# MINERÍA DE TEXTOS: LA NUEVA GENERACIÓN DE ANÁLISIS DE LITERATURA CIENTÍFICA EN BIOLOGÍA MOLECULAR Y GENÓMICA

## *TEXT-MINING: THE NEW GENERATION OF SCIENTIFIC LITERATURE ANALYSIS IN MOLECULAR BIOLOGY AND GENOMICS*

Carmen Gálvez, PhD - cgalvez@ugr.es
Facultad de Comunicación y Documentación - Universidad de Granada

**Resumen**
Una vez descifrado la secuencia del genoma humano, el paradigma de investigación ha cambiado dando paso a la descripción de las funciones de los genes y a futuros avances en la lucha contra enfermedades. Este nuevo contexto ha despertado el interés de la Bioinformática, que combina métodos de las Ciencias de la Vida con las Ciencias de la Información haciendo posible el acceso a la gran cantidad de información biológica almacenada en las bases de datos, y de la Genómica, dedicada al estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades. En este contexto, la minería de textos surge como un instrumento emergente para el análisis de la literatura científica. Una tarea habitual de la minería de textos en Biología Molecular y Genómica es el reconocimiento de entidades biológicas, tales como *genes*, *proteínas* y *enfermedades*. El paso siguiente en el proceso de minería lo constituye la identificación entre entidades biológicas, tales como el tipo de interacción entre *gen-gen*, *gen-enfermedad*, *gen-proteína*, para interpretar funciones biológicas, o formular hipótesis de investigación. El objetivo de este trabajo es examinar el auge y las limitaciones la nueva generación de herramientas de análisis de la información en lenguaje natural, almacenada en bases de datos bibliográficas, como *PubMed* o *MEDLINE.*

**Palabras-clave:** Minería de Texto. Bases de Datos Textuales. Procesamiento del Lenguaje Natural (PLN).

## 1 INTRODUCCIÓN

El Proyecto Genoma Humano (PGH) está acentuando la necesidad de formar nuevos tipos de biólogos capaces de tender puentes entre diferentes disciplinas y la reorganización de los institutos de investigación, donde interaccionen especialistas en diversos ámbitos de las Ciencias de la Vida y las Ciencias de la Información. En los últimos años estamos asistiendo a una convergencia entre la Informática Médica (procesamiento de información clínica) y la Bioinformática (procesamiento de información genética) habida cuenta de la cada vez más estrecha relación entre enfermedades y genes. La Informática Médica y la Bioinformática son dos disciplinas científicas independientes que han llevado caminos separados. La Informática Médica

tiene una experiencia de varias décadas en el desarrollo de aplicación informática en el procesamiento de la información clínica, mientras que la Bioinformática, que lleva a cabo la aplicación de la informática en el procesamiento de la información genética es más joven, pero se ha desarrollado de manera extraordinaria en los últimos años, debido principalmente a los logros obtenidos en el PGH. Por ello, se plantea la aparición de nuevas áreas como la Genómica Clínica y la Medicina Molecular, que plantean retos importantes en la investigación biomédica, como son por ejemplo el uso de nuevas técnicas diagnósticas y terapéuticas y la creación de nuevos fármacos personalizados.

La convergencia diversas áreas de conocimiento ha dado lugar al diseño e implementación de sistemas informáticos que soporten la integración de bases de datos heterogéneas con la información médica y genómica. La información que se busca puede estar disponible en bases de datos privadas o públicas, tales como *MEDLINE*[1], *PubMed*[2] o *GenBank*[3]. El usuario podrá conectarse directamente con estas bases de datos a través del interfaz de un servidor de términos genético-médicos, y que permitirá la navegación y búsqueda en múltiples bases de datos. La información recuperada será almacenada para su integración posterior con información clínica o sanitaria. En este contexto, la literatura científica en Biología Molecular y Genómica constituye el mayor repositorio de conocimiento, y un elemento esencial en los procesos de gestión de ese conocimiento, debido a que es la mayor y más fiable fuente de información biológica. El resultado final de todos los experimentos biológicos se publica en formato de texto, y se recoge en bases de datos bibliográficas o textuales, como *MEDLINE*. Además, se ha incrementado la distribución de información médica en diferentes tipos de documentos, y no sólo en artículos científicos, como registros médicos electrónicos, documentos web, tales como *CliniWeb*[4] y *CISMeF*[5], o informes electrónicos, tales como *ProMed-mail*[6].

De cualquier forma, es necesario precisar que lo que entendemos por PFH consiste en principio en la obtención de información estructural, pero lo realmente importante empieza después, dando *sentido biológico*, tanto funcional como evolutivo a tal cúmulo de información, es decir, extraer auténtico extraer y producir auténtico conocimiento científico en la Biología Molecular. La gran cantidad de datos que han de ser procesados adecuadamente está provocando que se impulsen nuevos enfoques, nuevos experimentos e hipótesis de trabajo en las Ciencias Biológicas, y en las que los métodos propios de las Ciencias de la Información tienen mucho que aportar. Se habla por ello de una *era post-genómica*, en la que se irán integrando los conocimientos acumulados en diversos 'atlas' del ser humano y de otros seres vivos, en los que se podrán interrelacionar de modo funcionalmente significativo diversos niveles de comprensión de la materia viva: génico, genómico, regulación, biología celular, fisiología o evolución. El impacto real de todo ello no se puede prever, pero no cabe duda que el PGH sienta las bases de un salto cualitativo y cuantitativo en nuestra visión del mundo vivo. La información sobre el PGH y la investigación post-genómica tiene un enorme

[1] Disponível em: <http://medline.cos.com/>
[2] Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed>
[3] Disponível em: <http://www.ncbi.nlm.nih.gov/Genbank/>
[4] Disponível em: <http://www.streamx.com.au/cliniweb.htm>
[5] Disponível em: <http://www.chu-rouen.fr/cismef/>
[6] Disponível em: <http://www.promedmail.org/>

potencial de aplicación clínica. Se espera que el PGH sirva como fuente de conocimiento para la comprensión de los fenómenos biológicos, y así dar lugar a nuevos métodos de diagnóstico y tratamiento para enfermedades con base genética.

