciales, se seleccionó un nodo al azar con grado intermedio. El proceso se repitió 1000 veces para obtener el tamaño promedio y el error reportado proviene de la desviación estándar de los mismos. La fracción final de la componente gigante de cada red se muestra en la tabla 3.

Se observa que la red *LIT* es la de mayor vulnerabilidad ante la remoción de nodos esenciales. Esta red junto con la *AP-MS* presentaron menor efecto en el caso de remoción de nodos no esenciales, distando un 0,13 y 0,08 respecto del tamaño luego de eliminar los nodos esenciales. En el caso de la red *AP-MS*, ya se había visto que el grado no era un buen índice para vulnerabilizar la red, por lo que ese podría ser el motivo de la disparidad en los tamaños.

	Esencial	No-esencial al azar
AP-MS	0.324	$0{,}405 \pm 0{,}020$
Y2H	0.624	$0{,}545 \pm 0{,}016$
LIT	0.281	$0{,}414 \pm 0{,}004$
LIT-REGULY	0.575	$0{,}529 \pm 0{,}004$

Tabla 3: *Efecto de remoción de nodos esenciales en el tamaño de la componente gigante de cada red y la remoción del mismo número de nodos no esenciales elegidos al azar dentro de una distribución de grado determinada.*

Como se vió en la figura 1, esta red contaba con un número considerable de nodos esenciales de menor grado que las otras. Sin embargo, el índice de conectividad de grado parecía ser un buen parámetro en el caso de la red *LIT* en función del gráfico 3, por lo que no se comporta como esperado.

Esencialidad: Módulos biológicos vs. Interacciones esenciales

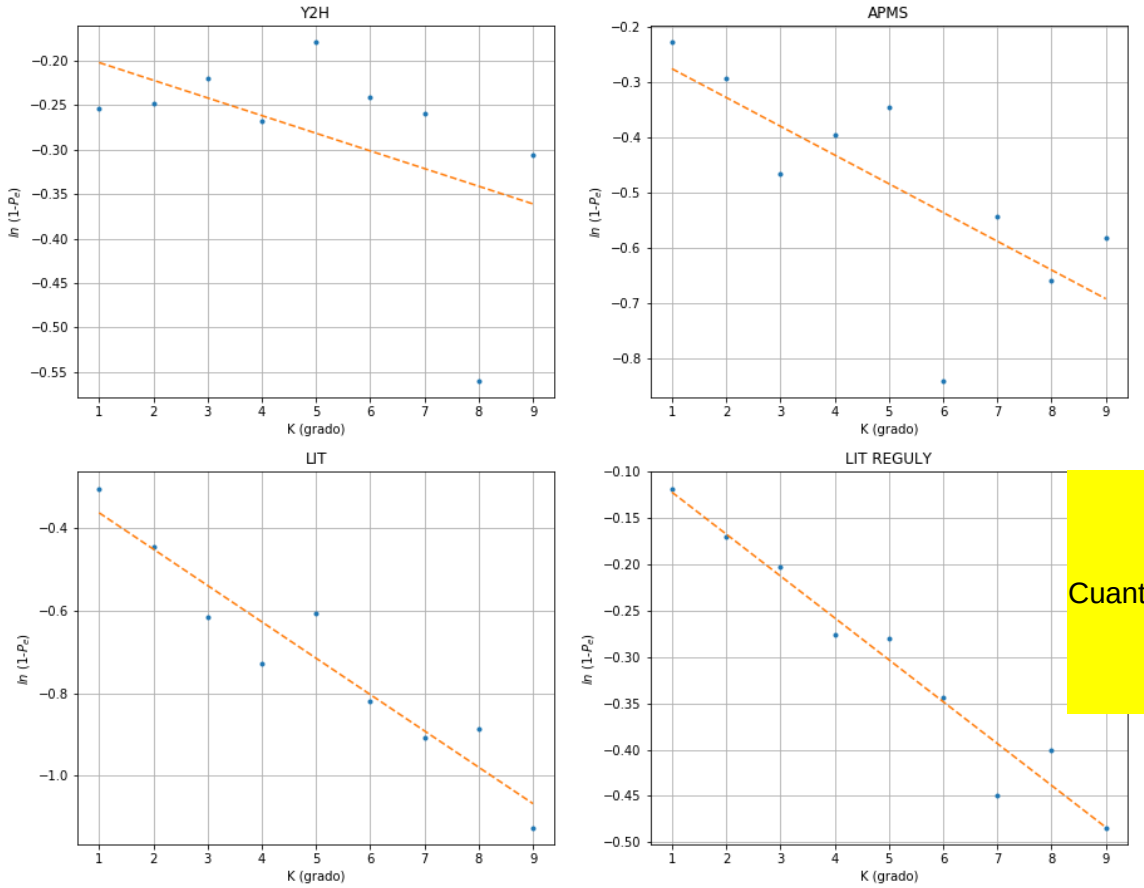
La principal hipótesis de He para su modelo de importancia de enlaces es que hay una relación entre dos "fuentes"de esencialidad en las redes de interacción de proteínas. Una proteína puede ser esencial por estar en interacción con otra mediante un enlace esencial, que se cuantifica con una probabilidad α o , con una dada probabilidad β, una proteína puede ser esencial en sí misma. La hipótesis de He es que ambas probabilidades son independientes, por lo que sus complementos (probabilidad de no ocurrir), también lo son y, por lo tanto, se las puede multiplicar o sumar sus logaritmos, entonces:

$$ln(1 - P_e) = kln(1 - \alpha) + ln(1 - \beta) \tag{1}$$

En la figura 4, se puede observar las fracciones de nodos esenciales en función del grado para las redes *Y2H*, *AP-MS*, *LIT* y *LIT-REG*. La red *Y2H* es la que posee el comportamiento menos claro, puede deberse a que suponer interacciones binarias lleve a estar ignorando algunos comportamientos de proteínas con alta conectividad. Esto podría subsanarse con datos de más interacciones.

En el resto de las figuras el comportamiento parece ser lineal. Si se observa el ajuste, las pendientes y ordenadas están bien definidas, por lo que el modelo de He parecería ser consistente, y la regla de centralidad-letalidad se cumple porque es más probable que nodos con grados más altos estén involucrados en más interacciones esenciales. Es decir, partiendo de que el modelo de He afirma que dichas interacciones esenciales se distribuyen de forma aleatoria por toda la red, es más posible que un hub sea esencial porque tiene mayor probabilidad de tener al menos un enlace esencial.

Sin embargo, Zotenko destaca que este modelo asume que, si dos proteínas no interactúan de forma directa, la probabilidad de que una de ellas sea esencial es independiente de que la otra lo sea, aún si poseen primeros vecinos en común. Para ver si esta hipótesis se sostiene en



Cuanto vale alpha y be

Figura 4: *Fracción de nodos esenciales como función del grado para las redes Y2H, APMS, LIT, LIT REGULY ajustados según la ecuación 1, los parámetros pendiente (m) y ordenada (b) resultantes fueron:* $m = -0{,}0199$ *y* $b = -0{,}1824$*, para Y2H,* $m = -0{,}0520$*,* $b = -0{,}2234$ *para APMS.* $m = -0{,}0883$*,* $b = -0{,}2737$ *para LIT y* $m = -0{,}0453$*,* $b = -0{,}0766$ *para LIT REGULY.*

las redes reales, se contaron cuántos pares de proteínas no interactuantes con 3 o más primeros vecinos en común eran pares donde ambas proteínas son esenciales o ambas no esenciales, y se comparó este valor con el esperado utilizando el ajuste lineal del modelo de He. La tabla 4 muestra la cantidad de pares totales de proteínas no interactuantes que comparten 3 o más primeros vecinos, la cantidad de estos pares que poseen la misma esencialidad, la cantidad esperada bajo el modelo de He de estos últimos, y el porcentaje que representa lo esperado respecto a los encontrados en cada red.

Puede verse que, salvo para la red *LIT*, hay una gran diferencia porcentual entre la cantidad de pares esperados y los contados en las redes. Esto refuta la hipótesis de probabilidades independientes de He, por lo que al igual que Zotenko, el modelo de He no es válido para estas redes. Para verificar que los resultados de la tabla 4 no fueran producto de la elección del grado de primeros vecinos con el cual se eligió contar pares, se repitió el análisis para 2 y

	Pares Totales	Pares Iguales	Pares Iguales Esperados	Porcentaje
Y2H	522	352	284.16	80.73
AP-MS	11613	5907	7772.68	131.58
LIT	730	389	399.67	102.74
LIT-REGULY	10777	6187	5633.31	91.05

Tabla 4: *Los pares totales son la cantidad de pares que no tienen interacción directa entre sí, pero comparten 3 o más primeros vecinos. Pares iguales hace referencia a la cantidad de estos pares en los que ambas proteínas son esenciales o no esenciales. Pares iguales esperados es la cantidad de estos últimos esperada según un ajuste lineal utilizando el modelo de He.*

4 primeros vecinos como umbral y se muestra en la tabla 5.

Umbral de corte : 2 vecinos

	Pares Totales	Pares Iguales	Pares Iguales Esperados	Porcentaje
Y2H	2258	1514	1275.87	84.27
AP-MS	15467	7740	9897.99	127.88
LIT	1858	1047	983.80	93.96
LIT-REGULY	43027	25898	22842.50	88.20

Umbral de corte: 4 vecinos

	Pares Totales	Pares Iguales	Pares Iguales Esperados	Porcentaje
Y2H	185	121	97.30	80.42
AP-MS	9314	4793	6444.29	134.45
LIT	383	195	217.31	111.44
LIT-REGULY	5342	3065	2864.08	93.44

Tabla 5: *Iteración del análisis con umbrales distintos.*

Comparando los resultados de cada caso, se ve que la red *Y2H* y la *LIT-REGULY* muestran tendencias similares independientemente del grado elegido para contar pares, y en ambos casos se espera un número menor de pares iguales a los que hay en cada una de las redes (las redes estudiadas por Zotenko presentan esta misma tendencia).

