umero de enlaces compartidos sobre el número de enlaces totales de la red correspondiente a la fila donde se encuentran.23

???

Trabajo Práctico 2

Facundo Emina Juan Pablo Fiorenza

Redes complejas con aplicaciones a sistemas biológicos
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

Resumen

En una red de interacción de proteínas la regla de centralidad - letalidad que afirma que los nodos de alta conectividad corresponden a nodos esenciales, nos permite a partir del análisis de la importancia topológica de un nodo predecir su comportamiento biológico. Para ello se realizó el análisis de cuatro redes de interacción de proteínas de levaduras. En primer lugar, se calcularon las características principales de la red y se estudió como cambiaba su topología al realizar la extracción de nodos con distintos valores de centralidad. Por último, se analizó la correlación que había entre pares de nodos que no estando conectados de manera directa pero si teniendo vecinos en común, compartían esencialidad.

1. Introducción

Una red se compone de un conjunto de nodos de múltiples conectividades enlazados entre sí. Se dice que el grado de cada nodo es la conectividad que posee, es decir, la cantidad de enlaces que lo vinculan con otros nodos. Existe un amplio campo de investigación en la biología donde se estudian diversos sistemas como redes complejas de interacciones, por ejemplo, los complejos proteicos. En particular, este trabajo se enfoca en realizar una serie de análisis sobre cuatro redes proteicas para determinar la correlación entre la importancia topológica de una proteína y su esencialidad a nivel funcional.

Se ha demostrado que una alta conectividad para un nodo (a estos nodos se los denominan *hubs*) está fuertemente vinculada con su importancia en la topología de la red. Esto quiere decir que al remover un hub de la red, la estructura de la misma se ve fuertemente afectada; por ejemplo, puede aumentar fuertemente el número de pares de nodos sin camino que los vincule o bien aumentar el diámetro de la red (distancia máxima entre dos nodos). A este fenómeno se lo conoce como Regla de Centralidad - Letalidad, donde un nodo hub se lo asocia que corresponde a un nodo esencial de la red de proteínas, por ende así como su remoción de la red produciría grandes cambios topológicos dichos cambios se correlacionan con un fuerte impacto en la funcionalidad del complejo proteico modelado.

En nuestro trabajo se analizó el impacto que tendría sobre el largo de la componente gigante de cada red la extracción de nodos que tuvieran distinto valor de centralidad, y de esta manera poder realizar una comparación con el ocasionado por la remoción de los hubs.

A que llaman hubs?La idea es comparar la

Finalmente, se calculó la cantidad de nodos que no estando conecta
en común, compartían esencialidad y se comparó dicho valor con el e
He y sus colegas [1] expresado en la siguiente ecuación:

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

Que es alpha? Y beta? y k? Que exp

2. Características de las redes a analizar

En el presente informe se analizaron cuatro redes de proteínas d
número de nodos, el número de enlaces, el promedio del grado de sus nodos, y por último el promedio del coeficiente de clustering.

Red	Número de nodos	Número de enlaces	Grado medio	<Ci>
AP MS	1622	9070	11,18	0.55
LIT	1536	2925	3.81	0,34
Y2H	2018	2930	2,90	0.153985
LIT Reguly	12222	11859	1,94	0,10

Tabla 1: Porcentaje de enlaces compartidos entre las redes

A su vez, se compararon todas las redes entre sí y se calculó la fracción de enlaces compartidos, es decir, la cantidad de pares de proteínas unidas que poseen en común cada red.

AP MS	0,14	0,03	0,28
0,44	**Y2H**	0,09	0,86
0,09	0,09	**LIT**	0,16
0,21	0,24	0,04	**LIT Reguly**

Tabla 2: Porcentaje de enlaces compartidos entre las redes

donde por ejemplo, el 14 % de los enlaces contenidos en la red *AP-MS* están también presentes en la red *Y2H*.

