Aplicación de la bibliometría en bases de datos no bibliográficas. El caso del *mycobacterium tuberculosis*.

Guzmán, MV.*; Calero, R.*; Milanés, Y.**; Ramirez, JC*

**Instituto Finlay. Centro de Investigación – Producción de Vacunas y Sueros. Ave. 27 No.19805, La Lisa, La Habana. Cuba. A.P. 16017 Cod. 11600. email. mvguzman@finlay.edu.cu // Universidad de las Ciencias Informáticas, UCI.*

Resumen

Al unir los fundamentos teóricos de la minería de datos y textos, el descubrimiento de conocimiento en bases de datos (BD) y la bioinformática con los conceptos bibliométricos, se ha incursionado en el estudio de las bases de datos biológicas no bibliográficas. La Bibliometría, como disciplina que asume como su materia prima a la información, es una herramienta válida para estudiar el contenido de casi cualquier base de datos. Esto es posible porque se apoya para los análisis en fórmulas matemáticas y estadísticas (Algoritmos) que le permiten relacionar partes de un conjunto enorme de datos y llegar a un nuevo conocimiento. A ello se le suma que bases de datos como el GenBank, BD Structure y la BD Taxonomy (todas producidas por el National Center for Biotechnology Information, NCBI). Estas BD están estructuradas por campos específicos como palabras claves, autores de las proteínas, referencias a los artículos o los artículos en Medline que la citan, etc. En este estudio se aplicaron los indicadores bibliométricos en el análisis de las proteínas que componen el genoma del mycobacterium tuberculosis específicamente las vinculadas a las proteínas de la membrana externa, se utilizaron los indicadores de co-ocurrencia de palabras y los indicadores de frecuencia. Se vincularon los resultados de las bases de datos entrezProteins a la base de datos bibliográfica MedLine.

Palabras claves: Bibliometría, bioinformática, minería de datos y textos.

Introducción

El mundo experimenta cambios tecnológicos fundamentales que tienen un impacto sobre la forma en que las personas trabajan, una muestra de ello es la confianza que han ido alcanzando los resultados que nos ofrecen las máquinas y la idea de considerar a la información como el más valioso de los recursos. En este contexto están inmersos desde las empresas de bienes raíces hasta la actividades de investigación y desarrollo.

Sin embargo, asimilar la tecnología y la información convirtiéndola en algo útil para la organización necesita de métodos y técnicas que ofrezcan esa posibilidad. Es decir, para que exista un verdadero beneficio los datos deben convertirse en nuevos conocimiento. Es por ello, que unido al crecimiento exponencial de la información, han surgido una serie de técnicas y procesos que posibilitan el análisis y la extracción del conocimiento. Tal es el caso del Descubrimiento de Conocimiento (Knowledge Discovery), la Minería de Datos (Data Mining), el Análisis Inteligente de Datos (Inteligent Data Análisis), el Análisis Exploratorio de Datos (Exploratory Data Análisis) y la Bioinformática (Fayyad, Piatetsky-Shapiro and Smyth, 1996), (Norton, 1999), (Tukey, 1977), (Berthold and Hand, 2000)

Todas ellas necesarias y con componentes multidisciplinares que pueden ser usados por las ciencias de la información o ser enriquecedores de ésta. La Bioinformática, por ejemplo, contempla al análisis de la literatura publicada en formato bibliográfico o textual como su área de investigación. Esto se hace evidente al consultar toda una serie de artículos publicados en la última década como: *Extending the mutual information measure to rank inferred literature relationships* publicado en *BMC Bioinformatics* por Jonathan D Wren (2004), este autor utiliza la co-ocurrencia de palabras usando un algoritmo al que llamó calculo del mínimo de las asociaciones compartidas. En el estudio se utilizó la BD Medline y explica que se puede utilizar para crear redes de asociaciones para la evaluación.

Otro de los ejemplos es el del grupo Alemán del *European Molecular Biology Laboratory*, (Shah, et al., 2003) que han estado trabajando también sobre el estudio de las asociaciones de palabras pero en este caso para responder a la problemática (que también es de interés de la bibliometría) de si las palabras claves de los artículos son mejores para representarlos que hacer minería de textos en el artículo completo, la pregunta es si vale la pena el esfuerzo. Se le suman el artículo del *Department of Pharmacology, University of Tennessee* que realizan estudios de minería de datos para construir redes de interés biológico usando los resúmenes de PubMed. (Chen and Sharp, 2004).

