CISTI 2007, Volume I, pp.469-480

Aproximación Bio-Bibliométrica a la Detección de Relaciones Biológicas entre Genes

Carmen Galvez[1], Félix Moya-Anegón[2]

cgalvez@ugr.es, felix@ugr.es

[1] Universidad de Granada, 18071, Granada, España.

[2] Universidad de Granada, 18071, Granada, España.

Resumen: La investigación bioinformática ha generado una gran cantidad de literatura biomédica almacenada en bases de datos tales como *MEDLINE*. La extracción de información de la literatura publicada se puede aplicar para detectar relaciones biológicas entre genes. La premisa del análisis Bio-Bibliométrico es la siguiente: *si dos símbolos de gen aparecen en el mismo documento es probable que estén relacionados* (por el principio de co-ocurrencia). Estos datos se pueden utilizar para calcular la '*distancia biobibliométrica*' entre pares de genes de un genoma completo. En este trabajo, realizamos un sencillo experimento basado en este planteamiento con el objetivo de extraer y visualizar información de la literatura biomédica relacionada con la enfermedad del linfoma. Las principales limitaciones de este método son la unificación de las diferentes variantes de nombres de gen, para que no se produzcan co-ocurrencias incorrectas, y la identificación del tipo de interacción genómica.

Palabras clave: Análisis Bio-Bibliométrico; Minería de Textos; Redes de Genes

1. Introducción

El Proyecto Genoma Humano (PGH) ha logrado elaborar un mapa que ubica a sus 30,000 – 40,000 genes (Venter et al., 2001). Con los datos de secuencias se pueden determinar la función de numerosos genes, permitiendo diferentes aplicaciones médicas y nuevos enfoques dentro de la biotecnología y la biología industrial. A raíz del PGH se ha generado una gran cantidad de literatura biomédica, que ha impulsado el diseño de potentes instrumentos para la obtención y el análisis de la información genética. A su vez, el vertiginoso incremento de las bases de datos, y la gran cantidad de conocimiento que reside en ellas, ha despertado el interés de la genómica, dedicada al estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades. La genómica se divide básicamente en tres grandes

ramas: (*i*) *genómica estructural* (orientada a la caracterización y localización de las secuencias que conforman el ADN de los genes); (*ii*) *genómica comparativa* (orientada a la comparación de los genomas animales con el genoma humano, para determinar sus diferencias y similitudes); y (*iii*) *genómica funcional* (orientada a recolección sistemática de información sobre la función de los genes). El objetivo de la genómica funcional sería llenar el hueco existente entre el conocimiento de las secuencias de un gen y su función para, de esta manera, desvelar el comportamiento de los sistemas biológicos.

Debido a que la mayor parte de la información sobre funciones e interacciones de genes se encuentra todavía en las bases de datos y en la literatura biomédica, las técnicas de Minería de datos (*data mining*) y Minería de Textos (*text mining*) han experimentado un auge como soporte para el descubrimiento del significado que poseen los datos almacenados. La Minería de Textos consiste básicamente en procesos automáticos para analizar textos en lenguaje natural con el propósito de descubrir información y conocimiento que es difícil de recuperar. Los procesos de Minería de Textos tienen una etapa necesaria de pre-procesamiento, al tratar datos no-estructurados, y usan diferentes técnicas tales como categorización, extracción de términos, *clustering* y visualización de los resultados. La aplicación de técnicas de Minería de Textos a la genómica funcional constituye un campo incipiente de investigación que comprendería tres grandes frentes (Tanabe, 2005): 1) *minería de relaciones*, o extracción de información, considerando dos o más entidades biomédicas; 2) *redes de genes basadas en la literatura*, o extracción de información a partir de la co-ocurrencia de nombres de gen; y 3) *knowledge discovery in database* (KDD), o extracción de conocimiento a partir de grandes conjuntos de datos.