Partiendo de que la mayoría de lo que se conoce sobre genes y genomas está sin descubrir en la literatura biomédica, el análisis de las bases de datos bibliográficas o textuales podría ayudar a interpretar determinados fenómenos, o detectar relaciones entre diversas entidades biológicas (YANDELL & MAJOROS, 2002). La aplicación de las técnicas de minería de textos al dominio de la Biología Molecular constituye una de las más recientes y prometedoras áreas de investigación para el análisis de los datos biológicos. El objetivo de este trabajo es examinar el auge que está experimentando la minería de texto como instrumento para el descubrimiento del significado que poseen la gran cantidad de datos biológicos almacenados en las bases de datos bibliográficas o textuales.

## 2 OBJETIVOS DE LA MINERÍA DE TEXTO EN BIOLOGÍA MOLECULAR

Debido a que la mayor parte de la información sobre funciones e interacciones de genes se encuentra en la literatura y en las bases de datos biomédicas, es necesaria la aplicación de nuevos y potentes métodos de procesamiento y acceso a la información. La **minería de datos** (*data-mining*) y la **minería de texto** o **minería textual** (*text-mining*) surgen como tecnologías emergentes que sirven de soporte para el descubrimiento de conocimiento que poseen los datos almacenados. La minería de datos se define como el descubrimiento de conocimiento, a partir patrones observables de datos estructurados, en bases de datos relacionales, se le denomina comúnmente *Knowledge-Discovery in Databases* (KDD). La minería textual se orientada a la extracción de conocimiento a partir de datos no-estructurados en lenguaje natural almacenados en las bases de datos textuales, se identifica con el descubrimiento de conocimiento en los textos y se le denomina comúnmente *Knowledge-Discovery in Text* (*KDT*). Tanto la minería de datos como la minería de texto son técnicas de análisis de información.

En el caso de la información textual, mediante el proceso de análisis se le agrega valor a la información hasta convertirla en conocimiento, sólo las computadoras pueden manipular rápidamente la gran cantidad de datos. La minería de texto es una herramienta de análisis encargada del descubrimiento de conocimiento que no existía explícitamente en ningún texto de la colección, pero que surge de relacionar el contenido de varios de ellos (HEARST, 1999). Según Hearst (1999) la minería de texto adopta un *enfoque semiautomático*, estableciendo un equilibrio entre el análisis humano y automático: antes de la etapa de descubrimiento de conocimiento es necesario procesar de forma automática la información disponible en grandes colecciones documentales y transformarla en un formato que facilite su comprensión y análisis. El procesamiento de grandes volúmenes de texto libre no-estructurado para extraer conocimiento requiere la aplicación de una serie de técnicas de análisis ya utilizadas en la Recuperación de Información (RI), el Procesamiento del Lenguaje Natural (PLN) y la Extracción de Información (EI), tales como la identificación y extracción de patrones, análisis de *clustering*, clasificación, o visualización de datos.

Las bases de datos biológicas pueden ser clasificadas en dos tipos (Stapley & Benoit, 2000): 1) bancos de datos estructurados, con registros sobre secuencias y estructuras moleculares, tales como las bases de datos *SwissProt*[7] o *GenBank*; y 2) bases de datos textuales no-estructuradas, con registros en lenguaje natural, tales como *PubMed* y *MEDLINE*. La relación entre estas dos formas de información estructura y no-estructurada es clave. El conocimiento sobre el genoma no se limita al ADN o las secuencia genómicas, hay una gran cantidad de información sobre estos genes, almacenada en formatos no-estructurados dentro de millones de publicaciones. Los biólogos pueden extraer medidas entre dos secuencias de ADN de un banco de datos, como *GenBank*, pero esta relación puede ser identificada y descrita semánticamente con relaciones conceptuales extraídas de *PubMed* o *MEDLINE*.

Generalmente, el conocimiento biológico en las bases de datos textuales puede ser descubierto a través de tres procesos básicos (LEROY & CHEN, 2005): 1) *aproximación top*-down, en la cual los investigadores formulan hipótesis que conducen a experimentos específicos, o se crean ontologías para describir la terminología y el conocimiento en un dominio dado; 2) *aproximación bottom-up*, que persiguen descubrir patrones interesantes o asociaciones en los datos existentes, que a su vez se usan para formular nuevas hipótesis, las técnicas de *clustering* son las que se usan de forma más frecuente para este propósito; y 3) *métodos híbridos*, que implican la combinación de varias técnicas y fuentes de conocimiento, tales como métodos de recuperación de información y análisis de co-ocurrencia, para obtener conjuntos de documentos que puedan ayudar a los investigadores a articular nuevas hipótesis.

