Análisis de la red de interacciones proteína-proteína de *Brucella melitensis*

C. Ríos, R. Sieira, A. Troiano, M. Villagran

Redes Complejas con Aplicaciones a Biología de Sistemas

Dpto. de Física, FCEyN, UBA

Resumen

Se estudió la red de interacciones proteína-proteína de la cepa Brucella Melitensis 16M, haciendo un análisis de funcionalidad biológica y esencialidad por comunidades.

1. Objetivos

En el presente trabajo se analizó una red de interacciones proteína-proteína (PPI) de la bacteria *Brucella melitensis* 16M. Una de las especies de bacterias patógenas intracelulares facultativas responsables de una enfermedad zoonótica conocida como brucelosis, que afecta a una amplia variedad de mamíferos salvajes y domésticos. Como objetivo general se propuso evaluar si la construcción de dicha red permite identificar clusters de proteínas relacionadas funcionalmente, para luego determinar si dichos clusters se hallan enriquecidos en proteínas esenciales.

En particular se propuso visualizar la red como un grafo, caracterizarla mediante algunos parámetros generales, identificar clusters, hacer un análisis de posibles relaciones funcionales dentro de los clusters mediante Gene Ontology, y hacer una estimación de esencialidad de los nodos de las comunidades.

2. Resultados y discusión

2.1. Construcción y caracterización de la red PPI

Se utilizó la base de datos STRING, la cual utiliza información de distintos tipos de fuentes (experimental, datos de la literatura, interacciones de proteínas ortólogas de distintos organismos, entre otras). Se descargaron los enlaces proteína-proteína correspondientes a la cepa *B. melitensis* 16M. Dichos enlaces cuentan con un peso de confianza experimental que abarca un rango de 0 a 1000. Para realizar un análisis estadístico sobre las características de las redes que se forman al seleccionar solo alguna subsección de la red total, tomando como criterio su confianza experimental, es útil conocer si la subred tiene características similares a la de la red completa. Por ejemplo, si la red completa presenta una distribución de grados con forma de ley de potencias en la cual existe un rango libre de escalas, es de

utilidad saber si las submuestras tienen un comportamiento similar, si la pendiente de la curva es la misma y si se conserva un rango con libertad de escalas. Como datos adicionales se puede clasificar la red dentro de las redes estándar, con pendiente $2 < \alpha < 3$, véase [1]. En la figura 1 se muestran distribuciones de grado para subredes, marcadas con distintos colores que representan el nivel de confianza experimental (exp), siendo celeste (exp $>$ 100), violeta (exp $>$ 200), rojo (exp $>$ 400), amarillo (exp $>$ 600) y verde (exp $>$ 800). Para estas distribuciones se realizaron ajustes a leyes de potencias con el algoritmo presentado en Alstott et al. (2014) [2], y los resultados se contrastaron con una distribución normal. Los resultados de este análisis se muestran en los cuadros 1 y 2 donde se presentan el grado mínimo y el grado máximo de las redes, el menor grado donde comienza la ley de potencias y el exponente de la ley de potencias. Al hacer la comparación con una distribución normal se obtienen la fuerza de la correlación (r) y el valor p, adicionalmente se muestra la distancia de Kolmogorov-Smirnov a la recta ajustada. Vale la pena mencionar que a pesar de que el grado máximo disminuye considerablemente al aumentar el rigor de selección, las pendientes de los ajustes siguen teniendo rangos que entran dentro de los valores esperados. El hecho de tener un mayor número de nodos de grado uno al aumentar el requisito experimental hace ver claramente que muchos de los enlaces presentados en la red completa carecen de sustento experimental suficiente. Adicionalmente, no solo aumenta el número de nodos con esta característica, también aumenta considerablemente la fracción de nodos totales que representan, esto se ve al comparar la red Exp $>$ 800, con un 77 % de sus nodos de grado uno, y la red total con solamente un 2 % de nodos de grado uno. Las grandes diferencias aquí expuestas, junto con la dificultad de realizar experimentos exhaustivos para la evaluación de la red completa dan mayor importancia a las predicciones que pueden salir de búsquedas como las presentadas en este trabajo.