Por último para finalizar con la caracterización global de las redes, se prosiguió a analizar la fracción de nodos esenciales entre los hubs de nuestra red. Para ello se definieron distintos valores de cut - off para el grado, tomándose como criterio para la selección de los hubs a aquellos nodos que presentaban un grado mayor a dicho cut - off. Luego, se calculó de esa cantidad de hubs la fracción que era esencial, realizándose sucesivas veces para valores de cut - off que fueron desde uno hasta el grado máximo de nuestra red, del cual se obtuvo el siguiente gráfico.

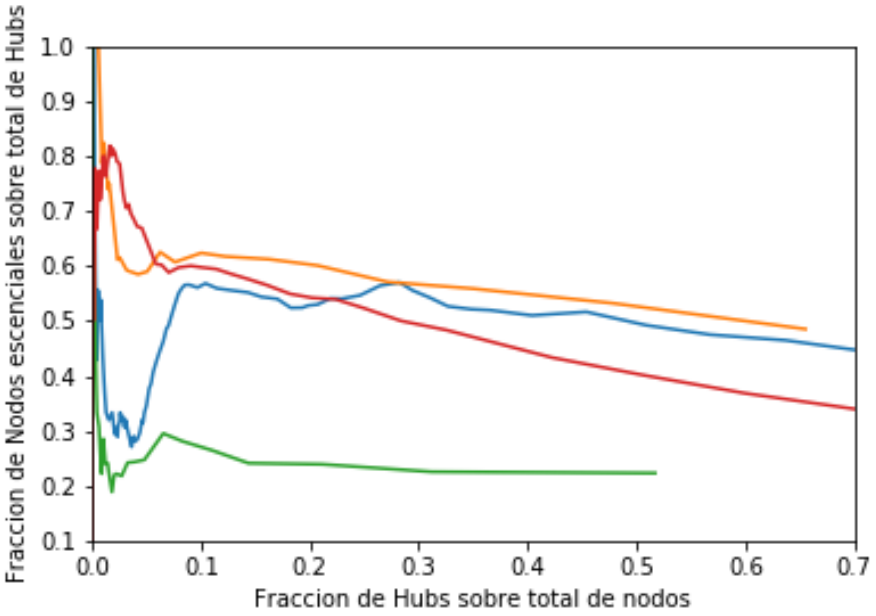

Figura 1: Relación entre hubs y escencialidad

Donde en el eje x se colocó la fracción de hubs para los distintos cut - off y en el eje y la proporción de nodos esenciales entre los nodos considerados hubs.

Los valores mas grandes de cut - off se encuentran mas cerca del cero en x, lo cual resulta obvio al observar que para valores grandes de este la proporción de nodos tomados como hubs era menor, siendo particularmente estos la fracción de nodos con mayor grado. Como puede observarse en dicha figura, la mayor proporción de nodos esenciales se dio en dicho margen, indicando que es dentro de los nodos con mayor grado que uno encuentra las proteínas esenciales. Lo cual se condice con la regla de centralidad-letalidad al mostrar la importancia que tiene para la topología de la red los nodos de alta conectividad.

Las proteinas esenciales no se encuentran necesariamente en los nodos de r

3. Análisis de vulnerabilidad

Con el fin de analizar la capacidad disruptiva en la topología de la red dada por los nodos esenciales se comparó el impacto topológico de la remoción de los nodos esenciales con el dado por la extracción de nodos clasificados por distintos índices de centralidad. Entre estos encontramos los valores de centralidad locales y de intermediación.

En los valores de centralidad locales, el índice esta mayormente determinado por la topología de su entorno, en nuestro caso realizamos la clasificación en función de dos de ellos, el *grado* y *centralidad de autovector*.

Existen también valores de centralidad de intermediación, donde esta viene dada por la capacidad del nodo de mantener la conectividad entre pares de otros nodos de la red, en nuestro caso se

utilizó para dicha clasificación la *centralidad de intermediación del camino mas corto.*

¿Recalculaban la centralidad cada vez o solo al comienzo o en bloques?

Luego de clasificados los nodos en función de dichos índices, se prosiguió a analizar de a uno el impacto que tenían en la conectividad de la red. Para esto se medía el largo de la componente gigante, luego se extraía el nodo con mayor valor de centralidad y se volvía a medir dicho largo, realizando este procedimiento sucesivas veces.

interesante, ¿bajo que metodo?