Una aproximación de Minería de Textos, basada en la construcción de redes a partir de la literatura biomédica, denominada *Bio-Bibliometría* (Stapley & Benoit, 2000) surge como un procedimiento para recuperar y visualizar información biológica teniendo en cuenta la co-ocurrencia de términos genómicos. Para la detección de relaciones biológicas, el análisis Bio-Bibliométrico parte de la siguiente premisa: *si dos genes están relacionados biológicamente es probable que los dos nombres de gen aparezcan en el mismo documento*. Stapley & Benoit (2000) computan el número de co-ocurrencias de cada par de genes en abstracts de *MEDLINE* y realizan un proceso de normalización de datos utilizando una medida de similaridad para calcular lo que ellos denominan '*distancia biobibliométrica*', definida como el *Coeficiente de Dice* de dos genes *i* y *j*:

$$d_{ij} = \frac{|i| + |j|}{|i| \cap |j|}$$

Con los datos de co-ocurrencia se construye una matriz simétrica que podría contener medidas de *similaridad/distancia* de todos los pares de genes de un genoma completo. Los nodos del gráfico representarían genes, mientras los enlaces que unen los distintos nodos indican las relaciones existentes entre ellos. En una etapa más ambiciosa, los nodos pueden constituir vínculos hipertextuales conectados a bases de datos de secuencias y expresiones genéticas, mientras que los enlaces pueden estar conectados a aquellos documentos de *MEDLINE* que los han generado.

Las redes de co-ocurrencia de genes no sólo proporcionarían información para asignar funciones biológicas a secuencias y expresiones genéticas, sino que se podrían utilizar para sugerir, y probar, hipótesis, o descubrir nuevo conocimiento. En un trabajo publicado en *Nature*, Blasoklonny & Pardee (2002) afirmaban que la Biología Molecular se mueve de una era de recopilación de datos a otra dirigida por hipótesis a través de la conexión de diversos hechos, aparentemente separados. Posteriormente, las hipótesis deberán ser probadas en términos de ensayos experimentales. Varios sistemas de minería de la literatura biomédica se basan en la probabilidad de co-ocurrencia en el mismo documento de genes relacionados funcionalmente, como *MedMine*r (Tanabe et al., 1999) y *PubGene* (Jenssen et al., 2001), y en el significado estadístico de estos datos para descubrir información sobre genes (Chaussabel & Sher, 2002; Wren et al., 2004). La naturaleza de estas interacciones se podrían anotar usando bio-ontologías, o terminologías del MeSH® (*Medical Subject Headings*).

En este estudio vamos a realizar una sencillo experimento basada en la '*distancia biobibliométrica*' para generar una visualización gráfica que permita detectar relaciones biológicas entre clusters de genes. Para apoyar nuestro análisis hemos adaptado el método del análisis de redes, usando paquetes software externos existentes, como *Ucinet 6.1* y *NetDraw 2.28* (Borgatti, Everett, & Freeman, 2002). Nuestro objetivo será detectar funciones moleculares entre los genes asociados a la enfermad del linfoma que co-ocurren en un mismo documento. Esta información se podrá utilizar para formular hipótesis generales sobre las funciones moleculares de un grupo de genes, o para hacer conjeturas sobre las posibles relaciones biológicas, desconocidas, de un determinado gen a través de las conexiones semánticas con otros genes relacionados.

2. Material

Los nombres de gen han sido seleccionados de *ENTREZ GENE* [1], una base de datos producida por *The National Center for Biotechnology Information* (NCBI) y especializada en información sobre genes. El conjunto de datos se limita al orgnanismo '*Homo sapiens*' y a los genes relacionados con la enfermedad 'lymphoma'. Un total de 369 símbolos oficiales para denominar a los genes fueron

recuperados. Por ejemplo, para el gen **BCL11B**, la búsqueda en *ENTREZ GENE* ofrece la siguiente información:

- Official Symbol: **BCL11B** and Name: **B-cell CLL/lymphoma 11B** (zinc finger protein) [*Homo sapiens*]
- Other Aliases: **CTIP-2**, **CTIP2**, **RIT1**, **hRIT1-alpha**
- Other Designations: **B-cell CLL/lymphoma 11B**; **B-cell CLL/lymphoma 11B/T-cell receptor delta constant region fusion protein**; **B-cell lymphoma/leukaemia 11B**; **BCL11B/TRDC fusion**; **zinc finger protein hRit1 alpha**
- GeneID: **64919**

Debido a la gran cantidad de términos y nomenclaturas utilizadas para la identificación una misma entidad genómica, la normalización de las variantes del nombre de un gen constituye una etapa de pre-procesamiento esencial para calcular una red de co-ocurrencias de genes en la literatura científica. Jenssen et al. (2001) estiman que alrededor del 40% de los errores en las redes de genes están provocados por una identificación incorrecta de las variantes de nombres. En esta misma línea, Blaschke & Valencia (2001) advierten que aproximadamente dos terceras partes de los errores en el diseño de redes de genes se deben al uso inconsistente de las nomenclaturas genómicas. Para la unificación de las diferentes denominaciones y alias de un mismo nombre de gen aplicamos un procedimiento semiautomático, basado en técnicas de equiparación de patrones y *Transductores de Estado-Finito* (Galvez & Moya-Anegón, 2006a; 2006b).

La selección de la literatura biomédica sobre la que vamos a realizar el análisis bio-bibliométrico se ha limitado a abstracts de la base de datos *MEDLINE* a través de una búsqueda con el término simple 'Lymphoma', sin sub-encabezamientos, seleccionado del índice de términos MeSH® (*Medical Subject Heading*). Un total de 2,157 registros fueron recuperados sobre el período de 1960 hasta la fecha de la consulta. Un total de 61 genes, de los 365 seleccionados, fueron encontrados en los abstracts de *MEDLINE*.

3. Método

Con los datos obtenidos creamos un fichero de frecuencias de genes y un fichero de co-ocurrencias de pares de genes (Tabla 1).

Tabla 1 – Co-ocurrencias de nombres de gen en abstracts de *MEDLINE*

Gen A	Gen B	Ocurrencias A	Ocurrencias B	Co-ocurrencias (A, B)
CD40	**IL-4**	166	84	9
CD40	**Fc**	166	396	14
IL-4	**Fc**	84	396	27
IgE	**IL-4**	84	84	10

A partir de los datos de frecuencias se realiza el proceso de construcción de una matriz de co-ocurrencias simétrica de 96 x 96 genes con valores absolutos (Tabla 2).

Tabla 2 – Matriz simétrica de co-ocurrencias de nombres de genes

	BL	BLD	Bcl-2	CD25	CD3	CD4	EBV
BL	0	0	0	28	0	0	0
BLD	0	0	0	0	0	0	0
Bcl-2	0	0	0	0	0	0	0
CD25	0	0	0	36	0	0	0
CD3	0	0	0	0	0	13	0
CD4	0	0	0	0	0	113	0
EBV	0	0	0	27	0	0	0

Con los valores de co-ocurrencias en estado puro podemos generar redes de genes. En el caso de las matrices simétricas no es necesario utilizar ninguna medida de similaridad, ya que la similaridad de dos genes ya viene expresada en la matriz por el número de co-ocurrencias – cuanto más relacionados se encuentren dos genes más próximos se hallarán entre sí. Sin embargo, por tratarse de una aproximación basada en la '*distancia biobibliométrica*', decidimos aplicar la medida de similaridad *Coeficiente del Coseno, o Índice de Salton* (Hamers et al., 1989). El índice coseno de dos genes *i* y *j* se definiría como:

$$sim(i, j) = \frac{C_{ij}}{\sqrt{C_i C_j}}$$

donde

C_{ij} es el número de co-ocurrencias de *i* y *j*

C_i es el número de ocurrencias de *i*

C_j es el número de ocurrencias de *j*

Con los datos de similaridad obtenemos una matriz normalizada de co-ocurrencias de genes. La representación gráfica de clusters de genes se construye desde la matriz, utilizando el programa de visualización de redes *NetDraw 2.28*. Para identificar el tipo de relación, etiquetamos los nodos de la red con códigos de la base de datos *GENE ONTOLOGY* (GO) [2] y con descriptores seleccionados del MeSH [3]. Los términos GO proporcionan un esquema estructurado de clasificación jerárquica, *Direct Acyclic Graph* (DAG), para describir los atributos de los genes. Los tres principios de organización de los términos GO son: *Molecular Function* (MF), *Cellular Component* (CC) y *Biological Process* (BP). El gráfico GO consiste en alrededor de 18,000 términos representados como nodos dentro de DAG, y conectados por relaciones representadas por enlaces. Dentro de cada jerarquía se pueden dar dos clases de relaciones: '*is-a*' o '*part-of*'. A su vez, cada término de la ontología tiene un determinado número de atributos. Por ejemplo, el gen **CD40** tiene asignados, entre otros, los siguientes atributos: término GO ('*CD40 receptor binding*'), GO IDs ('*GO:0005174*'), o función molecular (*GO:0003674*') (Figura 1).

El MeSH se divide en 2 partes bien definidas: lista alfabética y lista jerárquica. Cada descriptor es un enlace directo con el registro completo de cada término del MeSH. La lista alfabética es una relación de encabezamientos principales, secundarios vocabulario de entrada y referencias cruzadas, organizados en forma alfabética. La lista jerárquica consiste en el conjunto de encabezamientos principales agrupados en 15 categorías naturales de acuerdo con un ordenamiento jerárquico. A cada encabezamiento principal y referencia cruzada le sigue una expresión alfanumérica cuya finalidad es orientar al usuario hacia la estructura jerárquica ('*Tree Structures*') correspondiente al encabezamiento. La estructura jerárquica del MeSH puede visualizarse igualmente en línea.

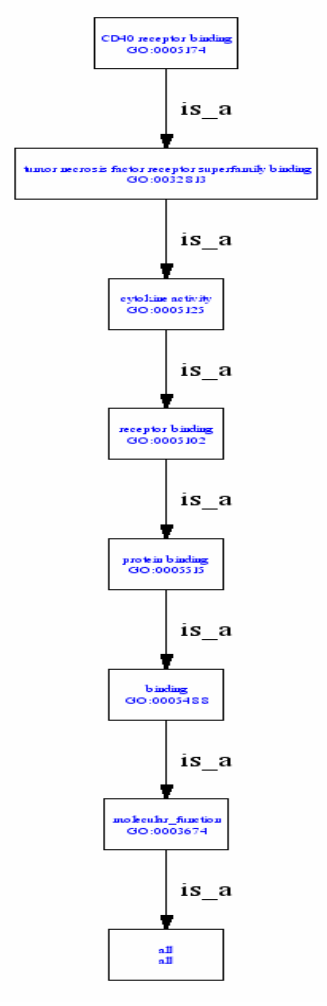

Figura 1 – Códigos GO asociados al gen **CD40**

4. Resultados

La visualización de los resultados del análisis bio-bibliométrico se refleja en 9 gráficos de redes de genes conectados. Los gráficos proporcionan una medio visual de representación esquemática de los términos genómicos dentro de un contexto. Los nodos del gráfico estarían determinados por los genes y el contexto estaría

determinado por el tipo de conexión, en este caso la información de co-ocurrencia de entidades genómicas. La Figura 2 muestra algunas de las redes generadas.