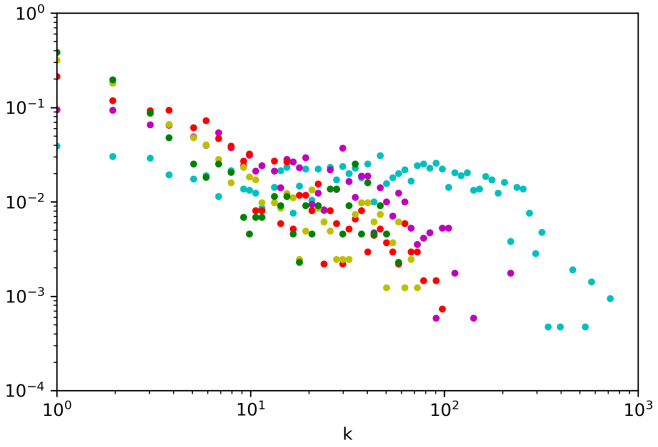

Figura 1: *Distribución de grados a diferentes cortes de confianza experimental. Los colores (confianzas experimentales) son celeste (exp $>$ 100), violeta (exp $>$ 200), rojo (exp $>$ 400), amarillo (exp $>$ 600) y verde (exp $>$ 800).*

Tabla 1: *Número de enlaces, de nodos, de nodos con grado uno y el nodo máximo de la red*

	Enlaces	Nodos	K=1	Kmax
TOTAL	364764	3058	64	1453
Ex > 0	68313	2107	164	687
Ex > 200	14064	1700	320	209
Ex > 400	5713	1360	580	94
Ex > 600	2853	816	520	70
Ex > 800	1523	437	336	56

Tabla 2: *Grado mínimo de la ley de potencias, el valor de α, la fuerza de la correlación r, el valor p de la comparación con una distribución normal y la distancia de Kolmogorov-Smirnov.*

	PL Kmin	PL alpha	R	p value	KS
TOTAL	152	4.12	-0,01	0,93	0.35
Exp > 0	30	2.35	-0,001	0,982	0.08
Exp > 200	11	2.62	-0,592	0,464	0.99
Exp > 400	1	1.60	-1,89	0,183	0.13
Exp > 600	1	1.87	0,036	0,958	0.06
Exp > 800	1	1.97	0,722	0,354	0.06

Tabla 3: *Parámetros generales de la red de B. Melitensis con score experimental mayor a 200.*

No. de nodos	1700
No. de enlaces	14064
$< k >$	16.6
k_{max}	209
k_{min}	1
Densidad	0.0097
C global	0.19
$< C >$	0.16
Diámetro	10

Luego de este análisis comparativo se eligió fijar la atención de este trabajo en la subred compuesta por enlaces con un score experimental mayor a 200. Para esta red se calcularon los parámetros que se observan en la tabla 3.

2.2. Clustering

Utilizando el algoritmo Infomap se identificaron 104 comunidades con tamaños que van desde los 2 a los 100 elementos, lo cual se muestra en la figura 2.

Figura 2: *Partición de la red según Infomap.*

2.3. Análisis de GO y esencialidad

Se obtuvieron los términos GO para las categorías *biological process* (BP), *cellular component* (CC) y *molecular function* (MF) para la cepa *B. melitensis* 16M a partir de la base de datos UniProt [3]. Utilizando un conjunto de scripts de Python desarrollados específicamente para este fin, se obtuvieron los ancestros de los términos GO y se calculó el factor de enriquecimiento como el cociente entre la proporción de genes asociados a un término particular de GO presente en cada comunidad con respecto a la proporción del número total de genes para dicha categoría en el genoma de la bacteria.

