Enriquecimiento mediante vías metabólicas de datos de Cromatografía Líquida- Espectrometría de Masas a través de análisis espectral de grafos

S. Picart[1,2], F.Fernández[1,2], M. Vinaixa[3,4,5], O. Yanes[3,4,5], A.Perera[1,2]

[1]Dpto. ESAII, Centre for Biomedical Engineering Research, UPC, Barcelona, España. [2]CIBERbbn. [3]Centre for Omics Sciences, Rovira i Virgili University, Reus, España. [4]CIBERDEM. [5]Department of Electronic Engineering, URV.

Resumen

Una de las técnicas experimentales más extendidas en el ámbito de investigación biológica y la química analítica es la Cromatografía Líquida – Espectrometría de Masas, CL/EM, cuya salida informa sobre los compuestos presentes en las muestras mediante una técnica de separación física acoplada a una separación en función de la relación carga-masa. Las técnicas de enriquecimiento de vías metabólicas son preciadas en el tratamiento de conjuntos extensivos de datos, puesto que traducen esta información sobre computestos en términos de vías metabólicas a la vez que reducen el ruido estadístico. Las vías metabólicas son fuente de conocimiento por su estrecha relación con los mecanismos biológicos.

Este trabajo propone una nueva técnica de enriquecimiento de datos obtenidos en CL/EM mediante una estrategia en dos bloques. El primero consiste en plasmar la base de datos Kyoto Encyclopedia of Genes and Genomes en grafos interpretables. El segundo trata de aplicar algoritmos de difusión de calor y PageRank sobre dichos grafos, con el objetivo de llevar a término el enriquecimiento.

Estos procedimientos se han aplicado en un caso real y sus resultados coinciden con los de validación funcional.

1. Introducción

La Cromatografía Líquida – Espectrometría de Masas, CL/EM, es el acoplamiento entre un cromatógrafo líquido y un espectrómetro de masas. Las moléculas que lo atraviesan ganan o pierden átomos o moléculas, volviéndose así aductos. Desafortunadamente, los patrones de su formación introducen incertidumbre. Aun así, la CL/EM proporciona un amplio rango de detección y se consolida como un procedimiento para ampliar los conocimientos sobre el metabolismo.

La CL/EM proporciona un espectro con masas, intensidades y tiempos de retención. Su tratamiento permite hallar los metabolitos significativamente afectados en un experimento caso-control, que en adelante serán referidos como compuestos significativos. Sin embargo, el investigador requiere de información a nivel de vías metabólicas para entender los sucesos biológicos presentes. Las técnicas de enriquecimiento de vías metabólicas le asistirán en su cometido.

Este trabajo se propone hacer un mejor uso de la información disponible en una base de datos de anotaciones, con la concepción de métodos que ofrezcan resultados completos e interpretables. La base de datos minada es KEGG [1], siglas de Kyoto Encyclopedia of Genes and Genomes, que ofrece anotaciones curadas para múltiples ómicas.

La técnica habitual para materializar el enriquecimiento es el test hipergeométrico. Éste otorga a cada vía metabólica un p-valor para indagar si la vía está sobrerepresentada, es decir, si está directamente involucrada con un gran número de los compuestos significativos. Dicha técnica ha sido ampliamente usada con datos de expresión génica y del proyecto Gene Ontology [2].

A continuación se referencian otros métodos en el estado de arte del enriquecimiento. ProbMetab [3] sugiere un enfoque bayesiano. MetaboAnalyst [4] es una interfaz web que usa técnicas como MSEA [5] y el test de sobrerepresentación. IPA [6] es un software comercial que propone agrupamientos en redes. Por último, también existe una perspectiva bayesiana combinada con el Análisis de Componentes Independientes [7].

Este trabajo aplica los algoritmos de difusión de calor y PageRank en grafos para hallar los nodos más relevantes y su interpretación biológica.

2. Materiales y métodos

La figura 1 muestra el proceso de trabajo seguido. Los puntos enfatizados en verde se han desarrollado bajo el entorno R [8] [9].