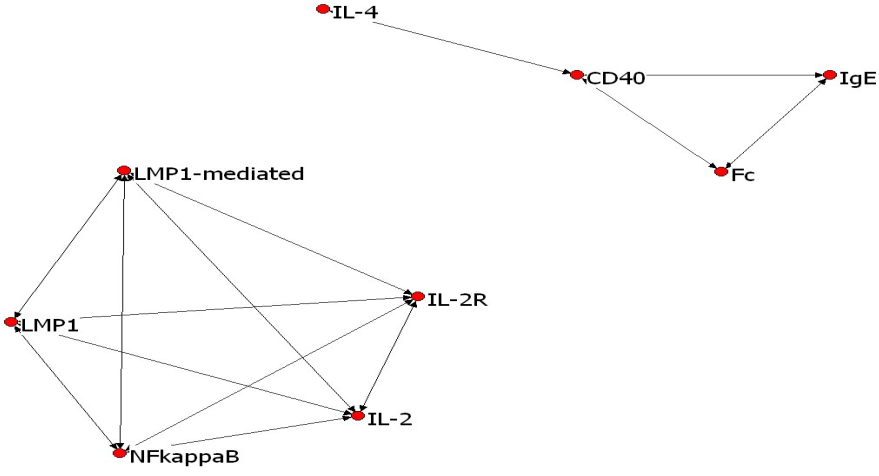

Figura 2 – Redes de genes

Seleccionamos una de las redes de genes y realizamos una anotación manual con los correspondientes códigos GO (Tabla 3), y con los términos extraídos de la clasificación jerárquica del MeSH (Tabla 4).

Tabla 3 – Términos y códigos GO asociados a los genes seleccionados

Términos GO	IL-4	CD40	IgE	Fc
protein binding GO:0005515	1	1	1	1
receptor binding GO:005102	-	1	-	-
immunoglobulin binding GO:0019865	-	-	1	1
binding GO:0005488	1	1	1	1
molecular function GO:0003674	1	1	1	1

Tabla 4 – Términos MeSH asociados a los genes seleccionados

Términos MeSH	IL-4	CD40	IgE	Fc
Biological Factors	1	1	-	-
Amino Acids, Peptides, and Proteins	-	1	-	-
Antigens	-	1	-	-
Peptides	1	1	1	-
Interleukins	1	-	-	-
Receptors, Cell Surface	-	1	-	1
Receptors, Tumor Necrosis Factor	-	1	-	-
Immunoglobulin Isotypes	-	-	1	-
Receptors, Immunologic	-	-	-	1

La anotación de la red seleccionada ha aportado la siguiente información:

- El cluster de genes **IL-4**, **CD40**, **IgE** y **Fc** comparte la *Función Molecular: GO:0003674.*

- El cluster de genes **IL-4**, **CD40** está relacionado con los términos MeSH pertenecientes a la estructura jerárquica: *Biological Factors [D23].*

- El cluster de genes **IL-4**, **CD40** y **IgE** está relacionado con los términos MeSH pertenecientes a la estructura jerárquica: *Amino Acids, Peptides, and Proteins [D12].*

Los resultados demuestran que si dos genes son co-citados en un registro de *MEDLINE* hay una relación biológica relevante entre ellos. El análisis de co-ocurrencia puede evidenciar muchas clases de interacciones pero es difícil caracterizar la naturaleza de las relaciones identificadas – que sería, por otra parte, lo deseable en aplicaciones tales como la construcción automática de redes de interacciones de genes. La anotación de las conexiones tanto por medio de la asignación manual de códigos GO, como de términos MeSH, es útil, pero limitada por la falta de términos para caracterizar todas las interacciones posibles, y por la reducida flexibilidad del propio esquema estructural de las bio-ontología, y de la clasificación jerárquica de los descriptores del MeSH.

5. Conclusiones

Los términos genómicos que co-ocurren de forma más frecuente en la literatura biomédica tienen más probabilidades de representar una relación biológica, según el significado estadístico de los términos propuesto por el análisis Bio-Bibliométrico. Estos datos se pueden visualizar gráficamente en agrupaciones de genes. Con el sencillo experimento realizado hemos comprobado que las redes de genes pueden revelar relaciones y funciones biológicas que se hallaban implícitas en la información existente, y pueden ayudar a descubrir nuevo conocimiento subyacente a partir de las relaciones compartidas ya existentes. Sin embargo, hemos constatado que los principales problemas de esta aproximación serían: (*i*) la identificación y unificación de las diferentes variantes de nombres de gen, para que no se produzcan co-ocurrencias incorrectas entre pares de genes; y (*ii*) la identificación y caracterización del tipo de relaciones entre genes, debido a que las interacciones son muchas veces vagas e imprecisas, o faltan códigos para caracterizar todas las relaciones posibles. En conclusión, la eficacia de este método dependerá de la superación de, al menos, dos obstáculos. *Primero*, la resolución del problema generado por la ambigüedad de los términos genómicos, a través de la unificación de todas las variantes, símbolos y sintagmas nominales con los que se denomina a un gen. *Segundo*, la clasificación de la naturaleza de la co-occurrencia de genes.