Estos son solo tres ejemplos de una nueva aplicación de la bibliometría o de los problemas que le pueden preocupar a los analistas. Sin embargo, y retomando lo planteado en documentos anteriores (Guzmán y Carrillo, 2004a) (Guzmán y Carrillo, 2004b), se puede apreciar, como estas nuevas disciplinas y técnicas emergentes utilizan tecnología de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biomedicina.

Específicamente, utilizan procedimientos de análisis e indicadores que ya han sido esgrimidos por la bibliometría desde hace varias décadas. Es decir la bibliometría, como disciplina asume, como su materia prima, a la información. Se apoya para los análisis en fórmulas matemáticas y estadísticas (Algoritmos) que le permiten relacionar partes de un conjunto enorme de datos y llegar a un nuevo conocimiento. A la Bibliometría, tampoco, le son ajenos los términos asociados a la gestión de bases de datos, normalización y a las técnicas asociadas a la minería de datos y textos, reconocimientos de patrones, etc. (Guzmán y Carrillo, 2004a) (Guzmán y Carrillo, 2004b).

Es por ello que tratando de buscar nuevas aplicaciones a la bibliometría y a los conocimientos de los profesionales de la información y con el propósito de apoyar a las actividades de investigación y desarrollo de organizaciones bio-farmacéuticas es que se han trazado como objetivos de este trabajo

1. Identificar las proteínas del *mycobacterium tuberculosis* que ha sido menos exploradas por los científicos.
2. Las que han sido más exploradas, identificar a que aplicaciones o líneas de investigación están asociadas.
3. Proporcionar datos al proyecto de búsqueda de una vacuna efectiva contra la tuberculosis.

Se utilizarán las tecnologías de manejo de información y particularmente los recursos del Análisis Inteligente de Datos como lo son las Redes Neuronales Artificiales (RNA) y la visualización de datos haciendo converger la Minería de datos con los métodos de la Cienciometría, la

Bibliometría y la Informetría. (Sotolongo y Guzmán, 2001), (Sotolongo, Guzmán y Carrillo, 2002), (Sotolongo, et al., 2001).

Materiales y Métodos:

Se confeccionaron tres bases de datos: una que contiene las proteínas asociadas al *mycobacterium tuberculosis* (específicamente las de la membrana externa por ser potencialmente más propicias para el diseño de nuevas vacunas) y sus descripciones provenientes de la BD *Protein* de la NCBI. Otra con los artículos de los autores que han publicado alguna aplicación relacionada a cada una de las proteínas de interés (referencias contenidas en PubMed). Se elaboró una clasificación con las proteínas para poder aplicar técnicas de minería de textos en los registros de Medline (autores).

En este ejercicio se trabajó primero con los tres conjuntos de datos por separado. El primero consta de 46259 proteínas (búsqueda realizada en BD) cada una de ellas se corresponde a un registro de la BD. Para el objetivo planteado solo se extrajeron los siguientes datos de las proteínas (Cuadro 1).

```
LOCUS            AAO17530 141 aa   Linear   BCT 26-JAN-2006
DEFINITION       65 kDa heat shock protein [Mycobacterium tuberculosis subsp
                 tuberculosis].
ACCESSION        AAO17530
VERSION          AAO17530.1  GI:27502337
DBSOURCE         accession AF547886.1
KEYWORDS
SOURCE           Mycobacterium tuberculosis subsp. tuberculosis
ORGANISM         Mycobacterium tuberculosis subsp. tuberculosis
                 Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
                 Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium
                 tuberculosis complex.
REFERENCE        1  (residues 1 to 141)
AUTHORS          Devulder,G., de Montclos,M.P. and Flandrois,J.P.
TITLE            A multigene approach to phylogenetic analysis using the genus
                 Mycobacterium as a model
JOURNAL          Int. J. Syst. Evol. Microbiol. 55 (PT 1), 293-302 (2005)
PUBMED           15653890
REFERENCE        2  (residues 1 to 141)
AUTHORS          Devulder,G., Pichat,C. and Flandrois,J.P.
TITLE            Direct Submission
JOURNAL      Submitted (20-SEP-2002) Dynamique des populations bacteriennes,
             Universite Claude Bernard Lyon1 UMR CNRS 5558 & C.H.U. Lyon Sud,
             Faculte de Medecine de Lyon Sud, BP 12, Oullins 69622, France
COMMENT      Method: conceptual translation.
```

Cuadro 1. Selección de los campos extraídos de la BD proteins para el estudio.