Para cada una de las 104 comunidades de la red se realizó un análisis de GO tomando en cuenta sólo los datos de enriquecimiento de términos GO con p-value menor o igual a $0,05$, corregidos por el método de Benjamini–Hochberg. Dado que el objetivo fue determinar posibles asociaciones funcionales entre proteínas, se analizaron los resultados de la categoría funcional BP (*biological process*). Para cada comunidad se eligió el término GO representativo más específico (de menor rango en la jerarquía de términos GO) presente en el mayor porcentaje de nodos, descartando aquellos términos observados en menos del 10 % de los

nodos. Como resultado se obtuvo enriquecimiento en términos GO específicos en 63 de las 104 comunidades (60 %). Estos resultados se muestran en la tabla A del Apéndice.

En las comunidades con término GO específico enriquecido por encima del punto de corte de 10 % se obtuvieron porcentajes de nodos asociados a dicho término con la distribución mostrada en la figura 3-a. Se observó un máximo de $96,8\,\%$ en la comunidad No. 12 (31 elementos) enriquecida en el término GO:1901362 "organic cyclic compound biosynthetic process", seguido por una distribución aproximadamente homogénea para el rango de 20 a 80 % de nodos asociados (ver fig. 3-a). A priori dicho resultado sugeriría que existen comunidades con un porcentaje sustancial de sus elementos contenidos dentro de una categoría GO definida. Sin embargo, como se puede observar en la fig. 3-b, la asociación a términos GO enriquecidos muestra una distribución con un alto sesgo hacia un número muy bajo de nodos, observándose una mayor frecuencia de 39 comunidades conteniendo entre 1 y 5 nodos asociados, seguida por 16 comunidades con 5-10 nodos asociados a términos GO.

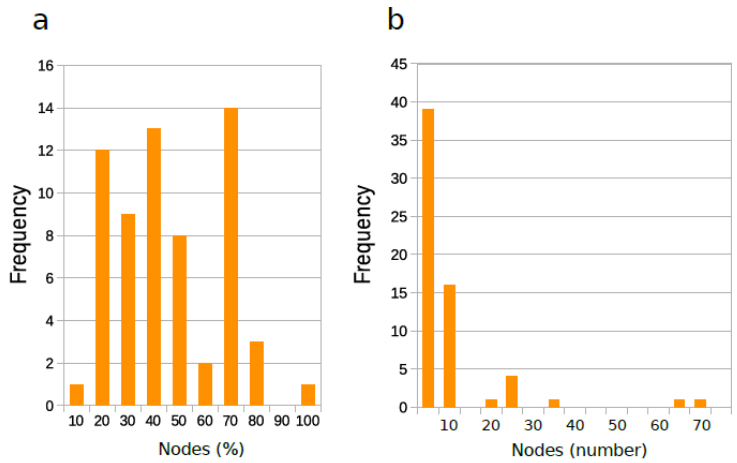

Figura 3: *En las 63 comunidades que poseen un término GO representado en más del* 10 % *de sus nodos: a) frecuencia de porcentaje del total de nodos de la comunidad asociados al término GO, b) frecuencia de número de nodos asociados al término GO en la comunidad.*

La ausencia de una base de datos que contenga las proteínas o interacciones entre proteínas que se determinen esenciales para la bacteria de *Brucella Melitensis* es un indicador de la falta de información que existe para dicho organismo. Una manera de subsanar este vacío es creando una lista de posibles candidatos a ser proteínas esenciales buscando sus contrapartes en algún organismo más estudiado. Se obtuvo una lista de proteínas esenciales de la cepa *Escherichia coli* MG1655 I de la base de datos Database of Essential Gene (DEG) [4]. Luego se utilizó la herramienta BLAST ([5]) para generar una *index database* y comparar la secuencia aminoacídica de cada una de las proteínas esenciales de *E. coli* MG1655 I en formato FASTA con las totalidad de secuencias predichas de proteínas de *B. melitensis* 16M. Los *hits* con mayor porcentaje de similitud y *E-value* $< 0,01$ se consideraron como ortólogos con misma función y carácter de esencialidad que en *E. coli* MG1655 I.

