Carmen Galvez[a], Félix Moya-Anegón[a]
a Scimago Research Group, University of Granada, Faculty of Library and Information Science, Granada, SPAIN

Identificación de Nombres de Genes en la Literatura Biomédica

Pages: 344-348

Current Research in Information Sciences and Technologies
Multidisciplinary Approaches to Global Information Systems

Vicente P. Guerrero-Bote (Editor)

Proceedings of the I International Conference on Multidisciplinary
Information Sciences and Technologies, InSciT2006
Mérida - SPAIN
October, 25th-28th, 2006

ISBN-10 CD-ROM: 84-611-3106-1
ISBN-13 CD-ROM: 978-84-611-3106-8

ISBN-10 Whole Edition: 84-611-3103-7
ISBN-13 Whole Edition: 978-84-611-3103-7

ISBN-10 Volume II: 84-611-3105-3
ISBN-13 Volume II: 978-84-611-3105-3

Open Institute of Knowledge
(Instituto Abierto del Conocimiento)
Antonio Álvarez 6
06005 Badajoz, SPAIN
Phone: +34 924221935
Fax: +34 924221935
http://www.instac.es

Identificación de Nombres de Genes en la Literatura Biomédica

Carmen Galvez[a1] , Félix Moya-Anegón[a]
[a] *Scimago Research Group, University of Granada, Faculty of Library and Information Science, Granada, SPAIN*

An enormous complexity arises in the identification of gene terms in biomedical literature. With the discovery of huge quantities of genes, and the Human Genome Project (HGP), the scientists have remained without easy and intuitive names. In the genomic information many forms of variation occur due to lack of standardization of gene names. Although nomenclature and ontological specifications are valuable for processing, efforts toward the systematic naming of genes have been made, but the difficulty still exists. The development of procedures that resolve these problems would benefit the progress of *molecular pathways*, the extraction of *gene-gene* and *gene-disease* interactions, the delimitation of the structure of the genomic research domain through *gene-document* relations and the knowledge discovery that is hidden in the biomedical literature. Our proposal relies on approximate pattern-matching techniques, adopted of Natural Language Processing (NLP), to find and filter gene variants matches. To perform the gene-matching, we apply Finite-State Transducers (FSTs). To implement our prototype system, we using publicly available gene and text databases, such as FlyBase (biological database of the *Drosophila genome projects*) and PubMed (*U.S. National Library of Medicine*).

Keywords: Bioinformatics, Gene Name Variations, Finite-State Transducers, Natural Language Processing.

1 INTRODUCCIÓN

Las Ciencias de la Vida han avanzado a una velocidad sin precedentes en los últimos años como resultado, en parte, del *Human Genome Project* (HGP). El número de publicaciones y bases de datos biomédicas disponibles electrónicamente se ha incrementado de forma exponencial. Esta situación hace muy difícil, casi imposible, que los científicos biomédicos puedan conocer los avances producidos en su dominio de interés. A esto, se añade que el establecimiento de comparaciones entre genes similares en diferentes organismos específicos y bases de datos es crucial para dar sentido a la inmensa cantidad de información genómica. Ante a esta situación, los biólogos moleculares y los comisarios, o *curators*, de las bases de datos biológicas, se enfrentan básicamente a tres grandes retos: (*i*) identificación de la información en la literatura biomédica; (*ii*) exploración y análisis de las bases de datos disponibles para ayudar a interpretar un fenómeno, o adoptar decisiones de acuerdo con las necesidades, tales como el análisis de las interacciones genómicas *gen-gen* o *gen-enfermedad* y su relación con el desarrollo de enfermedades – lo que se denomina comúnmente *genómica comparativa* y que incluye el diagnóstico clínico, la investigación de nuevos fármacos, la epidemiología, la informática médica, o el desarrollo de los test genéticos basados en el ADN; y (*iii*) delimitación y mapeo de datos aparentemente separados, tales como redes de *gen-gen*, *gen-literatura* o *gen-documento* que pueden ayudar a formular nuevas hipótesis de investigación, o ayudar a la visualización de la estructura del dominio de la investigación genómica. Todos estos retos, y otros muchos, se ven obstaculizados por la falta de sistemas homologados para denominar a los genes, amenazando con ello los beneficios que se pudieran derivar de la secuencia del HGP.