Se utilizó la metodología del ViBlioSOM (Sotolongo y Guzmán, 2001), (Sotolongo, Guzmán y Carrillo, 2002), (Sotolongo, et al., 2001), para la cual se integraron algunas herramientas diseñadas específicamente para emigrar los datos no bibliográficos al gestor de BD bibliográficas ProCite. Se obtuvieron mapas topográficos que ofrecieron una visión del problema y contribuyeron a revelar nuevos conocimientos.

Resultados

En este estudio no se profundizará en los indicadores de actividad obtenidos, aunque aportaron información a los gestores de proyectos, los hallazgos más interesantes fueron encontrados en la aplicación de los indicadores relacionales. Estos fueron empleados para encontrar las

asociaciones entre las proteínas y las características bioquímicas que ya han sido identificadas y trabajadas a nivel internacional. En estos casos se co-relacionaron las variables y resultados obtenidos en el procesamiento de la BD de proteínas con variables identificadas en el procesamiento de la BD Medline. Estos fueron, fundamentalmente:

1. Autor de la proteína y autor de los artículos en Medline.
2. Nombre de las proteínas (de membrana externa, porque son potencialmente más propicias para la elaboración de vacunas, en total fueron 471 proteínas) con los campos de sustancias y palabras claves.

En el primer caso de estudio se pudo comprobar que existen investigadores que tienen una productividad muy alta en la identificación de proteínas de la tuberculosis (TB). En total, se registraron 8091 autores diferentes. En la Tabla 1 solo se muestran los que tienen una frecuencia mayor a 100 proteínas identificadas.

Tabla 1. Registro de proteínas por autores, BD Entrez Proteins.

Autor	Nivel de Actividad
Carpenter,L	203
Eisen, JA	203
Fleischmann,RD	203
White,O.	203
Garnier,T	188
Camus, JC	129
Parkhill, J	123
Cole,ST	121
Brosch, R	114
Churcher,C	113

Para identificar los autores que han registrados proteínas con sus artículos en Medline se aplicaron técnicas de Minería de textos al confeccionar un listado de taxones con los autores de las proteínas y se aplicó al campo autor de la BD MedLine que contiene todos los registros de Vacunas Tuberculosis (19770 investigaciones). Esto permitió comprobar que los niveles de actividad en los registros de proteínas no coincidían con los niveles de investigación de estos mismos autores. Es decir Carpenter reporta en MedLine 41, White 188, mientras que otros de menor actividad como Birren,B.L (82 proteínas) tiene más investigación general reportada (73 registros) en MedLine que Carpenter. Aunque habría que profundizar en esta afirmación con otros estudios al respecto, se ha podido observar que no existe una relación directa entre la cantidad de proteínas registradas por un autor y la cantidad de artículos científicos que genera.

La otra aplicación (2), estuvo enfocada al descubrimiento de conocimiento relevante usando los datos de las proteínas y los campos sustancias-palabras claves. Esto permitió predecir probables determinantes de estas proteínas o asociaciones no esperadas como proteínas relacionadas a otros microorganismos del *mycobacterium*. Es decir, proteínas que están presentes tanto en el *mycobacterium tuberculosis* como en el *mycobacterium leprae* (Figura 1). Esto es fundamental para comenzar las investigaciones de un candidato vacuna que proteja no solo contra la tuberculosis sino también contra otras enfermedades de la misma familia como la lepra.