Por otro lado, la red *AP-MS*, si bien también muestra una tendencia independiente del grado que se elija, parece esperar más pares de igual esencialidad que los contados en la red. Dado el alto grado medio y coeficiente de clustering de la red, podría asumirse que se trata de una base de datos armada como se especificó, todas interactúan entre sí. De este modo, aumenta la probabilidad de que dos proteínas interactúen de forma directa, disminuyendo la cantidad que lo haga por medio de primeros vecinos únicamente. Además, He describe como una de las limitaciones de su modelo redes construidas de esta manera, ya que se crean interacciones entre proteínas esenciales de más, lo que causa una sobre-estimación de α. Ambas explicaciones justifican la tendencia vista para esta red.

La red LIT no parece mostrar una tendencia clara respecto la relación entre pares esperados y los contados en cada caso. Si consideramos que el grado medio de la red es de 3,81, al

buscar que dos proteínas compartan 3 o 4 primeros vecinos (o más) se terminan seleccionando un número de pares estadísticamente poco representativo, por lo que no se puede concluir que Zotenko falla para estos casos. Entonces, parecería haber un límite en el grado de primeros vecinos que podemos pedir como umbral para contar pares. En este sentido, es estadísticamente más representativo seleccionar pares que compartan 2 primeros vecinos. En este caso el porcentaje neto sí contiene información más conclusiva debido a que se seleccionan una mayor cantidad de grupos.

3. Conclusiones

Se pudieron caracterizar las redes a analizar con éxito, obteniendo sus parámetros de estructura. Con esa información, se analizó la relación entre el grado de los nodos y la esencialidad de los mismos. Como se estudió, los nodos de mayor grado son mayormente esenciales. Sin embargo, también se estudiaron otros parámetros topológicos para analizar la relación entre esencialidad para el sistema biológico y relevancia en la red. Jeong postuló que los hubs de una red son esenciales porque preservan la conectividad de la red, .ªcercando"nodos que de otra forma no interactuarían.Se realizó un análisis de vulnerabilidad ante cuatro parámetros topológicos con los cuales determinar la relevancia de un nodo en la red: grado, *betweenness*, autovectorial y de corriente. Al contrastar los resultados con la remoción azarosa de nodos y la remoción de todos los nodos esenciales se encontró que la integridad de la red es muy susceptible ante remoción de nodos de acuerdo a su grado y parámetro de conectividad autovectorial, exceptuando la red *AP-MS*. Por otro lado, el tamaño de la componente gigante de cada red ante la remoción de todos sus nodos esenciales contra el tamaño luego de remover nodos no esenciales al azar, con distribución de grado similar, fueron similares. De esta forma se corrobora la correlación entre grado y esencialidad de un nodo, pero los hubs no necesariamente son esenciales por brindar conectividad a la red, ya que la misma se rompe de igual manera ante remoción de nodos no esenciales.

Respecto a la refutación de Zotenko a la hipótesis de He, se pudo ver que el modelo de He presenta variaciones significativas al predecir la cantidad de pares no interactuantes que comparten más de k primeros vecinos y que comparten esencialidad, respecto a las encontradas en las redes reales. Se puede concluir que la hipótesis de partida que se basa en distribución aleatoria de enlaces esenciales no es válida. Como observación, parece existir algún tipo de limitación en cuanto a la cantidad de primeros vecinos utilizados como umbral para el conteo de pares, relacionada con el grado medio de la red; suonemos que esta se relaciona con el grado medio de la red y su coeficiente de clustering.

Referencias

[1] Ito T, Muta S Ozawa R Chiba T et al.(2001): *A comprehensive two-hybrid analysis to explore east protein interactome*, 8: 4569-4574.

[2] Jeong H, Barabasi AL Oltvai ZN(2001): *Lethality and centrality in pretein networks.*

[3] Greicius, Michael D. / Krasnow, Ben / Reiss, Allan L. / Menon, Vinod(2003): *Functional connectivity in the resting brain: A network analysis of the default mode hypothesis*, 1: 253–258.

[4] B, Zhang / S, Horvath(2005): *A General Framework for Weighted Gene Co- Expression Network Analysis.*

[5] Albert, Réka(2005): *Scale-free networks in cell biology*, 21: 4947–4957.

[6] He X, Zhang J(2006): *Why Do Hubs tend to be essential in protein networks?*

[7] Zotenko E, O'Leary P Przytycka T(2008): *Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality.*

[8] Dana Faber Cancer Research, Harvard Medical School (2013-2018): *CCSB Interactome Database, http://interactome.dfci.harvard.edu/index.php?page=home, [15 Oct 2018]*
.