El análisis de esencialidad mostró que tanto en las 63 comunidades que poseen término GO enriquecido como en el total de comunidades, se obtuvo una distribución de porcentajes de nodos esenciales mostrada en las figuras 4-a y 5-a, respectivamente, observándose un 100 % sólo en 3 comunidades, 2 de las cuales contienen los términos enriquecidos GO:0006720 y GO:0055085 ("isoprenoid metabolic process" y "transmembrane transport" respectivamente,

ver tabla B del Apéndice). Dichas comunidades poseen un total de 3 elementos, de modo acorde con lo observado para la distribución de número de nodos esenciales por comunidad (figuras 4-b y 5-b), donde existe una mayor frecuencia de 39 comunidades con 1-5 elementos, seguida por 16 comunidades con 5-10 nodos esenciales. Es decir, la distribución de porcentajes de nodos esenciales observadas en las figuras 4-a y 5-a también muestra un alto sesgo hacia comunidades con pocos elementos. Como excepción se identificaron las comunidades No. 0, 1, 7, 17 y 36, con 12-44 nodos esenciales representando entre un 46 y un 72 % del total de sus elementos (asociados a los términos GO:0044271, GO:0006082, GO:0045454, GO:0008643 y GO:0042886; pertenecientes a las categorías "cellular nitrogen compound biosynthetic process", "organic acid metabolic process", "cell redox homeostasis", "carbohydrate transport" y "amide transport", respectivamente).

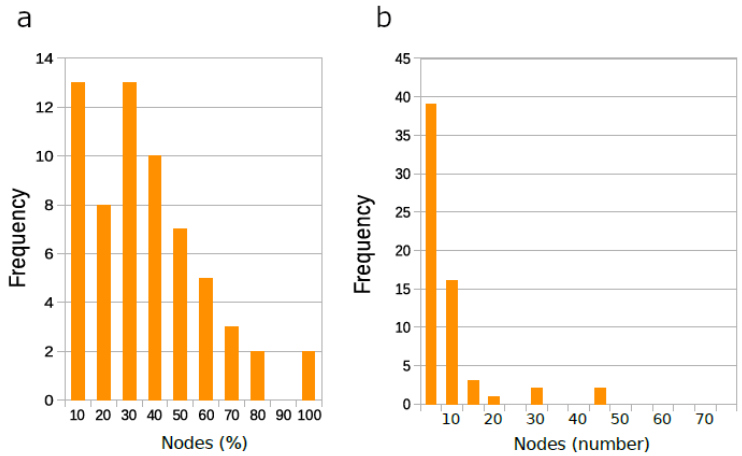

Figura 4: *En las 63 comunidades que poseen un término GO representado en más del* 10 % *de sus nodos: a) frecuencia de porcentaje de nodos esenciales en la comunidad, b) frecuencia de número de nodos esenciales en la comunidad.*

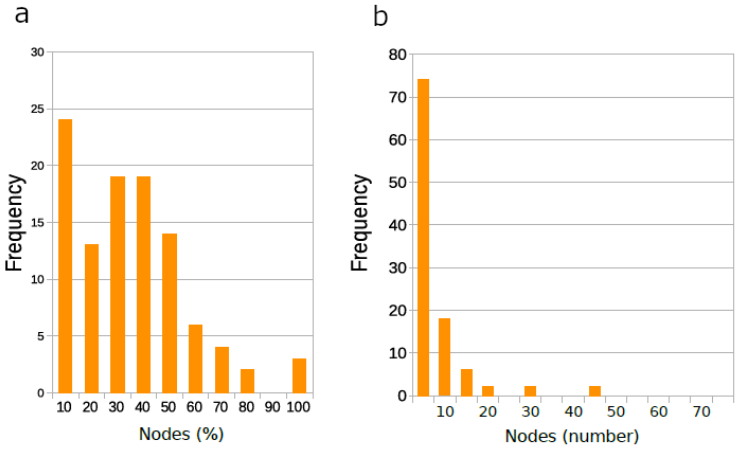

Figura 5: *En las 104 comunidades con o sin término GO asociado: a) frecuencia de porcentaje de nodos esenciales en la comunidad, b) frecuencia de número de nodos esenciales en la comunidad.*