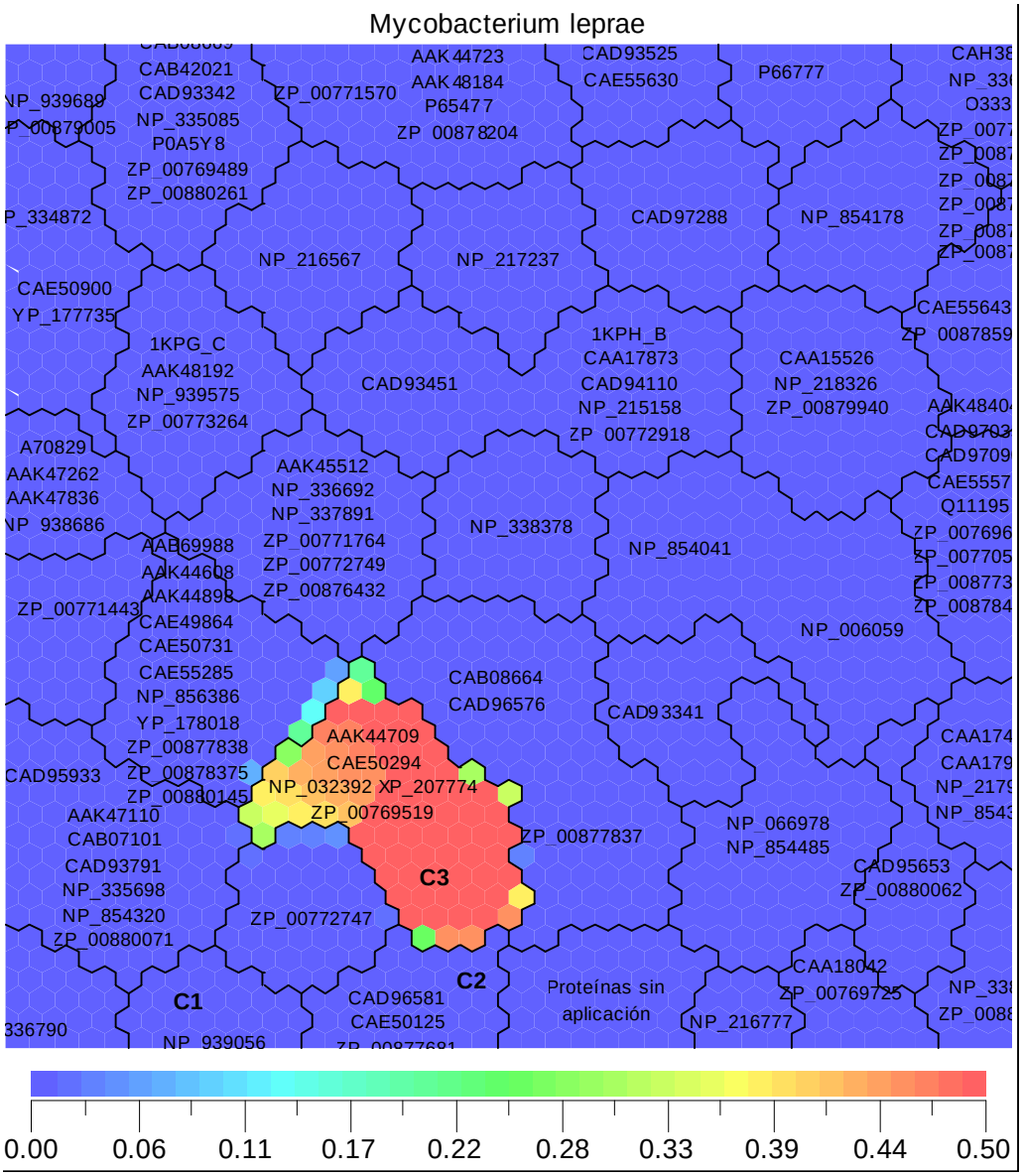

Figura 1. Mapa que representa todas las proteínas del mycobacterium tuberculosis (membrana externa). En el cluster 3 (C3) están concentradas aquellas proteínas que también fueron utilizados en estudios relacionados con el mycobacterium leprae (relacionado con descriptores de Medline).

Otro de los aspectos en los que se estaba interesado era en la búsqueda de elementos de la resistencia de la Tuberculosis ante los fármacos existentes (fundamentalmente antibióticos). En la Figura 2 se pueden apreciar las investigaciones asociadas a este parámetro.