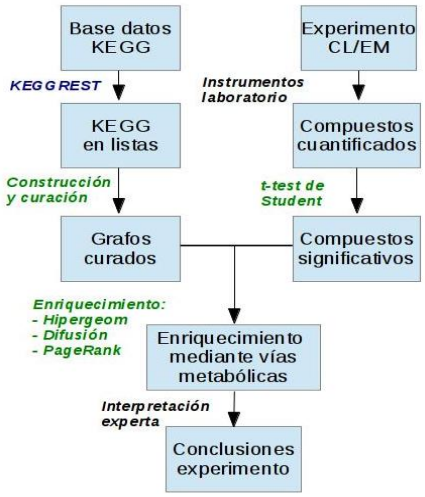

Figura 1. *Proceso de trabajo*

1) Construcción de grafos a partir de KEGG

KEGG contiene ciertas categorías de interés en el estudio del metabolismo. De entre ellas, se han seleccionado: vías metabólicas, módulos, enzimas, reacciones y compuestos.
Los compuestos son el continente de los metabolitos proporcionados por CL/EM. Las reacciones proporcionan información sobre los compuestos involucrados y la enzimas que la pueden catalizar. Asimismo, las enzimas son catalíticos biológicos macromoleculares y se han incluido por tener un papel relevante. Los módulos son ciertas reacciones que tienen una función juntas; suelen ser una pequeña porción de una vía metabólica. Finalmente, las vías metabólicas son conjuntos de reacciones entre compuestos, catalizadas por enzimas, con un significado biológico comprensible e identificable.
Los grafos se han elaborado usando las categorías listadas, véase la figura 2.

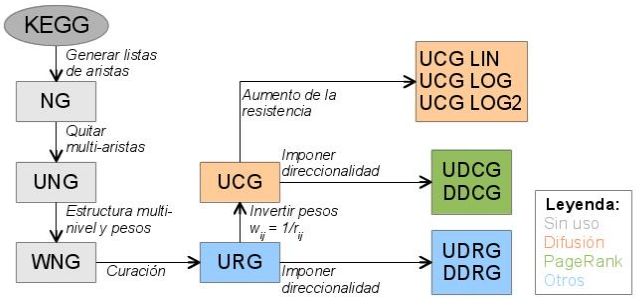

Figura 2. *Resumen de los grafos concebidos*

El primer grafo, **NG**, emerge usando las entradas de KEGG como nodos y sus referencias como aristas de peso unitario. Lo sucede **UNG** si se eliminan las multi-aristas. En este punto, se introduce una estructura multi-nivel, como esboza la figura 3, cuyo resultado es **WNG**, un grafo con pesos.

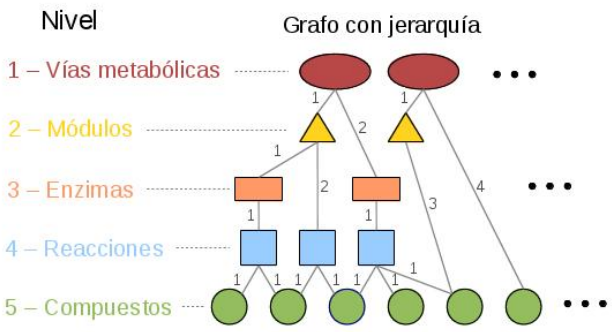

Figura 3. *Estructura multi-nivel con los pesos en las aristas, basados en la distancia entre niveles.*

El grafo **WNG** permite un proceso de curación sobre las aristas redundantes, dando lugar a **URG** y **UCG**. Los pesos de las aristas de estos grafos curados expresan disimilitud y similitud, respectivamente.

A continuación, ambos grafos pueden dirigirse ascendente (**UD**) o descendentemente (**DD**) para aplicar algoritmos específicos.
Por último, el grafo UCG contiene nodos poco informativos pero conectados en demasía, como es el caso del nodo "agua". Para mitigar este fenómeno se ha aumentado los pesos de las aristas en función de los grados de los nodos que relacionan, de forma lineal (**LIN**), logarítmica (**LOG**) o doblemente logarítmica (**LOG2**)

2) Materiales

Se dispone de un experimento caso-control de CL/EM para enriquecer. En los casos se ha modificado la vía metabólica del glutatión; el uso de marcadores isótopos lo ratifica y revela tres vías metabólicas adicionales afectadas.
Este trabajo persigue poder replicar tal resultado mediante el enriquecimiento. Después de la CL/EM, un Análisis de Componentes Principales, en figura 4, separa perfectamente caso de control. Seguidamente, un t-test (0.95) halla los compuestos significativos.