El reconocimiento de los nombres de los genes es el primer paso para la identificación y utilización del conocimiento codificado en la literatura y en las bases de datos biomédicas. La Biología moderna debe manejar una inmensa cantidad de datos que no pueden ser interpretados sin ayuda computacional. Estas cuestiones han despertado el interés de la recuperación de información (RI), del procesamiento de lenguaje natural (PLN), de la extracción de información (EI), y de los métodos estadísticos. A su vez, se están configurando áreas emergentes de interés, tales como el análisis *BioBibliométrico* [1], las técnicas de visualización [2], o el análisis de redes [3]. La Bioinformática, una disciplina que se encuentra en la intersección entre las Ciencias de la Vida y de la Información, proporciona los recursos necesarios para ayudar a la investigación biomédica, ocupándose del análisis de la información biológica, a través del procesamiento, extracción de información y consulta de datos. La Bioinformática integraría, por tanto, métodos lingüísticos, matemáticos, estadísticos y de las ciencias de la computación para analizar datos biológicos, bioquímicos y genómicos.

Uno de los desafíos a los que se enfrenta la Bioinformática es la enorme complejidad que presenta la identificación de los términos genéticos en los textos biomédicos, debido a la falta de una nomenclatura

[1] Corresponding Authors: C. Galvez & F. Moya-Anegón. Faculty of Library and Information Science, University of Granada, 18071 Granada, Spain. cgalvez@ugr.es; felix@ugr.es

común. En este trabajo se va a proponer un modelo prototipo, no utilizado hasta ahora en este ámbito científico, que facilite la identificación y unificación de los diferentes nombres de genes en los textos biomédicos. Con esta finalidad, vamos a presentar un procedimiento adoptado del PLN, basado en la aplicación de métodos de estado-finito. El material para poder desarrollar nuestra aplicación nos lo va a proporcionar la base de datos biológica FlyBase[2] (*Drosophila melanogaster*) y el sistema de búsqueda PubMed Search Engine[3] (proyecto desarrollado por la *National Center for Biotechnology Information, NCBI*, en la *National Library of Medicine, NLM*) que permite el acceso a bases de datos bibliográficas compiladas por la NLM, tales como *Medline*, *PreMedline*, *Genbak* y *Complete Genome*.

2 EL PROBLEMA DE LA DENOMINACIÓN DE LOS GENES

Con el avance de las investigaciones y el descubrimiento de cantidades ingentes de genes, los científicos se han quedado sin nombres fáciles e intuitivos. El resultado de todo ello es la utilización de nombres arbitrarios, en los que se abusa de abreviaturas y siglas. En los procesos de identificación de tales términos se producen básicamente dos tipos de inconsistencias:

- *Problemas de sinonimia*: un único nombre de gen puede tener un gran número de sinónimos, tales como el gen **AcCoAS**, con 9 alias (**CG9390**, **acetate-coenzyme A ligase, acetyl-CoA synthetase, acetyl CoA synthase, Acetyl CoA synthase, ACS, Acetyl-CoA synthase, Acetyl CoA synthetase, BEST:GH2840**).
- *Problemas de ambigüedad*: un único nombre de gen puede referirse a múltiples genes, o incluso puede ser la abreviatura de términos no-genéticos completamente diferentes: el gen **PSA** se refiere a los genes **Puromycin-Sensitive Aminopeptidase**, **Prostate Specific Antigen**, **PSoriatic Arthiritis**, **Phosphoserine Aminotransferase**, o a un término completamente diferente **Poultry Science Association**.

Ante la falta de denominaciones oficiales, el Consorcio de Ontología Genética, *Gene Ontology (GO) Project*[4], ha desarrollado vocabularios controlados que vinculan los genes de diferentes bases de datos genómicas sin necesidad de establecer un sistema homologado de denominaciones. Los términos GO proporcionan tres redes estructuradas de términos controlados para describir los atributos de los genes. Los tres principios de organización de los términos GO son: *función molecular*, *procesos biológicos*, y *componentes moleculares*. Con este sistema común se produce un vocabulario controlado que se puede aplicar a cualquier organismo. Muchas bases de datos de diferentes organismos asignan ya términos GO a cada gen y a sus productos. Sin embargo, aunque finalmente se implanten los términos GO, las bio-ontologías codifican sólo una pequeña fracción de información, y el problema todavía existe. Por esta razón, el desarrollo de herramientas capaces de identificar los nombres de los genes sigue siendo relevante para capturar información de la literatura biomédica y transferir esa información a las bases de datos biológicas, que deben ser continuamente actualizadas, en el proceso denominado por los biólogos '*data curation'*.