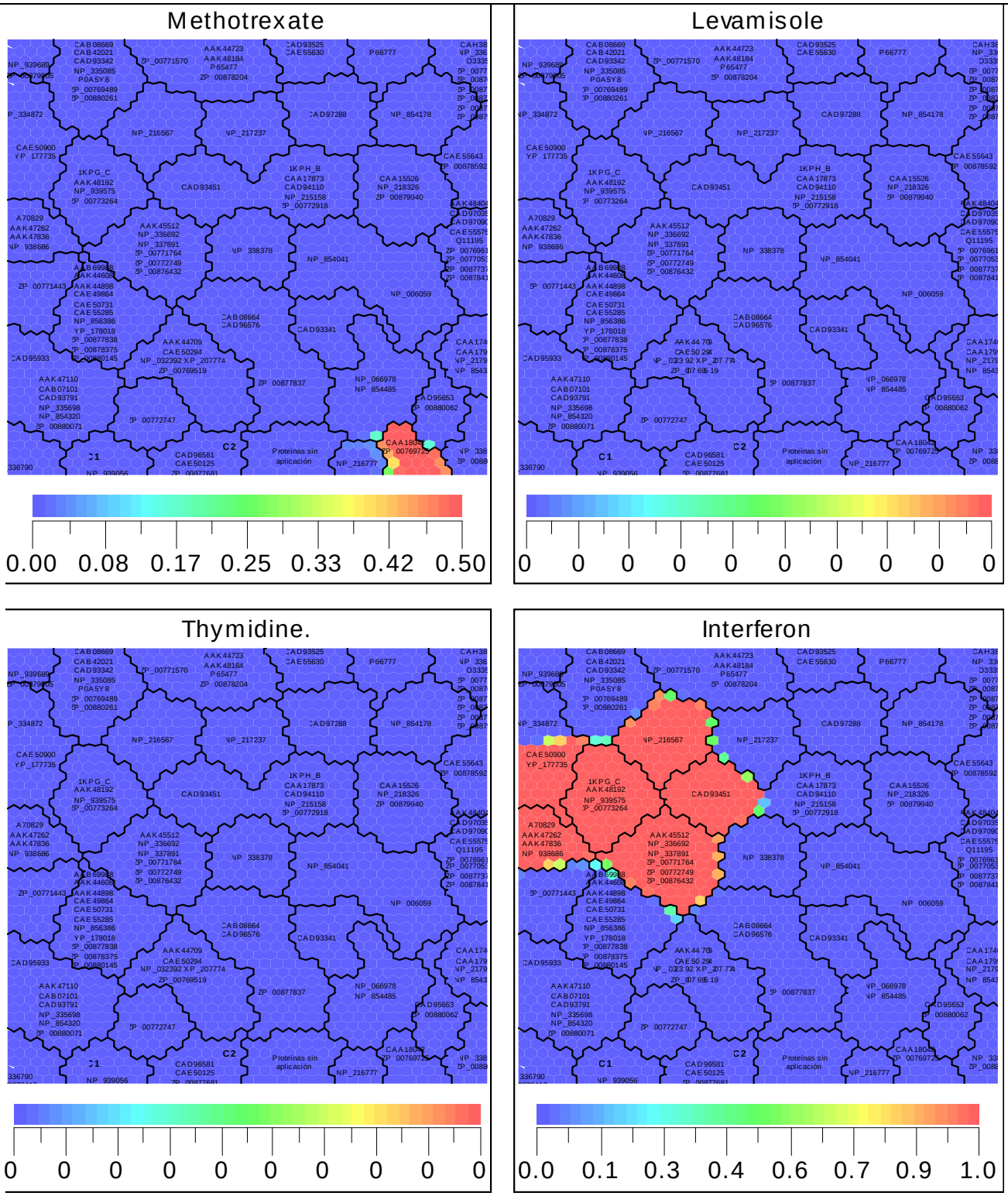

Figura 2. Estudios relacionados con la resistencia a los medicamentos Levamisole, Tymidine, Methotrexate) y al interferón.

En cada uno de los mapas observados, se puede apreciar que se han realizados varios estudios con el Interferón (hay 6 cluster que contienen 18 proteínas, cluster en rojo). Es importante señalar que no se esperaba la relación encontrada entre las dos proteínas (Cluster en rojo, mapa superior izquierdo) con el Methotrexate. Este es un medicamento utilizado fundamentalmente para el

tratamiento del cáncer (quimioterapia) y algunas experiencias los vinculan al uso del REMICADE (Medicamento para el tratamiento de artritis reumatoide). Sin embargo, no se encontraron estudios que lo vincularan directamente a la tuberculosis.

Por otra parte, se estudió la presencia de relaciones con la grasa (Oil). En la Figura 3 se señalan dos proteínas relacionados con este descriptor (Cluster 4).

