Herramientas bioinformáticas para el análisis de resultados de experimentos de expresión génica

R. Massanet Vila[1,2,3], J.J. Gallardo Chacón[3], T. Padró Capmany[4], L. Badimon[4], P. Caminal Magrans[1,2,3], A. Perera Lluna[1,2,3]

[1]Dept. ESAII, Universitat Politècnica de Catalunya (UPC), Barcelona, España; {raimon.massanet, pere.caminal, alexandre.perera}@upc.edu
[2]Centre de Recerca en Enginyeria Biomèdica (CREB), Barcelona, España;
[3]CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), España; {joan.josep.gallardo}@upc.edu
[4]Centro de Investigación Cardiovascular (CSIC-ICCC), Barcelona, España; {tpadro,lbadimon}@csic-iccc.org

Resumen

1 Introducción

1.1 Contexto

Los experimentos de expresión génica diferencial estudian las diferencias en la expresión de los genes de un organismo bajo diferentes condiciones y son una de las herramientas más utilizadas en la práctica clínica y farmacológica para el estudio de la relación entre genes y procesos biológicos o patologías [1, 2]. Estos experimentos generalmente utilizan técnicas basadas en hibridación o chips de expresión (*microarrays*) [3]. Estas técnicas tienen la ventaja de que tienen un rendimiento muy alto, ya que permiten evaluar la expresión de una cantidad elevada de genes de una sola vez. Sin embargo solamente analizan los genes que se hallan en el chip de expresión. Por tanto, no permiten el análisis de genes que no hayan sido seleccionados para formar parte del experimento.

En este estudio se propone una metodología que combina datos de expresión diferencial con información de expresión entre proteínas. El objetivo es enriquecer los resultados de un experimento de expresión con nuevos genes que no fueron tenidos en cuenta en el diseño del experimento pero que podrían tener cierta relevancia en los procesos que muestran expresión diferencial. El método se basa en el estudio de las vías proteómicas que conectan genes con expresión diferencial significativa.

1.2 Definiciones y notación

A lo largo de este texto se utiliza la notación clásica en teoría de grafos, algunas de cuyas definiciones y notaciones se describen a continuación. Un grafo $G = (V, E)$ es una estructura formada por un conjunto finito y no vacío $V = \{v_1, v_2, \ldots, v_N\}$ y un conjunto E de pares de elementos de V $E = e_1, e_2, \ldots, e_M$, de forma que cada $e_k = \{v_{k1}, v_{k2}\}$. El orden de un grafo es el número de elementos de V, $N = |V|$. La medida de un grafo es el número de elementos de E, $M = |E|$. Los elementos de V se llaman nodos y los de E se llaman aristas. Alternativamente, el conjunto de nodos de un grafo G puede denotarse por $V(G)$ y el conjunto de aristas por $E(G)$. Si $e = \{u, v\}$ es una arista se dice que los nodos u y v son adyacentes ($u \sim v$). La arista e es incidente con los nodos u y v y viceversa. Si las aristas son pares no ordenados, es decir, si $(u, v) = (v, u)$ entonces se dice que el grafo es no dirigido. En caso contrario se dice que el grafo es dirigido. De aquí en adelante este texto considera solamente grafos no dirigidos. El grado de un nodo v de un grafo G se denota por $g(v)$ y es el número de aristas que inciden en v: $g(v) = |\{u : (v, u) \in E(G)\}|$. Un camino en un grafo G es una secuencia alternativa de nodos y aristas de la forma $\{v_0, e_1, v_1, e_2, v_2, \ldots, v_{n-1}, e_n, v_n\}$ con $v_i \in V(G)$ para $i = 0, \ldots, n$ y $e_j \in E(G)$ para $j = 0, \ldots, n$ y donde los nodos v_i y v_{i-i} son adyacentes mediante la arista e_i. El número de aristas de un camino se denomina longitud del camino. Un camino mínimo entre un nodo origen v_o y un nodo destino v_d es un camino $v_0 = v_o$ y $v_n = v_d$ y no existe ningún otro camino entre v_o y v_d de longitud menor. En este texto se denota por SP_{uv} al conjunto de todos los caminos mínimos (*shortest paths*) entre los nodos u y v.