En relación con lo anterior, la minería de la literatura constituye un campo de investigación de la lingüística computacional que combina diversos procedimientos y técnicas de análisis de textos con el propósito de establecer relaciones entre entidades biológicas (como relaciones *gen-gen*, *gen-enfermedad*, *gen-proteína*, o *gen-drogas*) para interpretar funciones biológicas o formular hipótesis de investigación. La información textual, como la que se encuentra en *MEDLINE*, es una fuente infrautilizada de información biológica para los investigadores. Por esta razón, cada vez son más los sistemas dedicados a analizar resúmenes de *MEDLINE* para ofrecer servicios de información bio-relacionada.

El objetivo de la minería de textos en Biología Molecular y Genómica sería, por tanto, permitir a los investigadores identificar información de forma eficaz, descubrir relaciones no percibidas, ante el gran volumen de información disponible, y ayudar a descubrir conocimiento. Por otra parte, el interés creciente de esta rama de la lingüística computacional se refleja en el desarrollo de diversos proyectos de minería de la literatura, como *Suiseki* (BLASCHKE & VALENCIA, 2002), *MedMiner* (TANABE *et al*. 1999), *GeneCards* (SAFRAN *et al*., 2002), *XplorMed* (PEREZ-IRATXETA *et al*., 2001), *EDGAR* (RINDFLESCH *et al*., 2000), *BioBibliometrics* (STAPLEY & BENOIT, 2000), *GENIS* (FRIEDMAN *et al*., 2001), o *GIS* (CHIANG *et al*., 2004). También, son cada vez más frecuentes los congresos internacionales que reflejan el interés de la aplicación de las técnicas de minería a la Biomedicina y Biología Molecular, tales como ISMB (*Intelligent Systems for Molecular Biology*), ECCB (*European Conference on*

[7] Disponível em: <http://expasy.org/sprot/>

*Computational Biology*) o PSB (*Pacific Symposium on Biocomputing*).

## 3 FUNCIONES DE LA MINERÍA DE TEXTO EN BIOLOGÍA MOLECUALR Y GENÓMICA

El PGH fue el catalizador para el diseño de potentes instrumentos de obtención y análisis de la información genética. Una vez descifrado la secuencia del genoma humano, el paradigma de investigación ha cambiado dando paso a la descripción de las funciones de los genes y a futuros avances en la lucha contra enfermedades. Este nuevo contexto ha despertado el interés de la Bioinformática, que combina métodos de las ciencias biológicas y biomédicas con las ciencias de la información haciendo posible el acceso a la gran cantidad de información biomédica almacenada en las bases de datos, y de la genómica, dedicada al estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades. Las funciones esenciales de los proyectos que utilizan minería de textos en la investigación biomédica se focalizarían en el reconocimiento de entidades biológicas, categorización automática de los textos, identificación y extracción de la terminología tratada en los documentos, la extracción de relaciones y redes de conceptos, la visualización gráfica de estas relaciones, o la generación de hipótesis.

**Identificación y etiquetado de entidades biológicas**. Una de las áreas de investigación de la minería de la literatura biomédica es la identificación de los nombres y símbolos de las entidades biológicas. La identificación de nombres es un paso previo, que permitirá establecer posteriormente las posibles relaciones. Esta tarea aparentemente sencilla constituye un problema por varias razones. Primera, no existe un diccionario para la mayoría de las entidades biológicas, de esta forma los algoritmos de equiparación de texto, o *text-matching algorithms*, no pueden operar de forma eficaz. Segunda, un mismo nombre de entidad biológica puede referirse a entidades diferentes y, al contrario, una misma entidade biológica tiene varios nombres. A este problema se añade la dificultad que plantea el reconocimiento de entidades biológicas que tienen nombres compuestos por varias palabras. Por lo tanto, la identificación de forma automática de entidades biológicas en los textos en lenguaje natural es un área de interés del PLN, muchos trabajos han estado dedicados a esta tarea en el dominio biomédico (PROUX *et al*., 1998; FUKUDA *et al*., 1998; NOBATA *et al*., 1999; COLLIER *et al*., 2000; HUMPHREYS *et al*., 2000).

**Extracción y normalización de sinónimos, homónimos y abreviaturas.** Una vez identificadas las entidades biológicas, se tienen que resolver los problemas de sinonimia y abreviaturas, que podrían ser unificadas a continuación en alguna forma normalizada. La sinonimia surge cuando una misma entidad biológica tiene diferentes nombres, como el gen *Acf1*, con 13 alias (*CG1966*, *ACF*, *ATP*, *CAF*, *acf1*, *p170/p185*, *CHRAC*, *dACF*, *dCHRAC*, *ACF1*, *Acf-1*, *Acf*, *CHRAC-175*). Los problemas de homonimia y abreviaturas surgen cuando una misma entidad biológica puede referirse a múltiples entidades o puede ser la abreviatura de varias entidades, como la abreviatura de nombre de gen *PSA*, que se refiere a los nombres de genes '*Puromycin-Sensitive Aminopeptidase*', '*Prostate Specific Antigen*', '*PSoriatic Arthiritis*', '*Phosphoserine Aminotransferase*'. Varios trabajos de minería de textos se han dedicado a resolver estos problemas de ambigüedad (LIU *et al.*, 2002; YU *et al*, 2002; CHANG *et al*., 2002; YU & AGICHTEIN, 2003; TUASON *et al.*, 2004). Frente a estas investigaciones, el problema de la normalización de genes es un campo relativamente nuevo e inexplorado (CRIM *et al*., 2005; GALVEZ & MOYA-ANEGÓN, 2006b).

