BIGO: Mejora del análisis de enriquecimiento en grupos de genes

Aurelio López Fernández

Resumen—El análisis de enriquecimiento de genes permite hacer una validación, basada en conocimiento biológico previo, de los resultados obtenidos por técnicas de agrupación de genes (Clustering y Biclustering) sobre bases de datos de expresión genética. En este artículo se presenta BIGO, una herramienta que mejora dicho análisis aportando nuevas informaciones que permiten acotar mejor el estudio y generar nuevas conclusiones.

Palabras Clave—Análisis de enriquecimiento, Validación biológica, Ontologizer, BIGO, Clustering, Biclustering.

1. Introducción

La bioinformática surge por la necesidad de estudiar la cantidad masiva de información biológica que se genera en la actualidad. Esta disciplina pasa a ser una ciencia al aportar la capacidad de análisis y la creación de modelos predictivos para los sistemas biológicos [1]. Una de las aplicaciones de la bioinformática es el análisis de la expresión genética, es decir, el estudio de la cantidad de ARNm que genera un conjunto de genes a partir de un número determinado de muestras o condiciones experimentales (diferentes individuos, tejidos cancerosos/sanos...) [2].

La información correspondiente a la expresión genética se almacena en microarrays, proporcionando información de la actividad de un conjunto de genes en un momento determinado. Por tanto, los microarrays permiten la comprensión de la regulación de genes así como el desarrollo y evolución de las enfermedades, por ejemplo, el estudio de por qué algunas células aumentan de forma incontrolada en casos de cáncer [3].

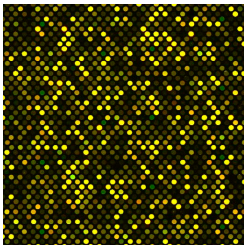

Figura 1: Sección de un microarray

En la figura 1, podemos observar un microarray en el que cada punto de color representa a un gen en particular (filas), mientras que su tonalidad hace referencia a la cantidad de ARNm expresado bajo una condición experimental concreta (columnas).

Esta información se dispone en matrices, M = $\{w_{ij} | 1 \leq i \leq n, 1 \leq j \leq m\}$, donde cada gen corresponde a una fila, F = $\{f_1, f_2, ..., f_n\}$, y cada muestra experimental a una columna, C = $\{c_1, c_2, ..., c_n\}$. Por lo que cada elemento de la matriz w_{ij} representa una cantidad de ARNm sobre un gen i ante una muestra experimental j.

$$M = \begin{pmatrix} w_{11} & w_{12} & ... & w_{1m} \\ w_{21} & w_{22} & ... & w_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & ... & w_{nm} \end{pmatrix}$$

Figura 2: Matriz de expresión genética

1.1. Clustering

El Clustering es de las técnicas más utilizadas para el estudio de la expresión genética y su objetivo es la obtención de agrupaciones de genes o muestras experimentales que compartan un gran porcentaje de sus características. Así, el objetivo puede ser, obtener una agrupación de genes en función de su expresión bajo determinadas condiciones o la obtención de una agrupación de condiciones basadas en la expresión de un número de genes. Estas agrupaciones se denominan clusters.

Una de las dificultades que han tenido que superar las técnicas de Clustering aplicadas a expresión genética es la adaptación a la estructura específica que tienen este tipo de matrices, ya que existe una gran diferencia de tamaño entre las dos dimensiones

Aurelio López Fernández, Escuela Politécnica Superior, Universidad Pablo de Olavide, E-mail: aurelio.lfdez@gmail.com

de la misma, es decir, más genes que condiciones [5].

La desventaja principal de las técnicas de Clustering es que las agrupaciones realizadas se basan en tan solo una dimensión. De tal manera, la agrupación de genes está basada en toda la dimensión de las muestras experimentales, mientras que la agrupación de muestras experimentales se basa en toda la dimensión de los genes. Se ha comprobado que en la naturaleza, un subgrupo de genes puede estar co-expresado y co-regulado bajo un conjunto de muestras experimentales pero su comportamiento podría variar bajo otro conjunto distinto [2].