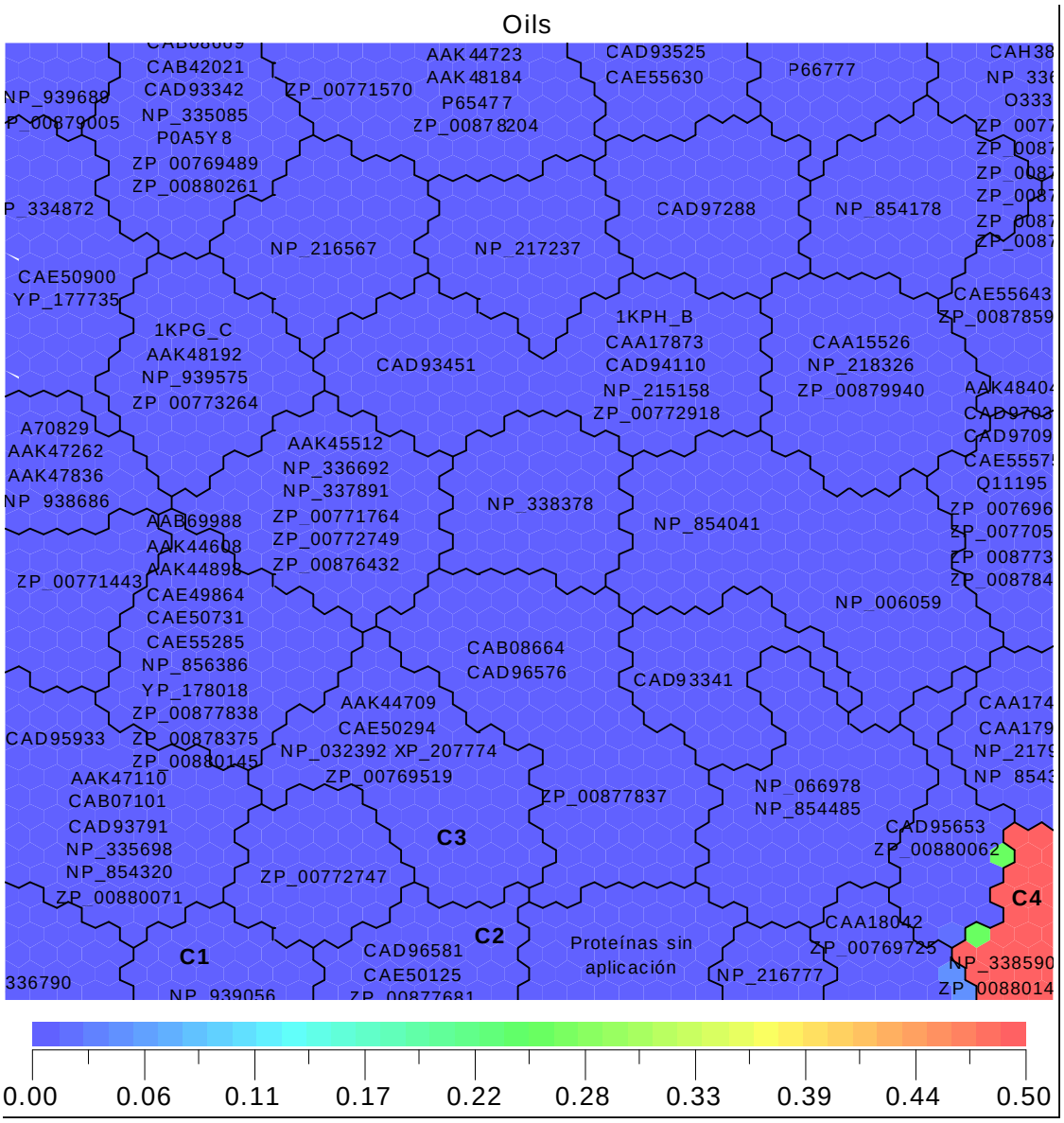

Figura 3. Las dos proteínas (Cluster 4, C4) asociadas a los estudios de producción y/o descomposición de grasas.

La relación de las grasas con la tuberculosis, es uno de los supuestos que los investigadores se han planteado para combatir la TB. Se estima que algunas proteínas identificadas dentro del genoma del *Mycobacterium tuberculosis* producen ciertas enzimas relacionadas con la producción y la descomposición de grasas. Algunos investigadores son de la opinión de que estas enzimas permiten que la bacteria de la TB sobreviva por largos períodos de tiempo. Esta

Figura, puede ser una demostración de que los presupuestos de los científicos ya están en fase de experimentación y que las proteínas referenciadas han sido las probadas bajo este paradigma.