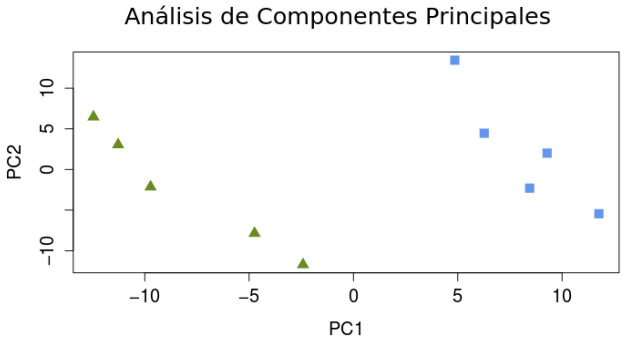

Figura 4. *ACP clasificando control (triángulos verdes) de caso (cuadrados azules). PC1 y PC2 explican el 55% de la varianza total, sugiriendo que ambos grupos presentan diferencias en ciertos compuestos.*

3) Enriquecimiento de CL/EM mediante grafos

3.1) Difusión de calor

El método de difusión de calor es la primera tentativa para mejorar el test hipergeométrico. La variante aquí usada se basa en la formulación en diferencias finitas [10], siendo un símil de la difusión en un objeto mallado y caracterizado (1). Concretamente, se estudia el estado estacionario (2).

$$T^{n+1} = T^n + DTC \cdot \left[KI \cdot T^n + KC \cdot TC + G \right] \quad (1)$$

$$T^{\infty} = -KI^{-1} \cdot \left[KC \cdot TC + G \right] \quad (2)$$

En la expresión (2) T^{∞} es la temperatura de los nodos, KI es la matriz de conductancias, KC es la matriz de conductancias hacia los nodos de contorno, TC es la temperatura de los nodos de contorno y G es la generación de calor para cada nodo.

La figura 5 detalla el uso del grafo UCG como objeto mallado y la imposición de condiciones de contorno. La salida del algoritmo contiene las temperaturas de cada nodo del grafo, incluidas las de las vías metabólicas.

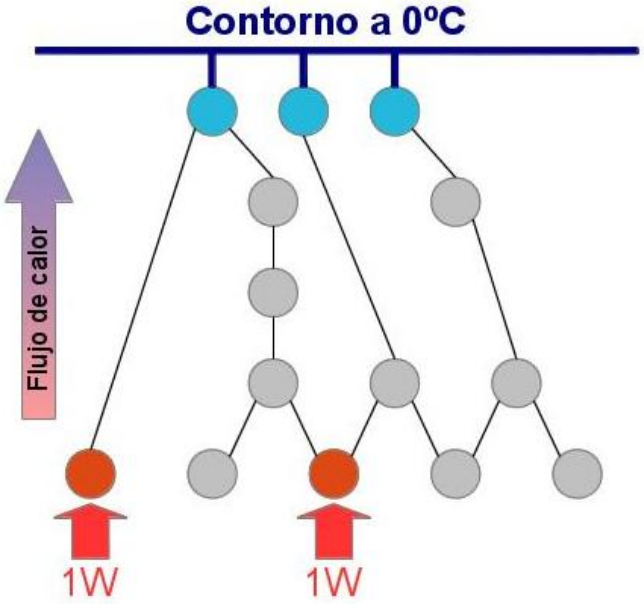

Figura 5. *Difusión de calor en el grafo UCG. Inicialmente la totalidad del grafo reposa en equilibrio a 0ºC, con sus vías metabólicas unidas a un contorno a 0ºC. A continuación, o bien se fuerza generación de calor en los compuestos significativos (ejemplificado en la figura) o bien se unen tales compuestos a otro contorno a, por ejemplo, 1ºC. En ambas alternativas, los nodos más calientes presentan un papel relevante.*

3.2) PageRank

El algoritmo PageRank [11] es otra alternativa estudiada al test hipergeométrico. PageRank es un algoritmo web que otorga una puntuación a cada página según su relevancia. Internet se convierte en un grafo cuyos nodos son las páginas, unidos por una arista dirigida en caso de existir un hipervínculo. La puntuación de una página depende del número de páginas apuntándola y, además, de la puntuación de las mismas.