Por ello, las técnicas de Biclustering se crearon con la finalidad de satisfacer este tipo de comportamiento.

1.2. Biclustering

El objetivo de las técnicas de Biclustering consiste en la identificación de subgrupos de genes y subgrupos de muestras experimentales que muestran patrones similares de comportamiento. Ello se consigue aplicando Clustering sobre dichos genes y muestras experimentales de manera simultánea, en lugar de realizarlo con una sola dimensión [2].

Al tratarse de subgrupos o submatrices, tenemos la posibilidad de que un gen o muestra experimental se encuentre en ninguno, uno o distintos biclusters. Por lo que nos proporciona una restricción mucho menor que los clusters, aumentando el número de posibles resultados y el solapamiento entre esas submatrices.

2. Validación biológica

La validación en Bioinformática puede ser agrupada en técnicas analíticas/matemáticas, que miden la calidad de los resultados en base a unas métricas que no se basan en ningún conocimiento previo, como por ejemplo en [6], y en técnicas que si se basan en ese conocimiento previo para determinar cómo de relevante es un resultado desde un punto de vista biológico, como por ejemplo en [7]. El conocimiento biológico previo es extraido de bases de datos biológicas disponibles en la web.

Gene Ontology, es una de las bases de datos más revelantes para la clasificación y asignación de funciones génicas y proteicas. Es una iniciativa centrada en unificar la representación de los genes y de sus productos de todas las especies. Gene Ontology está basado en términos, donde cada término GO dispone de un identificador único numérico (GO:xxxxxx), y un nombre asociado. Cada término GO es incluido dentro de una de las tres ontologías existentes: función molecular, componente celular o proceso biológico.

El análisis de enriquecimiento de genes es una de las técnicas de validación basadas en conocimiento biológico previo. Partiendo de una base de datos biológica, el objetivo de este análisis es la recopilación de aquellos términos biológicos que están relacionados con los genes del cluster o bicluster. Cada término biológico se pone en valor a partir de una medida estadística, el p-value, que nos indica la importancia de un término biológico con respecto a un conjunto de genes analizado, determinando si el resultado es positivo (valor próximo a 0), o se trata de un resultado más relacionado con el azar (valor más alejado a 0) [8].

Ontologizer es un software para llevar a cabo el análisis de enriquecimiento obteniendo, para cada grupo de genes analizado, los términos GO relacionados con dichos genes junto con el p-value asociado a cada término [9].

3. BIGO

El objetivo de BIGO es aprovechar la potencia de la herramienta Ontologizer y proporcionar información relevante a la validación de grupos de genes, clusters o biclusters, que dicha herramienta aporta.

BIGO procesa la validación generada por Ontologizer para obtener, por un lado, un ranking de términos biológicos, y por otro, un grafo que representa la relación entre los grupos de genes.

3.1. Ranking

El ranking se genera a partir de todos los términos biológicos encontrados en todos los grupos de genes analizados. El orden que se establece está basado en el número de veces que cada término biológico aparece entre los grupos de genes mencionados.

Nombre	Total grupos	Localizacion
positive regulation of phosphorylation	1	[13]
macromolecular complex	1	[5]
response to starvation	2	[25, 24]
regulation of gene silencing	3	[24, 4, 10]

Cuadro 1: Ejemplo de Ranking con distintos niveles.

El cuadro 1 representa un ejemplo de ranking donde cada línea corresponde a un término biológico y las columnas asociadas son las siguientes:

- Nombre: Nombre del término GO obtenido de Gen Ontology.
- Total grupos: Cantidad total de grupos de genes donde se encuentra el término en cuestión.
- Localización: Lista de los biclusters en los que aparece el término biológico.

En este ejemplo, la primera y segunda fila

corresponden a términos biológicos que aparecen en un único bicluster, en el 13 y el 5 respectivamente. Sin embargo, la última fila corresponde a un término biológico que aparece en tres biclusters, [24, 4, 10].