Después se calculó la importancia topológica de los nodos esenciales, extrayendo también de a uno y midiendo iteradamente el tamaño de la componente gigante.

Por último se realizó una curva que sirviera de referencia, extrayendo nodos de manera aleatoria y midiendo el cambio en el tamaño de la componente gigante. Esta última medición se realizó 1000 veces calculando luego un promedio de ellas.

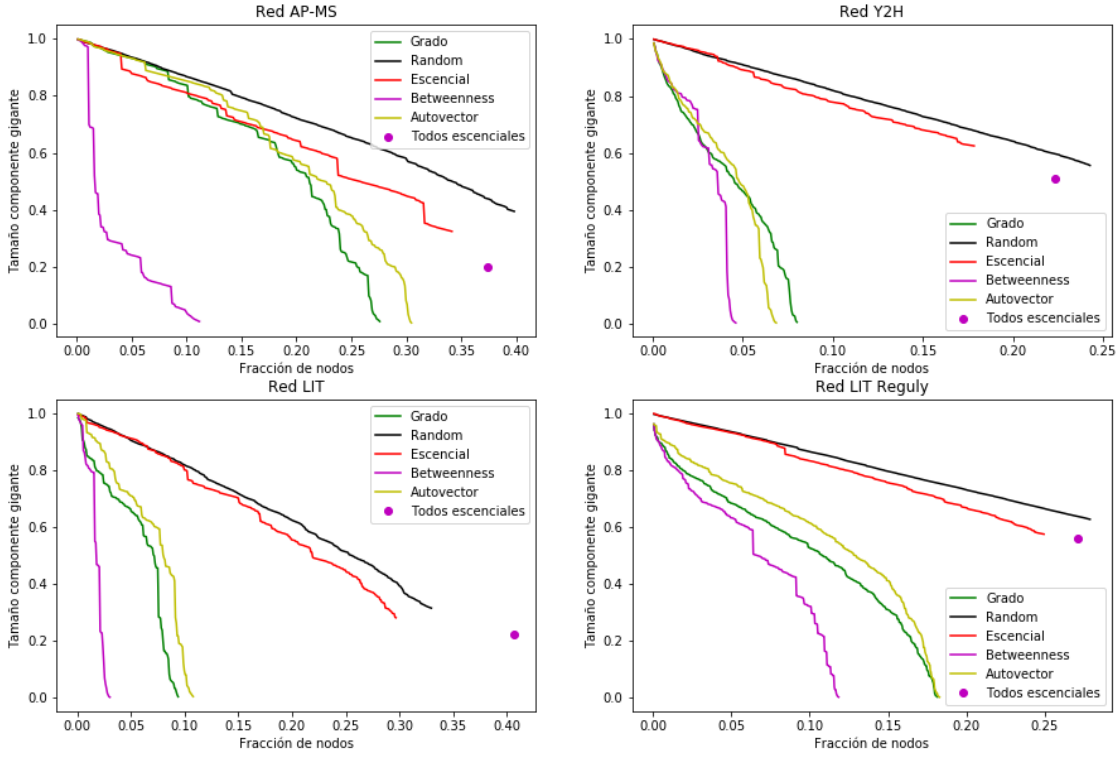

Figura 2: Relación entre hubs y escencialidad

Figura 3: Impacto de distintos valores de centralidad en el tamaño de la componente gigante para la red de proteínas.

Como resultado de dichos cálculos se confeccionaron los gráficos de la figura (2), en esta puede

observarse que existe una similitud en el impacto que cada índice de centralidad tuvo para las cuatro redes analizadas. En primer lugar, se puede afirmar que el impacto de los nodos de mayor grado sobre la topología de la red fue muchísimo mayor que el de los nodos esenciales, de los cuales resulta importante remarcar que fueron extraídos indistintamente de su grado. A su vez, comparando el impacto que tuvo la extracción de los nodos teniendo en cuenta su grado o su valor de autovector no se encontraron grandes diferencias en ninguna de las cuatro redes. Lo cual resulta esperable ya que ambos índices, al ser locales, analizan el entorno de cada nodo.