**Identificación de relaciones entre entidades biológicas a través de redes basadas en la literatura**. El objetivo de la extracción de relaciones es detectar ocurrencias, de un tipo específico de relación, entre pares de entidades biológicas. El tipo de relación puede ser por ejemplo asociación bioquímica, entre genes, proteínas o fármacos. Muchos trabajos han estado dedicados a la identificación de relaciones entre entidades biológicas. La co-ocurrencia de términos se usa para encontrar posibles relaciones entre genes (STAPLEY & BENOIT, 2000; JENSSEN *et al*., 2001; RAYCHAUDHURI *et al.*, 2002b) o proteínas (BLASCHKE & VALENCIA, 2001). Wren y Garner (2004) identifican genes relacionados analizando la cohesión y especificidad de la estructura gráfica a partir de las co-ocurrencias de genes en registros de *MEDLINE*.

**Generación de hipótesis y descubrimiento de conocimiento en las bases de datos textuales**. Blasoklonny y Pardee (2002) afirmaban en un artículo aparecido en **Nature** que la Biología Molecular se mueve de una era de recopilación de datos a otra dirigida por hipótesis, por la conexión de diferentes datos. Mientras la extracción de relaciones entre entidades biológicas se centra en la identificación de conexiones que se encuentran explícitamente en el texto, la generación de hipótesis se dirige a descubrir relaciones que no están presentes en el texto pero que se pueden inferir por la presencia de otras relaciones más explícitas. El objetivo de la generación de hipótesis sería revelar relaciones desconocidas dignas de ser investigadas posteriormente. La mayoría de los trabajos sobre generación de hipótesis parten de una idea original de Swanson (1986), en la que se proponía que las bases de datos de literatura científica permiten llevar a cabo descubrimientos por la conexión de conceptos, usando inferencia lógicas. La propuesta de Swanson, que se conoce como *modelo ABC* (WEEBER *et al*., 2003) es la siguiente: "*Si A influye en B, y B influye en C, entonces A puede influir en C*". En varios trabajos de Swanson (1987; 1988) se dan ejemplos del descubrimiento de nuevas hipótesis por la conexión manual de conceptos en la literatura científica. Posteriores investigaciones han tratado de automatizar este proceso (LINDSAY & GORDON &, 1999). Otros trabajos utilizan esta aproximación pero tomando términos MesSH (*Medical Subject Headings*), o conceptos del Metatesauro UMLS (*Unified Medical Language System*) (SRINIVASAN, 2004; SRINIVASAN & LIBBUS, 2004).

**Genómica funcional**. La genómica es la rama de la biología que se encarga del estudio de los genomas. Un genoma es el conjunto de información genética o ADN de un organismo. Básicamente, la genómica se divide en tres grandes ramas: a) *genómica estructural* (orientada a la caracterización y localización de las secuencias que conforman el ADN de los genes); b) *genómica comparativa* (orientada a la comparación de los genomas animales con el genoma humano, para determinar sus diferencias y similitudes); y c) *genómica funcional* (orientada a recolección sistemática de información sobre la función de los genes). La genómica funcional es un campo de la Biología Molecular que estudia cómo la información genómica define las funciones de los genes y proteínas en los organismos vivos. 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.

El análisis de las bases de datos textuales, y de la literatura biomédica, puede ayudar a detectar relaciones entre genes, o genes y enfermedades, interpretar determinados fenómenos, o establecer comparaciones entre genes similares de diferentes bases de datos. Todos estos procesos son cruciales para dar sentido a la inmensa cantidad de información genómica. Yandell y Majoros (2002) aseguraban que la mayoría de los que se conoce sobre genes y genomas está sin descubrir en la literatura biomédica. 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.

## 4 TÉCNICAS DE LA MINERÍA DE TEXTO EN BIOLOGÍA MOLECULAR Y GENÓMICA

Las técnicas de minería están dirigidas a procesar suficientes datos hasta descubrir patrones de relaciones útiles en un conjunto de datos, o hasta que confirmen o refuten una hipótesis. Es decir, estarían orientadas a descubrir el significado '*oculto*' que poseen los datos almacenados, hasta convertirlos en conocimiento para interpretar un fenómeno, o para la toma de decisiones. Al contrario que en los métodos tradiciones, basados en pruebas estadísticas, en los que se formula una hipótesis y se diseña posteriormente un experimento para captar los datos que prueben la hipótesis planteada, en los métodos de minería se procesan los datos con la finalidad de que de ellos surjan hipótesis, que posteriormente deberán ser probadas con los métodos científicos convencionales. Con este enfoque, las técnicas de la minería textual se estructuran básicamente en tres etapas:

- **Etapa de pre-procesamiento**, en la que los textos se transforman en algún tipo de representación estructurada que facilite su análisis.
- **Etapa de representación**, que dependerá de la técnica de pre-procesamiento utilizada y determinará a su vez el algoritmo de descubrimiento a utilizar.

- **Etapa de descubrimiento**, en la que a partir de una representación estructurada de la información, se aplican usa serie de algoritmos capaces de descubrir regularidades en los textos.