2 Materiales y métodos

2.1 Datos

Los análisis de expresión génica diferencial generan resultados como los que se muestran, a modo de ejemplo, en la Tabla 1. Estos resultados muestran, para cada proteína incluida en el estudio, un valor de expresión diferencial, obtenido mediante *software* estándar en el campo. Si este valor es mayor que 1 se considera que el gen en cuestión muestra una expresión mayor de

lo normal (SOBRE) y si es menor que 1 se considera que la expresión es menor de lo normal (SUB). Si la tendencia se repite en los diferentes experimentos realizados se asigna el gen a una clase u otra. Si en los diferentes experimentos el gen muestra tendencias diferentes se clasifica como ND, no diferencial.

	Id. Swissprot	Clase	Expresión
65	P10809	SOBRE	1.83
11	P09601	SOBRE	1.62
16	P07101	ND	0.57
103	P35354	ND	1.56
46	P04792	SUB	0.43
38	Q16658	SUB	0.36

Tabla 1: Fragmento del resultado de expresión génica diferencial utilizado en este trabajo.

Un gen se considera que está significativamente sobreexpresado si su valor de expresión es mayor que 1.5. De forma simétrica, se considera que está significativamente subexpresado si su valor de expresión es menor que 0.67.

Los datos contienen un total de 107 valores de expresión para un número similar de proteínas. De estos valores, un 43% fueron clasificados como ND, 35.5% como SOBRE y un 21.5% como SUB. Un total de 22 valores muestran expresión diferencial significativa. [1]

2.2 Metodología

El método propuesto combina los resultados de un experimento de expresión génica con información semántica y con información de interacción entre proteínas. El objetivo es identificar los procesos biológicos que muestran una expresión diferencial, así como otros posibles genes que no aparecen en el estudio de expresión y que podrían ser susceptibles de mostrar expresión diferencial.

El método consta de tres pasos diferenciados. En primer lugar se identifican los genes con expresión diferencial significativa. En el segundo paso se buscan todos los caminos de longitud mínima que conectan estos genes. En el tercer y último paso se forma una red de proteínas a partir de esos caminos y se estudia su coherencia semántica. Todo el método fue desarrollado utilizando el lenguaje de programación estadística R [4]. A continuación se detalla cada uno de estos pasos.

El primer paso fue identificar los genes con expresión diferencial significativa y obtener el identificador correspondiente en la base de datos de IntAct. Se identificaron 22 proteínas con expresión diferencial. Se construyó un grafo G sin interacciones formado por estas proteínas. A continuación se procedió al enriquecimiento de G con nuevas interacciones. En cada iteración se añadió un nuevo nivel de interacciones. Es decir, se añadieron a G nuevas proteínas si éstas interactuaban con alguna de las proteínas $V(G)$. El proceso se repitió hasta que G estuvo formado por una sola componente conexa. A este grafo conexo se lo denominó G_c. Esta red tiene la particularidad de que enlaza todas las proteínas que muestran expresión diferencial significativa en el experimento de expresión.

En el segundo paso se buscó para todo par de nodos de $V(G_c)$, (u,v), el conjunto de caminos de longitud mínima con origen en u y destino en v. A este conjunto se lo denominó SP_{uv}. A continuación se creó el grafo G_{sp} definido por la unión de todos los SP_{uv}:

$$G_{sp} = \bigcup_{u,v \in V(G_c)} SP_{uv} \qquad (1)$$

Este grafo tiene la particularidad de que contiene todas las vías proteómicas mínimas entre cualquier par de proteínas de G_c.

Finalmente, en el tercer paso, se estudiaron los valores de expresión diferencial para las proteínas $V(G_{sp})$ y se evaluó su coherencia semántica. Para el cálculo de la coherencia semántica se obtuvieron las etiquetas semánticas de las proteínas $V(G_{sp})$ de la base de datos *Gene Ontology Annotation* (GOA) [5]. A continuación se calculó la similitud semántica entre todo par de proteínas de $V(G_{sp})$ usando la biblioteca para R *GOSemSim* [6]. Para realizar el cálculo se utilizó la medida de Wang [7]. La similitud total entre dos proteínas se estimó como la media de las tres similitudes semánticas desde los puntos de vista contemplados por *Gene Ontology* (GO) [8]:

$$sim(u,v) = \frac{\sum_{a \in \{BP,MF,CC\}} sim_a(u,v)}{3} \qquad (2)$$

donde a es el aspecto desde el que se calcula la similitud semántica y puede tomar los valores: BP (proceso biológico), MF (función molecular) y CC (componente celular). La coherencia semántica C de la red G_{sp} se calculó como la media de todas las similitudes semánticas dos a dos:

$$C(G_{sp}) = \frac{\sum_{v_i,v_j \in V(G_{sp}) \wedge j>i} sim(v_i,v_j)}{N(N-1)/2} \qquad (3)$$

donde N es el número de proteínas en la red $|V(G_{sp})|$. Con el fin de evaluar la significación de esta medida se construyó una distribución nula de coherencias semánticas $D_N(C)$. Para ello se seleccionaron de la base de datos GOA 50 muestras aleatorias de N proteínas y se calculó su coherencia semántica. Para obtener una estimación de la plausibilidad de que $C(G_{sp})$ perteneciera a $D_N(C)$ se modeló la distribución de la última con métodos basados en *kernel* y se obtuvo un p-valor sobre esta distribución [9].

3 Resultados

El grafo G_c generado a partir del conjunto de 22 proteínas con expresión diferencial está formado por

[1] Adquisició? Preguntar a JJ o a TP.

4439 nodos y 18243 interacciones. El tamaño de esta red hace muy difícil su estudio y comprensión, así como su visualización. Además, por construcción, G_c puede contener un número elevado de nodos no relacionados con los procesos relevantes.

El grafo G_{sp}, formado por todos los caminos mínimos de G_c está formado por 127 nodos y 281 interacciones. Esta red se muestra en la Figura 1.

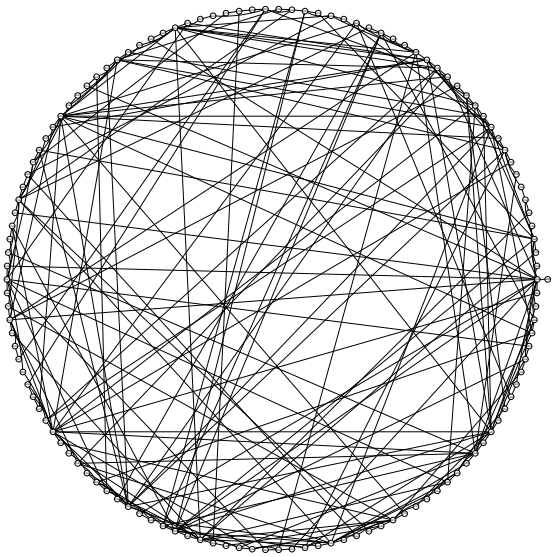

Figura 1: *Grafo formado por la unión de todos los caminos mínimos entre pares de nodos del grafo G_c.*

De las 127 proteínas contenidas en G_{sp} solamente 23 de ellas aparecen en los resultados de expresión diferencial del estudio inicial. Dado que los resultados del experimento contenían 22 proteínas con expresión diferencial significativa, esto significa que solamente se añadió una proteína adicional a la red de interacciones. La proteína añadida había sido clasificada como subexpresada, pero con un valor de expresión diferencial no significativo. Por tanto, la red G_{sp} no contiene ninguna proteína que en el experimento fuera clasificada como no diferencial (ND). Este resultado indica que las proteínas con resultados de expresión no diferenciales, no forman parte de las vías proteómicas de las proteínas con expresión diferencial conocida. En cambio sí forma parte de esas vías una proteína con expresión diferencial no incluida en la construcción de la red. Esto sugiere que los procesos biológicos afectados por el experimento tienen un grado de solapamiento relativamente bajo con el resto de procesos biológicos, desde un punto de vista de redes de interacción entre proteínas.

Los resultados sugieren que la red G_{sp} contiene las vías proteómicas que están siendo afectada por las condiciones del estudio. Las 104 proteínas de $V(G_{sp})$ apuntan a posibles futuros marcadores para nuevos estudios de expresión génica que confirmen estos resultados. La Tabla 3 muestra las proteínas de $V(G_{sp})$ con grados elevados. Los nodos de grados elevados en la red G_{sp} representan proteínas que se encuentran en numerosas vías proteómicas que conectan las proteínas de G_c. Por tanto podrían tener un papel destacado en los procesos biológicos contenidos en la red.