En futuros trabajos pensamos seguir profundizando en los factores frustrantes de esta interesante aproximación, y en la posible adaptación de este tipo de análisis a otras áreas de investigación, tales como el descubrimiento de relaciones potenciales entre *gen-enfermedad*. Por último, debido a que la mayor parte de los datos sobre interacciones de genes se encuentran en la literatura biomédica, podríamos añadir que sería deseable que este tipo de análisis se aplicara al texto completo de los documentos, y no sólo a los *abstracts* de *MEDLINE*. También sería deseable que las redes de genes construidas mediante esta metodología se generaran de forma dinámica, y no en esquemas *off-line*. De cualquier forma, la representación gráfica de clusters de genes sólo sería el punto de partida para que los expertos en la materia investiguen estas relaciones.

Notas

1. Disponible en: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>
2. Disponible en: <http://www.geneontology.org/>
3. Disponible en: < http://www.nlm.nih.gov/mesh/MBrowser.html>

Referencias

Blaschke, C. & Valencia, A. (2001). Can Bibliographic Pointers for Known Biological Data Be Found Automatically? Protein Interactions as a Case Study. *Comparative and Functional* Genomics, *2*, 196-206.

Blasoklonny, M. V. & Pardee, A. B. (2002). Conceptual Biology: Unearthing the Gems. *Nature*, 416:373.

Borgatti, S., Everett, M. & Freeman, L. (2002). *Ucinet 6.0 for Windows*. Harvard: Analytic Technologies.

Chaussabel, D. & Sher, A. (2002). Mining Microarray Expression Data by Literature Profiling. *Genome Biology*, *3*(10), *Research0055*.

Galvez, C. & Moya-Anegón, F. (2006a). Extracción y Normalización de Entidades Genómicas en Textos Biomédicos: Una Propuesta Basada en Transductores Gráficos. In *Proceedings of the 1st Iberian Conference on Information Systems and Technologies - CISTI 2006* (Esposende, Portugal, Escola Superior de Tecnologia), 697-709.

Galvez, C. & Moya-Anegón, F. (2006b). Identificación de Nombres de Genes en la Literatura Biomédica. In *Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006* (Mérida, Spain, Open Institute of Knowledge, INSTAC), 344-348.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity Measures in Scientometric Research: The Jaccard Index versus Salton's Cosine Formula. *Information Processing & Management*, *25*(3), 315-318.

Jenssen, T.-K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression. *Nature Genetics*, *28*(1), 21-28.

Stapley, B. J. & Benoit, G. (2000). Biobibliometrics: Information Retrieval and Visualization from Co-Occurrence of Gene Names in Medline Abstracts. In *Proceedings of Pacific Symposium on Biocomputing*, 529-540.

Tanabe, L. (2005). The Genomic Data Mine. In Chen, H., Fuller, S. S., Friedman, C. & Hersh, W. (Eds.), *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York: Springer.

Tanabe, L., Scherf, U., Smith, L., Lee, J., Hunter, L. & Weinstein, J. (1999). MedMiner: An Internet Tex-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling. *BioTechniques*, 27(6), 1210-1217.

Venter, J. C. et al. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304-1351.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V. & Hamer, H. R. (2004). Knowledge Discovery by Automated Identification and Ranking of Implicit Relationships. *Bioinformatics*, *20*(3), 389-398.