Todas las etapas están muy interrelacionadas, así pues, la primera etapa condicionaría el descubrimiento de los patrones que la minería de texto puede realizar. Las técnicas más usadas en minería textual son los vectores de temas, que muestran el nivel temático del texto, la secuencia de palabras que permite descubrir patrones en el texto y las tablas de datos, que permite descubrir interrelaciones entre entidades. En el ámbito biomédico, los proyectos de minería de texto adoptan un conjunto de técnicas agrupadas que incluyen en esencia: 1) pre-procesamiento de los documentos, en el que los textos se analizan y se elimina la información textual irrelevante, tales como técnicas de *stemming* y *lematization*; 2) etiquetado, identificación y extracción de las entidades biológicas, utilizando técnicas de equiparación de patrones, o *pattern-matching*; y 3) identificación de relaciones entre las entidades biológicas a través de análisis de co-ocurrencia, técnicas de *clustering*, clasificación automática y visualización gráfica.

**Pre-procesamiento de los documentos**. Las técnicas de pre-procesamiento de textos implican la eliminación de información textual que es no relevante para resolver la finalidad el proyecto de minería. Esta fase representa alrededor del 80% del esfuerzo global de las aplicaciones de minería (GLENISSON *et al*., 2005). El pre-procesamiento incluye la eliminación de palabras vacías y la unificación de los términos restantes mediante técnicas de *stemming* (PORTER, 1980). 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. Se estima que alrededor del 40% de los errores los proyectos de minería biomédica basados en redes de genes están provocados por una identificación incorrecta de las variantes de nombres (JENSSEN *et al*., 2001).

**Etiquetado, identificación y extracción de entidades biológicas**. Uno de los mayores obstáculos de la minería biomédica es la identificación de las entidades biológicas, especialmente las denominaciones de los genes. Hay múltiples designaciones para los mismos genes, y genes sin relación funcional entre sí llevan el mismo nombre. Los intentos por imponer denominaciones comunes en diferentes especies están encontrando una gran resistencia. Hay métodos que proponen dar a los genes números de identidad únicos, pero no pueden prosperar si las revistas científicas no obligan a los autores a adoptar este sistema. Las principales revistas científicas como *Nature*, *Nature Genetics* y *Science*, exigen a los autores que indiquen el número de acceso al banco genético *GenBank* en los artículos que describen un gen por primera vez, pero parece improbable que se imponga la utilización de ese número de identidad (PEARSON, 2001). Numerosos trabajos se han dedicado a la identificación de los nombres de los genes usando métodos del PLN, tales como métodos basados en reglas, uso de diccionarios o equiparación de patrones (HATZIVASSILOGLOU *et al*., 2001; LIU *et al*., 2001; TUASON *et al.*, 2004; SCHIJVENAARS *et al*., 2005; GALVEZ & MOYA-ANEGÓN, 2006a; 2006b).

**Análisis de *clustering*, categorización automática y visualización gráfica**. Los

algoritmos de minería se dividen generalmente en métodos no-supervisados, tales como algoritmos de *clustering* y técnicas de visualización, y métodos supervisados, tales como clasificación de documentos en una serie de categorías preestablecidas, o en ontologías creadas previamente. Los algoritmos de *clustering* agrupan las muestras de entrada en una serie de grupos, atendiendo a diferentes criterios, uno de los más habituales lo constituyen las relaciones de co-ocurrencia. En el caso de un banco de datos, los biólogos pueden establecer relaciones binarias numéricas entre entidades por alineamiento, o medidas de co-ocurrencia numérica, entre secuencias de ADN. En el caso de un corpus textual, los biólogos pueden establecer relaciones binarias semánticas entre entidades por medio de la co-ocurrencia de términos, como propone la *Bio-Bibliometría* (STAPLEY & BENOIT, 2000). Aunque la forma más simple de detectar relaciones entre entidades biológicas es calcular la co-occurencia de términos o símbolos, las interacciones entre se pueden visualizar en mapas o redes biológicas (NG & WONG, 1999; BLASCHKE & VALENCIA, 2002; GALVEZ & MOYA-ANEGÓN, 2007). Por otra parte, la técnica de categorización automática más utilizada en la minería textual biomédica consiste en clasificar textos biomédicos asociando entidades biológicas con términos seleccionados de ontologías, como los códigos *Gene Ontology* (GO)[8] (RAYCHAUDHURI *et al*., 2002a).

## 5 EVALUACIÓN DE LAS TÉCNICAS DE MINERÍA TEXTUAL EN EL DOMINIO BIOLÓGICO

A pesar de la importancia de los sistemas de minería textual biomédica para ayudar a los investigadores a extraer conocimiento de la literatura, y facilitar nuevos descubrimientos de una forma eficaz, estos sistemas no se utilizan de forma masiva. Una limitación para su desarrollo es la falta de métodos de evaluación sistemáticos, comparables a las métricas utilizadas en los sistemas de RI (HERSH, 2005). Uno de los problemas reside en que la mayoría de los trabajos realizados se focalizan en la evaluación de aspectos parciales, o sólo se dirigen a la evaluación de componentes muy específicos de tales sistemas, tales como la identificación de entidades biológicas, clasificación de documentos, o detección de relaciones entre entidades. Además de las dificultades ocasionadas por estas microevaluaciones, otro obstáculo lo constituye el hecho de que lo investigadores utilizan diferentes colecciones de prueba para evaluar sus sistemas, dando lugar también a la obtención de diferentes resultados según los corpus textuales en los que se aplican los experimentos.