3. Conclusiones

Según Zotenko *et al.* (2008) [6], en redes biológicas de PPI existen grupos de proteínas relacionadas funcionalmente entre sí, denominados *complex biological modules* (COBIMs), donde la mayoría de las proteínas esenciales se encontrarían preferencialmente en COBIMs enriquecidos en proteínas esenciales, denominados *essential* COBIMs (ECOBIMs). Parte de nuestra motivación para realizar el presente trabajo incluyó la identificación de clusters de proteínas de *Brucella* que pudieran cumplir con la hipótesis de Zotenko *et al.* (2008). Dado el pequeño tamaño de la comunidad mundial de investigación básica en *Brucella*, no existe una masa crítica de datos para la construcción de una lista de proteínas esenciales. Nuestro análisis propuesto para una red PPI en *B. melitensis* en principio permitiría identificar proteínas que interactúan entre sí, y de acuerdo a su grado de enriquecimiento en una función biológica particular y a su proporción de proteínas esenciales se podría inferir la posible esencialidad del resto de las proteínas que los componen, de acuerdo con la definición de COBIMs o ECOBIMs.

Como resultado observamos que en el total de comunidades de la red PPI de *B. melitensis*, tanto en proteínas asociadas funcionalmente como en proteínas posiblemente esenciales se observaron sesgos hacia comunidades muy pequeñas que contienen menos de 5 elementos. Como excepción se obtuvieron 5 comunidades de mayor tamaño (17 a 94 nodos), donde un alto porcentaje de sus elementos se hallan asociados a un término GO específico y poseen una alta probabilidad de ser esenciales. En dichos casos, de acuerdo a la hipótesis de Zotenko *et al.* (2008) se podría especular que dentro de estas comunidades existe una proporción de proteínas esenciales mayor a la estimada en base a la comparación con la información obtenida de *E. coli* MG1655 I en la base de datos DEG, lo cual podría determinarse experimentalmente. Sin embargo, consideramos que la presencia de 5 clusters con posible carácter de ECOBIMs sobre un total de 104 representa un número demasiado bajo, por lo cual especulamos que la red de PPI de *Brucella* no se ajustaría a la hipótesis de Zotenko *et al.* (2008) o bien los enlaces proteína-proteína obtenidos a partir de la base de datos STRING podrían contener una alta cantidad de información basada en observaciones en proteínas ortólogas de organismos no relacionados con *Brucella*, poco representativas de la biología de dicha bacteria.

Referencias

[1] Albert-László Barabási and Reka Albert. "Emergence of Scaling in Random Networks". Science, 286:509-512, October 1999.

[2] Jeff Alstott, Ed Bullmore and Dietmar Plenz. "Powerlaw: a Python Package for Analysis of Heavy Tailed Distributions". PLoS ONE, 9:e85777, January 2014.

[3] `https://www.uniprot.org/`

[4] `http://essentialgene.org/`

[5] `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

[6] E. Zotenko, J. Mestre, D. P. O'Leary, T. M. Przytycka. "Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection bet-

ween the Network Topology and Essentiality". PLoS Computational Biology, August 2008, Volume 4, Issue 8, e100001140.

Apéndice

- Tabla A: `https://github.com/andres1074/Redes-TP4/blob/master/tabla_go.pdf`
- Tabla B: `https://github.com/andres1074/Redes-TP4/blob/master/tabla_esencialidad.pdf`