PageRank tiene un trasfondo de marchas aleatorias en grafos. El navegante empieza en un nodo al azar, dictado por la *distribución a priori*, y escoge pasos en el grafo. En cada paso se plantea continuar la marcha con probabilidad *d* (*factor de amortiguamiento*), o bien abandonarla, con probabilidad *1-d*. Así, PageRank computa una distribución de probabilidad sobre todos los nodos del grafo, coincidente con las puntuaciones.

En este trabajo se toma el grafo UDCG, pues permite que las marchas empiecen en el nivel inferior (compuestos) y avancen hacia el superior (vías metabólicas). La distribución a priori es tal que las marchas se inician únicamente en los compuestos significativos. El factor de amortiguamiento se concreta en referencia al valor usado en el artículo del PageRank y se modifica ligeramente según las características del grafo.

El método expuesto otorga puntuaciones a todos los nodos, en particular a las vías metabólicas.

3. Resultados

El grafo URG, piedra angular del trabajo, tiene las características de la tabla 1. Asimismo, la figura 6 plasma su distribución de grados, con notable similitud a las redes libres de escala.

Propiedad	Valor
Número de nodos	22 725
Número de aristas	57 703
Componentes conexas	21
Tamaño de la mayor componente conexa	22 644

Tabla 1. *Propiedades del grafo URG. La mayor componente conexa es prácticamente todo el grafo, indicador favorable.*

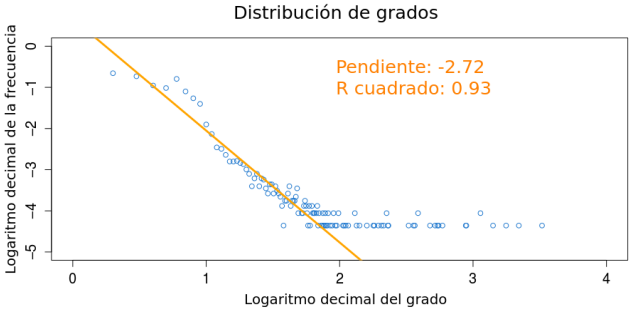

Figura 6. *Distribución de grados del grafo URG.*

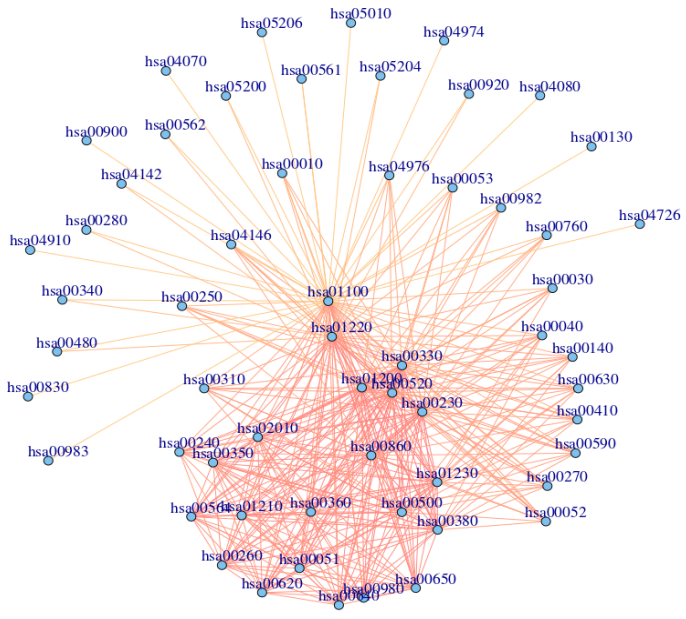

Figura 7. *Se ha graficado las menores resistencias equivalentes (percentil 1%). Algunas vías metabólicas se muestran íntimamente relacionadas, sugiriendo así que tienen relación biológica. Por ejemplo, se aprecia que la vía metabólica hsa01100 está conectada a muchas otras, lo cual se confirma sabiendo que se trata de una visión global del metabolismo en KEGG.*

Dando un paso más allá, se puede asimilar el grafo URG a una red eléctrica cuyas resistencias se identifican con los pesos de las aristas. A continuación, se puede calcular la resistencia equivalente entre cualquier pareja de nodos [12], incluyendo vías metabólicas. La representación de esta función de disimilitud se aprecia en la figura 7.