Una iniciativa relacionada con la evaluación de información *ad hoc* adaptada a artículos biológicos se produjo en *Text Retrieval Conference* (TREC), bajo la organización del *US Nacional Institute for Standards and Technology* (NIST). En esta conferencia internacional se presentaron diferentes trabajos sobre la evaluación de sistemas de recuperación de información en el dominio genómico. Anualmente, desde 2003 se celebra *TREC Genomics Track*[9] con una gran variedad de resultados sobre diferentes corpus y en los que han predominado los trabajos sobre la identificación de entidades biológicas, uno de los aspectos más estudiados de los proyectos de minería biomédica. A su vez, debido a la falta de colecciones normalizadas para la evaluación de los

[8] Disponível em: <http://www.geneontology.org/>
[9] Disponível em: <http://ir.ohsu.edu/genomics/>

sistemas de minería, se ha creado el corpus GENIA (KIM *et al*., 2003), consistente en 2000 resúmenes de la base de datos MEDLINE, con más de 400000 palabras y alrededor de 100000 anotaciones, que han sido codificadas de forma manual para los términos biológicos.

La colección GENIA se ha usado también por muchos investigadores para la identificación de entidades biológicas en el *Internacional Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (JNLPBA) (KIM *et al*., 2004). Otras colecciones de pruebas desarrolladas por investigadores de MITRE han servido para testear diferentes sistemas de minería biomédica: *Knowledge Discovery from Database (KDD) Challenge Cup* (YEH *et al*., 2003) y *Critical Assessment of Information Extraction in Biology* (*BioCreAtIvE*) (HIRSCHMAN *et al.,* 2005). El propósito de KDD fue analizar cómo las técnicas de la minería de textos pueden ayudar a los encargados del mantenimiento (o *curators*) de las bases de datos biológicas. Por su parte, los resultados de *BioCreAtIvE* se focalizan en dos tareas: *Task 1A y Task 1B,* dirigidas a la extracción de nombres de genes y proteínas de los textos, y su equiparación a un identificador de gen normalizado; y *Task 2.1* y *Task 2.2*, dirigidas a la anotación funcional de entidades biológicas usando términos GO (*Gene Ontology*), y la extracción posterior de aquellos fragmentos de los textos científicos que los contienen.

De cualquier forma, la mayoría de los intentos de evaluación de los sistemas de minería textual biomédica se han realizado generalmente en pequeñas colecciones desarrolladas por grupos de investigadores de forma individual. Es necesario, por tanto, mejorar los recursos y los parámetros de valoración de estos sistemas, tanto en lo que se refiere a la normalización de las colecciones de prueba, como en unificación de las medidas que se utilicen para evaluar tales proyectos (HIRSCHMAN *et al.,* 2002). La superación de estas barreras no sólo ayudarían a determinar qué procedimientos son los más adecuados en el campo biomédico, sino que proporcionarían un *insight* sobre cómo mejorar tales sistemas.

## 6 EL FUTURO DE LA MINERÍA TEXUAL EN LA ERA POST-GENÓMICA

La minería de texto es una poderosa herramienta de análisis para la extracción de conocimiento a partir de datos biológicos no-estructurados. Los sistemas de minería textual biomédica se enfrentan a grandes retos, entre ellos se encuentra la necesidad de procedimientos que permitan la detección correcta de entidades biológicas, debido a la complejidad y falta de unificación de las nomenclaturas biomédicas. Además, es necesario establecer una métrica de evaluación común y normalizada, como los que existen para la evaluación de los sistemas de RI, que se utilice a su vez sobre a las mismas colecciones de documentos, de forma que se pueda comparar la eficacia de tales sistemas para realizar determinadas tareas. No obstante, y a pesar de estas limitaciones, nos encontramos ante un prometedor instrumento de análisis de información en el que confluyen, debido la complejidad propia del dominio de conocimiento, diversos campos de la biomedicina, la RI y el PLN. El futuro de esta tecnología se encontraría, por tanto, en aproximaciones multidisciplinares, en la que investigadores de diversos ámbitos puedan realizar un esfuerzo coordinado para alcanzar el potencial científico completo que plantean los proyectos de minería textual en las diversas áreas de las Ciencias de la Vida junto a las Ciencias de la Información

**REFERENCIAS**


BLASCHKE, C.; VALENCIA, A. Can bibliographic pointers for known biological data be found automatically? protein interactions as a case study. **Comparative and Functional Genomics**, v. 2, p. 196-206, 2001.

BLASCHKE, C.; VALENCIA, A. The frame-based module of the SUISEKI information extraction system. **IEEE Intelligent Systems**, v. 17, n. 2, p. 14-20, 2002.

BLASOKLONNY, M. V.; PARDEE, A. B. Conceptual biology: unearthing the gems. **Nature**, v. 416, p. 373.

CHANG, J. T.; SCHÜTZE, H.; ALTMAN, R. B. Creating an online dictionary of abbreviations from MEDLINE. **Journal of the American Medical Informatics Association**, v. 9, n.6, p. 612-20, 2002.

CHIANG, J. H.; YU, H. C.; HSU, H. J. GIS: a biomedical text-mining system for gene information discovery. **Bioinformatics**, v. 20, n. 1, p. 120-121, 2004.

COLLIER, N.; NOBATA C.; TSUJII, J. Extracting the names of genes and gene products with a Hidden Markov Model. **Proceedings COLING 2000**, p. 201-207, 2000.