Por último, en cuanto al cálculo del índice de centralidad de intermediación del camino mas corto (betweenness en la figura), se observa que su impacto topológico fue el mayor de todos, permitiéndonos concluir que la capacidad de un nodo de mantener la conectividad entre pares de otros nodos resulta ser la característica mas importante, de entre las analizadas, a la hora de preservar la topología de la red.

Esto demuestra que a la hora de observar la importancia de un nodo en preservar la estructura de la red, resulta más importante su centralidad, independientemente del índice analizado, que su esencialidad a nivel biológico, incluso no puede apreciarse en ninguna red grandes diferencias entre el impacto de los nodos esenciales y la extracción de nodos al azar.

Para continuar con nuestro análisis comparativo analizamos el impacto que tendría sobre el tamaño de la componente gigante la extracción de todos los nodos esenciales al mismo tiempo y luego la comparamos con el impacto que tendría extraer una misma cantidad de nodos que tuvieran el mismo grado que los nodos esenciales pero siendo estos últimos no esenciales.

Red	**Esenciales**	**No esenciales**
AP MS	0,32	0,54
LIT	0,28	0,69
Y2H	0,62	0,76
LIT Reguly	0,57	0,71

¿Cuántas veces realizaron al

Tabla 3: Proporción de nodos en la componente gigante por la extracción de todos los nodos esenciales y una misma cantidad de nodos no esenciales pero del mismo grado que estos.

Si bien lo esperado según los resultados de Zotenko y sus colegas [2] era que los valores no fueran muy distintos entre las proteínas esenciales y las no esenciales cuando estas tenían el mismo grado, como se observa en la Tabla 3 la diferencia entre la fracción de nodos en la componente gigante luego de extraer todos los nodos esenciales es menor en la red Y2H que en todas las restantes. Permitiéndonos aproximarnos a la noción de que la topología de la red se ve afectada por la extracción de nodos de un mismo grado, indistintamente de si estos son o no esenciales.

Particularmente se observa de la figura 2 que la extracción de los nodos con alto grado presentan un fuerte impacto en la topología de la red, y por lo dicho anteriormente puede desestimarse que esto se deba a su carácter de esencial.

4. Esencialidad: Módulos biológicos vs Interacciones esenciales

Para explicar la regla de centralidad-letalidad, He y sus colegas [1] propusieron un modelo en el que una proteína es esencial debido a que pertenece a un enlace esencial o bien a otros factores (mayormente de origen biológico). En esta sección, se sometió a este modelo a las redes de estudio (Y2H, LIT, LIT REGULY).

Dado que He propone que la probabilidad de que un nodo sea esencial depende de la probabilidad de pertenecer a un enlace esencial (dependiente de la conectividad del nodo) junto con la probabilidad de que algún factor de otro orden la haga esencial, se determinó las probabilidades de que un nodo pertenezca a un enlace esencial (α) y la probabilidad de que algún otro factor haga esencial a un nodo (β) que propone en su trabajo. Para ello, se realizó un recableado aleatorio de la red 10000 veces manteniendo la distribución de grado (*ver apéndice*) y calculamos el valor medio de enlaces entre proteínas esenciales (al que llamamos m). Conociendo el número de enlaces entre proteínas esenciales de la red original (al que llamamos n) y el número total de enlaces (E), se calculó $\alpha = (n - m)/E$. Si suponemos que n es significativamente mayor a m, podemos apreciar que α representa la probabilidad de que un nodo sea esencial, dado que n-m representa el exceso de enlaces entre proteínas esenciales.

enlace

Una vez calculado α, se calculó β de la siguiente manera: se removió toda la información de la red en cuanto a esencialidad de los nodos; se distribuyó $n - m$ enlaces esenciales de forma aleatoria; a cada nodo enlazado por éstos se les asignó el atributo de esencialidad. Luego, se calculó la probabilidad de que un nodo sea esencial asignando aleatoriamente este atributo a los nodos restantes hasta alcanzar el mismo número de nodos esenciales que la red original tenía. Se repitió el proceso 10000 veces para tener una mejor estadística (*ver apéndice*). El análisis no fue realizado sobre la red AP-MS debido a que , al realizar las simulaciones, el número de enlaces esenciales de la red original (n) era menor al número de enlaces medio obtenido (m). Intuimos que esto se debe a que los enlaces de la red representan la pertenencia a un mismo grupo proteico [1].