Los primeros términos del ranking nos permitirán centrar las conclusiones del estudio biológico sobre aquellos términos que realmente distinguen a un grupo de genes de otros. Por otro lado, los últimos términos del ranking facilitan la localización de las stop-words, es decir, funciones biológicas consideradas muy genéricas, ya que aparecen en un elevado número de genes, y que no deben ser tenidas en cuenta en el análisis final. En conclusión, este ranking ayuda a acotar de manera más precisa la validación generada por Ontologizer, permitiendo conclusiones más certeras.

3.2. Grafo

El grafo se obtiene a partir del ranking generado y representa la relación existente entre los grupos de genes en función del número de términos biológicos que comparten.

Cada nodo del grafo es un grupo de genes identificado por un número único, mientras que la arista unirá dos nodos si dichos grupos de genes comparten términos biológicos. El peso de la arista corresponde al número de términos biológicos compartidos entre sí.

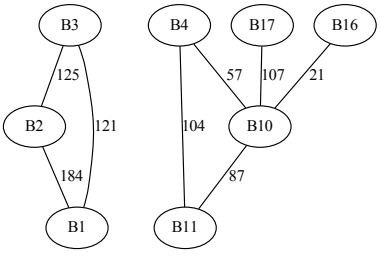

Figura 3: Representación gráfica de grupos de genes.

En el ejemplo de la figura 3 se observa que existen dos grandes grupos de genes bien diferenciados y no relacionados entre sí. Además, los biclusters incluidos en cada grupo están muy relacionados entre ellos debido a que comparten un gran número de términos biológicos entre sí.

El grafo tiene como objetivo la representación gráfica de las relaciones entre los distintos grupos de genes. Además, nos permite conocer si en nuestro resultado existen grupos de genes bien definidos e independientes entre sí. Además, aquellos grupos muy relacionados son también interesantes, en el caso en que no compartan una elevada proporción de sus genes.

CONCLUSIONES

En este artículo se ha expuesto una nueva herramienta para aumentar la información obtenida por el análisis de enriquecimiento obtenido por Ontologizer.

BIGO se basa en la obtención de un ranking a partir de los términos biológicos detectados de todos los grupos de genes, y posteriormente, un grafo que representa la relación entre esos grupos de genes.

Futuros trabajos permitirán añadir más información útil y transformar BIGO en una herramienta web accesible a cualquier investigador.

REFERENCIAS

[1] Norberto Diaz, "Tesis Doctoral: Similitud funcional de genes basada en conocimiento biológico".

[2] S.C. Madeira and A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," IEEE Transactions on computational Biology and Bioinformatics, vol. 1, no. 1, pp. 24 - 45, Jan/Mar 2004.

[3] P. Baldi and G.W. Hatfield "DNA Microarrays and Gene Expression. From Experiments to Data Analysis and Modelling," Cambridge University Press, 2002.

[4] G. Kerr, H.J. Ruskin, M. Crane and P. Doolan "Techniques for clustering gene expression data," Computers in Biology and Medicine, 38, pp. 289 - 293, Mar 2008.

[5] D. Jiang, C. Tang and A. Zhang "'Cluster Analysis for Gene Expression Data: A Survey,"' vol. 16, no. 11, pp 1370 - 1386, Nov 2004.

[6] C. van Rijsbergen. Information Retrieval. Second Edition, Butterworths, 1979.

[7] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics, 8:111+, March 2007.

[8] Rempher K.J. and Urquico K. "The P value: What it really means," American Nurse Today, 2(5), pp 13 - 15. 2007.

[9] Bauer S, Grossmann S, Vingron M and Robinson PN. "Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration." Bioinformatics (Oxford University Press), 24(14), pp. 1650 - 1651, 2008.

Aurelio López Fernández estudia 3º de Grado en Ingeniería Informática de Sistemas de Información en la Universidad Pablo de Olavide. Su interés investigador incluye el análisis inteligente de datos, la computación biomédica y biológica, el reconocimiento de patrones y las bases de datos. Desde 2013 es alumno interno en el Departamento de Lenguajes y Sistemas Informáticos. En 2011 obtuvo el Premio Extraordinario de Formación Profesional por la Junta de Andalucía.