CRIM, J.; MCDONALD, R.; PEREIRA, F. Automatically annotating documents with normalized gene lists. **BMC Bioinformatics**, v. 6, n. 1, p. 13-19, 2005.

FRIEDMAN, C.; KRA, P.; Yu, H.; KRAUTHAMMER, M.; RZHETSKY, A. GENIS: a natural-language processing system for the extraction of molecular pathways from journal articles. **Bioinformatics**, v. 17, n. 1, p. 74-82, 2001.

FUKUDA, K.; TSUNODA, T.; TAMURA, A.; TAKAGI, T. Toward information extraction: identifying protein names from biological papers. **Proceedings of the Pacific Symposium on Biocomputing**, p. 705-716, 1998.

GALVEZ, C.; MOYA-ANEGÓN, F. Aproximación *Bio-Bibliométrica* a la detección de relaciones biológicas entre genes. **II Conferęncia Ibérica de Sistemas e Tecnologias de Informação - CISTI 2007**, p. 469-480, 2007.

GALVEZ, C.; MOYA-ANEGÓN, F. Extracción y normalización de entidades genómicas en textos biomédicos: una propuesta basada en transductores gráficos. **I Conferęncia Ibérica de Sistemas e Tecnologias de Informação - CISTI 2006**, p. 697-709, 2006b.

GALVEZ, C.; MOYA-ANEGÓN, F. Identificación de nombres de genes en la literatura biomédica. **Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies - InSciT2006**, p. 344-348, 2006a.

GLENISSON, P.; GLÄNZEL, W; PERSSON, O. Combining full-text analysis and bibliometric indicators. a pilot study. **Scientometrics**, v. 63, n. 1, p. 163-80, 2005.

HATZIVASSILOGLOU, V.; Duboue, P. A.; RZHETSKY, A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. **Bioinformatics**, v. 17, p. 97-106, 2001.

HEARST, M. Untangling text data mining. **Proceedings of ACL'99: the 37th Annual Meeting of the Association For Computational Linguistic ACL**, p. 3-10, 1999.

HERSH, W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. **Briefings in Bioinformatics**, v. 6, n. 4, p. 344-356, 2005.

HIRSCHMAN, L.; PARK, C.; TSUJII, J.; WONG, L.; WU, C. H. Accomplishments and challenges in literature data mining for biology. **Bioinformatics**, v.18, n. 12, p. 1553-1561, 2002.

HIRSCHMAN, L.; YEH, A.; BLASCHKE, C.; VALENCIA, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. **BMC**

**Bioinformatics**, v. 6 (Suppl. 1), 2005.
HUMPHREYS, K.; DEMETRIOU, G.; GAIZAUSKAS, R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. **Proceedings of the Pacific Symposium on Biocomputing (PSB-2000)**, p. 505-516, 2000.
JENSSEN, T.-K.; LAEGREID, A.; KOMOROWSKI, J.; HOVIG, E. A literature network of human genes for high-throughput analysis of gene expression. **Nature Genetics**, v. 28, n. 1, p. 21-28, 2001.
KIM, J. D. ; T. OHTA; Y. TATEISI ; J. TSUJII. GENIA corpus - semantically annotated corpus for bio-textmining. **Bioinformatics**, v. 19, p. 180-182, 2003.
KIM, J. D.; OHTA, T.; TSURUOKA, Y.; TATEISI, Y.; COLLIER, N. Introduction to the biol-entity recognition task at JNLPBA. **Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)**, p. 70-76, 2004.
LEROY, G.; CHEN, H. Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. **Journal of the American Society for Information Science and Technology**, v. 56, n. 5, p. 457-468, 2005.
LINDSAY, R. K.; GORDON, M. D. Literature-based discovery by lexical statistics. **Journal of the American Society for Information Science and Technology**, v. 50, n. 7, p. 574-587, 1999.
LIU, H.; JOHNSON, S. B.; FRIEDMAN, C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. **Journal of the American Medical Informatics Association Online**, v. 9, p. 621-636, 2002.
LIU, H.; LUSSIER, Y. A.; FRIEDMAN, C. Disambiguating ambiguos biomedical terms in biomedical narrative text: an unsupervised method. **Journal of Biomedical Informatics**, v. 34, p. 249-261, 2001.
NG, S.; WONG, M. Toward routine automatic pathway discovery from on-line scientific text abstracts. **Proceedings of Genome Informatics**, p. 104-112, 1999.
NOBATA, C.; COLLIER, N.; TSUJII, J. Automatic term identification and classification in biology texts. **Proceedings of the 5th Natural Language Processing Pacific Rim Symposium**, p. 369-374, 1999.
PEARSON, H. Biology's name game. **Nature**, v. 411, p. 631-632, 2001.
PEREZ-IRATXETA, C.; BORK, P.; ANDRADE, M. A. XplorMed: a tool for exploring MEDLINE abstracts. **Trends in Biochemical Sciences**, v. 26, n. 9, p. 573-575, 2001.
PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, p. 130-137, 1980.
PROUX, D.; RECHENMANN, F.; JULLIARD, L. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. **Proceedings of Genome Informatics**, p. 72-80, 1998.
RAYCHAUDHURI, S.; CHANG, J. T.; SUTPHIN, P. D.; ALTMAN, R. B. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. **Genome Research**, v. 12, p. 203-214, 2002a.
RAYCHAUDHURI, S.; SCHÜTZE, H.; ALTMAN, R. B. Using text analysis to identify functionally coherent gene groups. **Genome Research**, v. 12, p. 1582-1590, 2002b.
RINDFLESCH, T. C.; TANABE, L.; WEINSTEIN, J. N.; HUNTER, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. **Pacific Symposium on Biocomputing**, p. 517-528, 2000.
SAFRAN, M.; SOLOMON, I.; SHMUELI, O.; LAPIDOT, M.; SHEN-ORR, S.;