Una vez obtenidos estos coeficientes, se graficó la probabilidad de que un nodo de grado k sea esencial observada en la red, junto con la estimación que nos brinda la ecuación 1. En la figura 5 se muestran los resultados obtenidos. Es importante aclarar que para nodos de grado tal que no hubiese más de 10 en la red se los eliminó del análisis puesto que no resultan estadísticamente significativos.

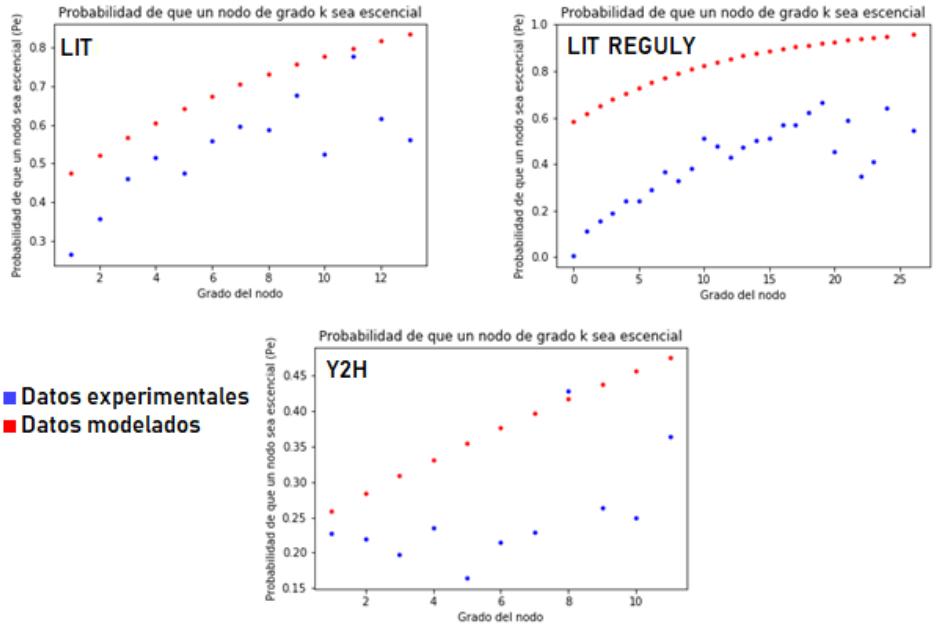

Figura 4: Probabilidad observada de que un nodo de grado k sea esencial vs. la probabilidad de que un nodo sea esencial estimada a partir del modelo

Es posible ver cómo los datos predichos por el modelo se encuentran bastante alejados de las probabilidades observadas. Esto probablemente se deba a fallas en el código.

A partir de la ecuación (1), se graficó *log(1-Pe) vs k*, siendo $1 - Pe$ la probabilidad de que un nodo no sea escencial y k el grado. Se ajustó los puntos por una recta cuya pendiendte sería $log(1 - \alpha)$ y su ordenada al origen $log(1 - \beta)$. En la figura (5) se muestran los ajustes y en la tabla (4) los valores de α y β obtenidos mediante el ajuste y mediante las simulaciones, mostrando su diferencia porcentual.