Consideraciones finales

La búsqueda de nuevos medicamentos y alternativas para tratar la tuberculosis es un tema de salud urgente, la infección tuberculosa acaba cada año con casi tres millones de personas en el mundo. La OMS considera que ya estamos ante una emergencia global de tuberculosis que podría provocar 200 millones de casos nuevos y 70 millones de muertes antes del año 2020, si no se encuentra alguna solución eficaz para frenar la expansión de esta epidemia.

Por otra parte, los medicamentos actuales: la vacuna BCG que se está utilizando en muchos países tiene una eficacia limitada (alrededor del 50%). Además, la enfermedad se manifiesta cada día más resistente a los antibióticos y el SIDA ha hecho que los casos de TB aumenten aún entre población no infectada con VIH (OMS, 2005). En el día mundial contra la tuberculosis (2005), Olivier Brouant jefe de la misión de Médicos sin Fronteras en la India plateo "Los mejores fármacos existentes contra la tuberculosis fueron desarrollados entre las décadas de 1940 y de 1960. Los médicos no podemos estar satisfechos con los tratamientos disponibles" (OMS, 2005b).

Todo ello justifica, cualquier estudio que pueda aportar datos para el desarrollo de futuro fármacos. Las información sobre la asociación entre las proteínas y la grasa podría ser útil para actuar el desarrollo de un fármaco que actué contra las enzimas que producen la grasa y bloque el mecanismo de supervivencia que tiene este microorganismo. Así mismo se llama la atención sobre la Methotrexate, esta es una sustancia sobre la que no se han encontrado manifestaciones directas de aplicación en el tratamiento de la TB, por lo que se recomienda profundizar en el experimento que esta asociación. Y se llama la atención sobre un nuevo tratamiento o fármaco alternativo para tratar la enfermedad.

Por último, se considera que los indicadores bibliométricos y algunos conocimientos manejados por la bibliometría son validos para ser aplicados aún en bases de datos de información biológica que no tengan una estructura bibliográfica pero que sí reflejan el resultado de una investigación.

Bibliografía citada

- Berthold, M., Hand, DJ. (2000) Intelligent data analysis, Springer, 2000
- Chen, H., Sharp, BM (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5:147
- Fayyad, U.; Piatetsky-Shapiro, G., Smyth, P. (1996) “From data mining to knowledge discovery: an overview”, in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P and Uthurusamy, R. (Eds), *Advances in knowledge Discovery and Data Mining*, MIT Press, Cambridge, M.A.
- Guzmán, MV.; Carrillo, H. (2004). La Bibliometría como una herramienta de la Bioinformática. En: *Congreso Internacional INFO´2004*. Palacio de las Convenciones. Abril del 2004.
- Guzmán, MV.; Carrillo, H. (2004). Minería de datos con Redes Neuronales Artificiales: aplicación en vacunas tuberculosis. *X Convención Internacional y Feria Informática´2004. I Congreso de Bioinformática*. Palacio de las Convenciones, La Habana, mayo del 2004.

- Norton, MJ. (1999) “Knowledge Discovery in Database”, *Library Trends*, 48(1):9-2.
- OMS, Organización Mundial de la Salud (2005). La morbilidad y la mortalidad por tuberculosis relacionadas con el VIH alcanzan ya niveles alarmantes en África. Ginebra:OMS. Documento Online. Acceso: 12 enero del 2006. http://www.who.int/mediacentre/news/releases/2005/pr14/es/
- OMS, Organización Mundial de la Salud (2005b). World TB Day 2005. Ginebra: OMS. Documento Online. Acceso: 12 enero del 2006. http://www.stoptb.org/events/world_tb_day/2005/
- Shah, PK.; Perez-Iratxeta, C.; Bork, P.; Andrade, MA. (2003). Information extraction from full text scientific articles: Where are the keywords?. *BMC Bioinformatics*, 4:20
- Sotolongo, G., Guzmán, MV. (2001). Aplicaciones de las redes neuronales. El caso de la bibliometría”, *Ciencias de la Información*. 2001; 32(1):27-34.
- Sotolongo, G., Guzmán, MV., Carrillo, H. (2002) ViBlioSOM: visualización de información bibliométrica mediante el mapeo autoorganizado, *Revista Española de Documentación Científica*, 2002, 25(4):477-484.
- Sotolongo, G., Guzmán, MV., Saavedra, O.; Carrillo, H (2001). Mining Informetrics Data with Self-organizing Maps, in: M. Davis, C.S. Wilson, (Eds.), “*Proceedings of the 8 th International Society for for Scientometrics and Informetrics*”, ISBN:0-7334-18201. Sydney, Australia July 16-20. Sydney: BIRG; 2001: 665-673.
- Tukey, JW. (1977). “*Exploratory data analisys*”, Addison Wesley, 1977.
- Wren, Jonathan D (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, 5:145.