ADATO, A.; BEN-DOR, U.; ESTERMAN, N.; ROSEN, N.; PETER, I.; OLENDER, T.; CHALIFA-CASPI, V.; LANCET, D. GeneCards 2000: towards a complete, object-oriented, human gene compendium. **Bioinformatics**, v. 18, p. 1542-1543, 2002.

SCHUEMIE, M. J.; WEEBER, M.; SCHIJVENAARS, B. J. A.; VAN MULLIGEN, E. M.; VAN DER EIJK, C. C.; JELIER, R.; MONS, B.; KORS, J. A. Distribution on information in biomedical abstracts and full-text publications. **Bionformatics**, v. 20, n. 16, p. 2597-2604, 2004.

SRINIVASAN, P. Text mining: generating hypotheses from MEDLINE. **Journal of the American Society for Information Science and Technology**, v. 55, p. 396-413, 2004.

SRINIVASAN, P.; LIBBUS, B. Mining MEDLINE for implicit links between dietary substances and diseases. **Bioinformatics**, v. 20 (Suppl. 1), p. 1290-1296, 2004.

STAPLEY, B. J.; BENOIT, G. Biobibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts. **Proceedings of Pacific Symposium on Biocomputing**, p. 529-540, 2000.

SWANSON, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. **Perspectives in Biology and Medicine**, v. 30, n. 1, p. 7-18, 1986.

SWANSON, D. R. Migraine and magnesium: eleven neglected connections. **Perspectives in Biology and Medicine**, v. 31, p. 526-557, 1988.

SWANSON, D. R. Two medical literatures that are logically but not bibliographically connected. **Journal of the American Society for Information Science**, v. 38, n. 4, p. 228-233, 1987.

TANABE, L. The genomic data mine. En: H. CHEN, H.; FULLER, S. S.; FRIEDMAN, C.; HERSH, W. (Eds.). **Medical informatics: knowledge management and data mining in biomedicine**. New York: Springer, 2005.

TANABE, L.; SCHERF, U.; SMITH, L.; LEE, J.; HUNTER, L.; WEINSTEIN, J. MedMiner: an Internet tex-mining tool for biomedical information, with application to gene expression profiling. **BioTechniques**, v. 27, n. 6, p. 1210-1217, 1999.

TUASON, O.; CHEN, L.; LIU, H.; BLAKE, J.; FRIEDMAN, C. Biological nomenclatures: a source of lexical knowledge and ambiguity. **Proceedings of the Pacific Symposium on Biocomputing**, p. 238-249, 2004.

WEEBER, M.; VOS, R.; KLEIN, H.; DE JONG-VAN DEN BERG, L. T. W.; ARONSON, A.; MOLEMA, G. Generating hypotheses by discovering implicit associations in the literature: a case report for new potential therapeutic uses for Thalidomide. **Journal of the American Medical Informatics Association**, v. 10, n. 3, p. 252-259, 2003.

WREN, J. D.; GARNER, H. R. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. **Bioinformatics**, v. 20, n. 2, p. 191-98, 2004.

YANDELL, M. D.; MAJOROS, W. H. Genomics and natural language processing. **Nature Reviews Genetics**, v. 3, p. 601-610, 2002.

YEH, A. S.; HIRSCHMAN, L.; MORGAN, A. A. Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. **Bioinformatics**, v. 19 (Suppl. 1), p. 331-339, 2003.

YU, H.; AGICHTEIN, E. Extracting synonymous gene and protein terms from biological literature. **BMC Bioinformatics**, v. 19, n. 1, p. 340-349, 2003.

YU, H.; HRIPCSAK, G.; FRIEDMAN, C. Mapping abbreviations to full forms in biomedical articles. **Journal of the American Medical Informatics Association**, v. 9, p. 262-272, 2002.

**ABSTRACT**
Since human genome sequences were first decoded, the paradigm of investigation has changed leading to the description of the functions of the genes and to future advances in the fight against diseases. This new context has awoke the interest of the Bioinformatics, that combines methods of the Life Science with the Information Sciences, making the access to the great quantity of biological information stored in the databases, and of the Genomics, dedicated to the study of the interactions of the genes and its influence in the development of diseases. In this context, the text mining arises like an emerging instrument for the analysis of the scientific literature. A habitual task of text-mining in Molecular Biology and Genomics is the recognition of biological entities, such as genes, proteins and diseases. The following step in the process of text-mining constitutes it the identification among biological entities (such as the type of interaction among *gene-gene*, *gene-disease*, *gene-protein*) to interpret biological functions, or to formulate research hypothesis. The objective of this work is to examine the growth and the limitations the new analysis tools of the information in natural language, stored in unstructured textual databases or bibliographical databases, such as *MEDLINE* or *PubMed*.

**KEYWORDS:** Text-Mining. Textual Databases. Natural Language Processing (NLP).

*Originais recebidos em: 19/10/2007*
*Texto aprovado em: 13/03/2008*