La coherencia semántica obtenida para la red ($C(G_{sp})$) fue de 0.21. La tabla 2 muestra la descriptiva de la distribución nula de la coherencia semántica ($D_N(C)$). Esta distribución no muestra diferencias significativas con una distribución normal (p-valor de 0.42 en una prueba de Kolmogorov-Smirnov). El valor obtenido para la probabilidad de que $C(G_{sp})$ pertenezca a $D_N(C)$ fue de $1.93e-17$. El valor de coherencia obtenido para la red G_{sp} es claramente y estadísticamente mayor de lo normal. Este resultado indica que las proteínas de la red conforman un conjunto relativamente homogéneo de proteínas involucradas en los mismos procesos biológicos, con funciones moleculares similares y en componentes celulares próximos. La red obtenida podría contener los procesos biológicos afectados por las condiciones del experimento de expresión génica realizado, o parte de ellos.

Min	**1er Q**	**Media**	**3er Q**	**Max**
$9.08e^{-3}$	$2.58e{-2}$	$4.16e^{-2}$	$5.24e^{-2}$	$1.02e^{-1}$

Tabla 2: *Estadística descriptiva de la distribución nula de la coherencia semántica.*

4 Conclusión

En este trabajo se ha mostrado que la combinación de diferentes herramientas bioinformáticas como teoría de grafos aplicada al estudio de redes de interacciones entre proteínas o teoría de la información aplicada al estudio de anotaciones semánticas de genes, pueden ser de gran utilidad en el análisis de resultados de experimentos de expresión génica. La metodología aplicada ha devuelto resultados estadísticamente coherentes desde el punto de vista semántico y desde el punto de vista de la expresión génica. Estas herramientas dependen enormemente de la completitud de las bases de datos de las que obtienen la información. A medida que estas bases de datos incorporan nueva información y depuran la que ya contienen, las herramientas aquí descritas deberían obtener mejores resultados y ser más útiles.

5 Agradecimientos

Los autores agradecen el apoyo recibido por parte del Ministerio de Educación y Ciencia a través del programa Ramón y Cajal y TEC2007-63637/TCM así como del Instituto de Salud Carlos III a través de la iniciativa CIBER-BBN en Bioingeniería , biomateriales y nanomedicina.

	Id. Uniprot	Proteína	Gen
1	O15264	Mitogen-activated protein kinase 13	MAPK13
2	O60739	Eukaryotic translation initiation factor 1b	EIF1B
3	P01106	Myc proto-oncogene protein	MYC
4	P04406	Glyceraldehyde-3-phosphate dehydrogenase	GAPDH
5	P04792	Heat shock protein beta-1	HSPB1
6	P06753	Tropomyosin alpha-3 chain	TPM3
7	P07355	Annexin A2	ANXA2
8	P08238	Heat shock protein HSP 90-beta	HSP90AB1
9	P08670	Vimentin	VIM
10	P10809	60 kDa heat shock protein, mitochondrial	HSPD1
11	P13569	Cystic fibrosis transmembrane conductance regulator	CFTR
12	P25786	Proteasome subunit alpha type-1	PSMA1
13	P30480	HLA class I histocompatibility antigen, B-42 alpha chain	HLA-B
14	P40337	Von Hippel-Lindau disease tumor suppressor	VHL
15	P49720	Proteasome subunit beta type-3	PSMB3
16	P61981	14-3-3 protein gamma	YWHAG
17	P62993	Growth factor receptor-bound protein 2	GRB2
18	Q14164	Inhibitor of nuclear factor kappa-B kinase subunit epsilon	IKBKE
19	Q15047	Histone-lysine N-methyltransferase SETDB1	SETDB1
20	Q16658	Fascin	FSCN1

Tabla 3: *Nodos de grado elevado en* G_{sp}.

Referencias

[1] J. Chen, "Key aspects of analyzing microarray gene-expression data," *Pharmacogenomics*, vol. 8, no. 5, pp. 473–482, 05/01 2007.

[2] J. DeRisi, V. Iyer, and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, October 24 1997.

[3] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, October 20 1995.

[4] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org

[5] D. Barrell, E. Dimmer, R. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The goa database in 2009–an integrated gene ontology annotation resource," *Nucl.Acids Res.*, p. gkn803, October 2008.

[6] G. Yu, "Gosemsim: Go-terms semantic similarity measures," R package version 1.2.0.

[7] J. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, May 15 2007.

[8] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[9] Harrell FE. Jr. with contributions from many other users., "Hmisc: Harrell miscellaneous," 2009, r package version 3.7-0. [Online]. Available: http://CRAN.R-project.org/package=Hmisc