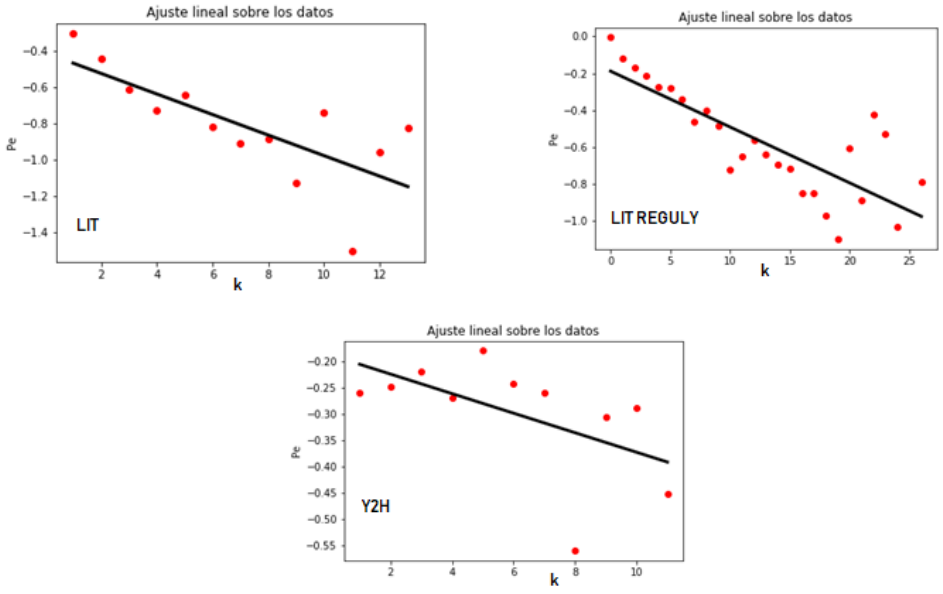

Figura 5: Ajuste lineal a la distribución de probabilidad de no ser un nodo escencial en función del grado.

Observando los resultados de la tabla 4 se puede ver una fuerte discrepancia entre las probabilidades obtenidas en la simulación y aquellas obtenidas con el ajuste. Esto probablemente se deba a algún problema en el código utilizado, por lo que no incluiremos estos datos en el posterior análisis.

Red	**Alpha (%) (ajuste)**	**Beta (%) (ajuste)**	**Alpha (%) (simulación)**	**Beta (%) (simulación)**	**Diferencia porcentual Alpha(%)**	**Diferencia porcentual Beta (%)**
LIT	6,72	35,26	9,19	42,12	26,88	16,29
LIT Reguly	12,23	61,21	8,38	58,29	45,94	5,01
Y2H	4,19	34,91	3,40	23,29	23,24	49,87

Tabla 4: Valores de coeficiente α y β para cada red según el ajuste y la simulación

Por lo dicho al comienzo de esta sección se puede inferir que según el modelo de He et al. si dos proteínas no interactúan entre sí, la esencialidad de cada una de ellas no esta determinada por la esencialidad de la otra. De esta manera, si dos proteínas interactúan con los mismos vecinos pero no de manera directa entre sí, las predicciones del modelo también deberían ser válidas.

En nuestro caso, se analizaron para cada red aquellos pares de nodos que no estaban conectados directamente y compartían tres o mas vecinos. Luego, se observó cuantos de esos pares compartían esencialidad, es decir, eran ambos esenciales o ambos no esenciales. Dicho cálculo fue luego comparado con el valor de pares de nodos esperados, que compartieran esencialidad no estando conectados, según el modelo de He y sus colegas reflejado en la ecuación (1) (Ver Apéndice).

¿Para Y2H usaron más de 3 vecinos en común o más de uno? Porque el número es muy grande.

Redes	**Número total de pares**	**Pares de la misma especie**	**Pares de la misma especie estimados con el modelo (ajuste lineal)**	**Diferencia porcentual (%)**
Y2H	23073	15087	11558,09	1,24
LIT	730	389	560,74	1,30
LIT Reguly	10777	6187	7123,14	1,36

Tabla 5: Pares de nodos disjuntos que comparten esencialidad. Red real y predicciones del modelo de He.

¿Diferencia porcentual entre quién y quién? ¿

Como se observa en la Tabla (5) los pares de nodos que siendo disjuntos comparten tres o mas vecinos y a su vez son de la misma especie, presentan diferencias porcentuales mayores al 1 %. Estas diferencias son aun mayores a las que presenta Zotenko en su estudio [2], cuyas diferencias porcentuales son del orden del $0,16\,\%$ y las considera estadísticamente significativas. Esto permite desestimar la hipótesis propuesta por el modelo de He et al. de que la esencialidad de las proteínas que no interactúan de manera directa (representado en nuestra red por un enlace) es independiente.