Respecto a los algoritmos, las tablas que siguen demuestran que los tres métodos descubren las cuatro vías afectadas entre las quince mejores puntuaciones.

El test hipergeométrico retorna una lista de vías metabólicas con sus respectivos p-valores corregidos mediante FDR. La tabla 2 selecciona las vías metabólicas afectadas acompañada de una breve descripción y de su posición en la ordenación.

Vía metabólica	Metabolismo involucrado	p-valor	Posición
hsa00480	Glutatión	4.66e-8	1
hsa00270	Cisteína y metionina	4.08e-7	2
hsa00330	Arginina y prolina	3.48e-6	4
hsa00250	Alanina, aspartato y glutamato	2.28e-3	14

Tabla 2. *Resultados del test hipergeométrico.*

El resultado de la difusión de calor con aumento lineal en el grafo UCG e imponiendo contorno a 10ºC para los compuestos significativos se sintetiza en la tabla 3. Este método permite la obtención de información adicional sobre el grafo, como pueden ser las enzimas más calientes para una vía metabólica.

Vía metabólica	Enzimas relacionadas		T [ºC]	Posición
hsa00480	EC: 2.3.2.4	EC: 1.8.1.13	2.080	1
hsa00270	EC: 3.5.1.31	EC: 1.8.4.1	1.424	3
hsa00330	EC: 2.3.1.1	EC: 2.7.2.8	1.458	2
hsa00250	EC: 2.3.1.17	EC: 3.5.1.15	1.226	6

Tabla 3. *Resultado de la difusión de calor con el grafo UCG penalizando linealmente los pesos y usando un contorno a 10ºC.*

El cálculo de las puntuaciones PageRank en el grafo UDCG para las vías metabólicas, con normalización a suma unitaria, se reúne el la tabla 4. La distribución a priori se ha escogido uniforme en los compuestos significativos. El artículo original sugiere un factor de amortiguamiento $d=0.85$ para el navegante, pero como UDCG permite cuatro pasos a lo sumo se ha fijado $d=0.7$.

Vía metabólica	Puntuación	Posición
hsa00480	5.53e-2	2
hsa00270	3.17e-2	5
hsa00330	2.83e-2	6
hsa00250	2.33e-2	8

Tabla 4. *Resultados de PageRank.*

4. Conclusiones

Los métodos propuestos han hallado las vías metabólicas afectadas en el experimento realizado. Tanto la difusión de calor como el PageRank han estado al nivel del test hipergeométrico, referencia en el ámbito. Adicionalmente, ambos permiten generar una salida más completa con información del resto del grafo. Las enzimas relacionadas con las vías metabólicas son un ejemplo de ampliación deseable, también aplicable para módulos, reacciones y compuestos. En definitiva, estas herramientas aportan resultados de mejor interpretabilidad para el investigador.

5. Agradecimientos

Esta investigación ha disfrutado de las subvenciones 2014 SGR 1063, TEC2010-20886-C02-02 y TEC2010-20886 C02-01. CIBER-BBN es una iniciativa de ISCIII, España. F. Fernández-Albert agradece a EVALXARTA-UB y a la Agència de Gestió d'Ajuts Universitaris i de Recerca, AGAUR (Generalitat de Catalunya) por su ayuda financiera.

6. Referencias

[1] H. Ogata [et al.] (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1): 29-34.

[2] Da Wei Huang, Brad T. Sherman and Richard A. Lempicki (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1): 1-13.

[3] Ricardo R. Silva [et al.] (2014). ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*, 30(9): 1336-1337.

[4] Jianguo Xia [et al.] (2009), MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, 37: W652-60.

[5] Jianguo Xia and David S. Wishart (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(suppl 2): W71-77.

[6] Apostolos Zaravinos [et al.] (2014). Ingenuity Pathway Analysis (IPA).

[7] Jan Krumsiek [et al.] (2012). Bayesian Independent Component Analysis recovers pathway signatures from blood Metabolomics data. *Journal of Proteome Research*, 11(8): 4120-4131.

[8] R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

[9] Gábor Csárdi and Tamás Nepusz (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

[10] Lluis Albert Bonals (2011). Transferència de calor. Apunts de classe.

[11] Lawrence Page [et al.] (1999). The PageRank citation ranking: bringing order to the Web.

[12] R. B. Bapat (2004). Resistance matrix of a weighted graph.