Bibliografía recomendada

- Chakravarti DN, Fiske MJ, Fletcher LD, Zagursky RJ. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine* 19:601-12 (2001).
- De Groot AS, Martin W. From immunome to vaccine: epitope mapping and vaccine design tools. *Novartis Foundation Symposium.* 254:57-76 (2003).
- Flower D. Vaccines in silico, The growth and power of immunoinformatics. *The Biochemist.* August:17-20 (2004).
- Grandi G. Antibacterial vaccine design using genomics and proteomics. T*rends in Biotechnology* 19(5):181-8 (2001).
- Zagursky RJ, Russell D. Bioinformatics: Use in Bacterial Vaccine Discovery. *BioTechniques* 31:636-59 (2001).

Datos de los autores:

Maria V. Guzmán. Master en Gestión de Información en la Organizaciones. Cátedra UNESCO. Universidad de La Habana y la Universidad de Murcia, España (Junio del 2000). Master en Información y Documentación. Universidad Carlos III de Madrid, España. (1996). Graduada de Licenciatura en Información Científico-Tecnológica. Universidad de La Habana, Cuba. (1992). Investigadora auxiliar y profesora de la universidad de La Habana. Miembro de la Cátedra UNESCO. Premio Nacional de la Academia de Ciencias, 2001. Consejo Asesor del Observatorio Cubano de Ciencia y Tecnología, Equipo Editorial VacciMonitor. Coordinadora del Seminario Internacional de Bibliometría ¨Gilberto Sotolongo Aguilar. Coordinadora de un

proyecto regional de la Naciones Unidas y coordinadora por la parte cubana de un proyecto en el marco de los proyectos bilaterales Cuba-México. Actualmente es J´ de Gestión de Información del Instituto Finlay.

Romel Calero Ramos. Lic. Ciencias de la Computación, Universidad de La Habana, Cuba (Julio 2001). Webmaster del Proyecto de la Naciones Unidas "Red Latinoamericana de Información Científico-Técnica en Vacunas". Ha participado en varios eventos científicos nacionales e internacionales en calidad de delegado y exponente. Ha obtenido varios resultados en los Forum de C-T a nivel de base. Tiene varias publicaciones nacionales e internacionales. Ha recibido varios cursos de postgrado en temas relacionados con las Bibliotecas Digitales, Gestión de Información y Programación distribuida.

Yusnelkis Milanés Guisado. Nacida el 23 de Julio de 1981, graduada de la especialidad de Bibliotecología y Ciencias de la Información en la Universidad de la Habana en el curso 2003/2004. Título de Oro y graduada más completa de su especialidad en Investigación y Docencia. Ha participado en eventos nacionales y en el I Simposio Internacional de Prospectiva y Vigilancia tecnológica. Ha recibido cursos de postgrado a diferentes niveles. Actualmente labora en la Universidad de las Ciencias Informáticas y colabora con el Proyecto Red Iberoaméricana de Información Científica Técnica en vacunas del Instituto Finlay. Profesora adjunta de la Universidad de la Habana y alumna de la maestría en BCI que se imparte en la Facultad de Comunicación.

Juan Carlos Ramírez Gómez. Nacido el 7 de Diciembre de 1975. Licenciado en Microbiología (1998). Universidad de La Habana, Cuba. Diplomado en Biología Computacional (2005). Universidad de La Habana, Cuba. Maestría en Bioquímica Mención Inmunología (2005). Universidad de La Habana, Cuba. Es Investigador Agregado y labora en el Departamento de Biología Molecular.