5. Conclusión

En lineas generales, se pudo verificar parte de los resultados obtenidos por Zotenko [2]. Se observó para las cuatro redes que la fracción de nodos esenciales sobre el total de Hubs es mayor cuando el $k_{cut-off}$ es menor, lo que muestra una correlación entre la conectividad del nodo y su esencialidad. Sin embargo, analizando la vulnerabilidad de la red se demostró que a partir de distintos parámetros de centralidad, uno puede definir la importancia de un nodo. Observando las curvas obtenidas en la figura (2), el impacto de la esencialidad de los nodos sobre la topología de la red es mucho menor que el observado para los distintos índices de centralidad analizados.

Por otro lado, se modelaron las redes Y2H, LIT y LIT Reguly utilizando el modelo de He et al. Se trató de simular el modelo pero los resultados obtenidos difieren en gran medida del ajuste lineal realizado a los datos obtenidos.

Asimismo, se calculó y comparó con el modelo de He y sus colegas, el número de nodos que compartieran esencialidad pero sin estar enlazados entre sí y teniendo en común tres o más vecinos (más de 1 para Y2H) observándose una diferencia entre dichos resultados, siendo por ejemplo 15087 los pares de la misma especie registrados con el recuento directo de la red Y2H y 11558.09 los predichos según el modelo de He y sus colegas, obteniéndose una diferencia mayor incluso que la observada por Zotenko [2].

6. Apéndice

6.1. Modelo de He

A continuación, dejamos explicitado el código utilizado al recablear la red para calcular las probabilidades α y β. Al realizar el recablado aleatorio para calcular lo que se hizo fue eliminar la

información de enlaces y tomar el nodo de mayor grado (k) en la red; luego, se le asignó k enlaces arbitrarios (sin autoenlaces posibles). Posteriormente, se repitió el proceso con el siguiente nodo de mayor grado, teniendo en cuenta los nuevos enlaces generados, y así sucesivamente. Con tener en cuenta los enlaces nuevos generados, nos referimos a que existirán k nodos a los cuales habrá que asignarles (k-1) enlaces, pues ya tienen uno nuevo asignado. Realizar este proceso en orden permite un recableado más efectivo, dado que empezar por los nodos de más bajo grado (o bien, elegirlos de forma aleatoria) puede generar que nodos de bajo grado se conecten entre sí, dejando a nodos de alto grado sin poder recablearse completamente.

Cálculo de α

```
def coef_alfa(g):

    #Calculo el n de la red original (IBEP original)
    n = 0
    for enlace in g.edges():
        if g.node[enlace[0]]['escencialidad'] == 'escencial' and g.node[enlace[1]]['escencialidad'] == 'escencial':
            n = n + 1

    lista_de_m = []

    for _ in range(10):
        nuevos_edges = []
        nodo_grados=[]
        for node in g.nodes:
            nodo_grados.append([node,g.degree[node]])
            #Para cada iteracion tengo que volver a crear
            #la lista de [nodos,grados]
            if g.degree(node) == 0:
                #Elimino nodos de grado 0
                nodo_grados.remove([node,g.degree[node]])
        nodo_grados.sort(key=lambda grado: grado[1], reverse=True)
        while len(nodo_grados) > 1:
            while nodo_grados[0][1] > 0:
                #agarro el nodo de grado mas alto
                nuevo_vecino = random.choice(nodo_grados[1:])
                if [nodo_grados[0][0],nuevo_vecino[0]] not in nuevos_edges:
                    nuevo_vecino[1] = nuevo_vecino[1] - 1
                    nodo_grados[0][1] = nodo_grados[0][1] - 1
                    nuevos_edges.append([nodo_grados[0][0],nuevo_vecino[0]])
                    #mi nueva lista de enlaces puede tener menos enlaces que la original,
                    #pero no sera significativo
                    if nuevo_vecino[1] == 0:
                        nodo_grados.remove(nuevo_vecino)
            #Reordeno la lista con los nodos redefinidos y con los nodos con k = 0 fuera
            nodo_grados.remove(nodo_grados[0]) #(IV)
            nodo_grados.sort(key=lambda grado: grado[1], reverse=True)

        #me fijo cuantos enlaces escenciales tengo
        m = 0
        for enlace in nuevos_edges:
            if g.node[enlace[0]]['escencialidad'] == 'escencial' and g.node[enlace
```

```
    [1]]['escencialidad'] == 'escencial':
                m = m + 1
        lista_de_m.append(m)

    m = sum(lista_de_m)/len(lista_de_m)
    alfa = ((n-m)/(g.number_of_edges()))*100

    return [alfa,n,m,nro_escenciales]
```

Cálculo de β

```
def coef_beta(g,n,m,nro_escenciales):

    nro_de_vueltas = 10000
    betas = []

    for node in g.nodes:
        betas.append([node,g.degree(node),0])

    for _ in range(nro_de_vueltas):
        #Lista de nodos con grados ordenada de mayor a menor sin nodos con grado cero
        nodo_grados=[]
        for node in g.nodes:
            nodo_grados.append([node,g.degree[node]])
        random.shuffle(nodo_grados)

        #Asigno (n-m) enlaces aleatorios y escencialidad a los nodos con esos enlaces
        nuevos_enlaces = []
        i = 0
        while len(nuevos_enlaces) < int(n-m):

            nodo = random.choice(nodo_grados)
            if nodo_grados[i][0] != nodo [0] and [nodo[0],nodo_grados[i][0]] not in nuevos_enlaces:
                nuevos_enlaces.append([nodo_grados[i][0],nodo[0]])
                i = i + 1

        #pongo a los nodos escenciales en una lista
        nodos_escenciales=[]
        for nod in nuevos_enlaces:
            nodos_escenciales.append(nod[0])
            nodos_escenciales.append(nod[1])
        nodos_escenciales = list(set(nodos_escenciales))

        #pongo a los nodos  no escenciales en una lista
        no_escenciales=[]
        for node in nodo_grados:
            if node[0] not in nodos_escenciales:
                no_escenciales.append(node[0])

        #Selecciono m nodos aleatoriamente y les asigno escencialidad
        random.shuffle(no_escenciales)

        #Creo una lista que contenga a los nodos escenciales debido a factores que
```

```
    no son los enlaces escenciales
        nuevos_escenciales = []
        if len(nodos_escenciales)<nro_escenciales:
            i = 0
            while i < nro_escenciales:
                nuevos_escenciales.append(no_escenciales[i])
                i = i + 1

        for i in range(len(betas)): #Beta para cada nodo[i] sera sum(r)/10000
            if betas[i][0] in nuevos_escenciales:
            #la probabilidad ser  el promedio de la cantidad de veces que cada nodo
            #fue escencial por selecci n aleatoria en las 10000 vueltas
                betas[i][2] = betas[i][2] + 1/nro_de_vueltas
    Beta = 0
    for i in range(len(betas)):
        Beta = Beta + betas[i][2]
    Beta = Beta*100/len(betas)

    return Beta
```

7. Tabla 5

Básicamente, a partir del modelo lo que uno quiere calcular es la probabilidad de que un par de nodos sea esencial o no simultáneamente; esto es la suma de la probabilidad de que ambos sean esenciales más la probabilidad de que ambos no lo sean, lo que resulta en la ecuación siguiente definida para todo par de nodos (n,m).

$$P = \sum_{n,m} P_n P_m + (1 - P_n)(1 - P_m) \tag{2}$$

Siendo P_n la probabilidad de que el nodo n sea escencial.

8. Referencias

[1] He X, Zhang J (2006) Why do Hubs tend to be essential in protein networks? PLoS Genetics 2 (6):e88.

[2] Zotenko E, Mestre J, O´Leary D.P., Przytycka T.M. (2008) PLoS Computational Biology 4 (8): e1000140