Nuestro objetivo en este trabajo se dirige únicamente a las inconsistencias producidas por los problemas de sinonimia, dejando fuera, por ahora, los problemas de ambigüedad originados cuando un mismo término se refiere a entidades genómicas distintas. Son varios los procedimientos de identificación de nombres de genes que se han desarrollado [4]. La mayor parte de los sistemas se podrían agrupar en tres grandes enfoques: a) *métodos basados en diccionarios*, que implican la aplicación de técnicas de equiparación aproximada de cadenas, tales como el sistema *BLAST* [5]; b) *métodos basados en reglas,* que usan analizadores sintácticos, tales como el sistema *Kex* [6] y el sistema *Yapex* [7]; y c) *métodos basados en técnicas estadísticas*, que emplean aproximaciones estadísticas, tales como Modelos de Makov y procedimientos probabilísticos para la identificación y clasificación de nombres de genes [8]. Nuestra propuesta para la identificación y unificación de las variantes de los nombres de los genes, consistirá en el desarrollo de un procedimiento para normalizar las diferentes denominaciones aplicando matrices binarias y redes de transición.

3 MATERIAL Y MÉTODO

El proceso de unificación de genes en nuestro sistema prototipo requiere tres etapas: 1) obtención de una lista de nombres de genes, en la que cada entrada de la lista representaría un gen específico, que contendría tanto el identificador único para ese gen como un conjunto de formas diferentes o alias, por medio

de las cuales el gen puede ser mencionado; 2) obtención de una muestra de documentos biomédicos en los que aparezcan los nombres de los genes seleccionados; y 3) desarrollo de un proceso que nos permita equipar las diferentes denominaciones de los genes, en los documentos biomédicos, a un identificador de gen único.

Para conseguir la lista de nombres de genes realizamos una consulta aleatoria en la base de datos biológica FlyBase. Por medio de este recurso conseguimos nombres de genes y listas de sinónimos, junto con su correspondiente identificador único en FlyBase. La Fig. 1 muestra una parte de la entrada FlyBase para el gen *AcCoAS*, en la que se distinguen, entre otros datos: el nombre completo del gen, **Full name**, el nombre o símbolo oficial, **Official Symbol**, un enlace a los **Sinónimos** del gen, el identificador único **FlyBase ID** asignado por la base de datos, y los términos GO que describen el gen, según la estructura **Función Molecular**, **Proceso Biológico** y **Componente Molecular**. Por otra parte, la muestra de documentos biomédicos se obtuvo a partir de una consulta en PudMed limitada al campo *Abstract* (AB), usando sólo el **Full name** de la lista de nombres de genes seleccionados en FlyBase. Un total de 9,605 registros fueron recuperados de PubMed, de los cuales guardamos como archivo de texto los primeros 500 registros del resultado total de la búsqueda.

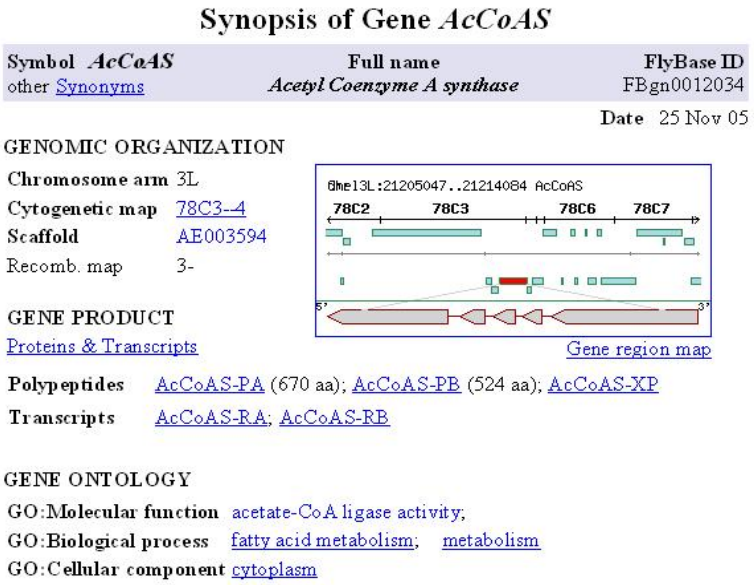

Fig. 1. Entrada de la base de datos biológica FlyBase para el gen *AcCoAS*

Con el material proporcionado por FlyBase y PubMed, proponemos un procedimiento capaz de establecer una correspondencia de las diferentes denominaciones de un gen con una forma unificada, que en este caso se va a obtener del identificador único del gen. La propuesta se basa en la aplicación de *Finite-State Transducers* (FSTs), considerados modelos matemáticos de un sistema con *input* y *output*, caracterizado como un conjunto de estados y conjunto de redes de transición de un estado a otro. Un FST acepta el nombre de un gen si se produce una transición, etiquetada con símbolos de entrada, desde el estado inicial a un estado final, etiquetada con símbolos de salida. Los transductores se encargan de establecer relaciones entre lenguajes regulares. Para computar las relaciones, el transductor etiqueta las transiciones con dos símbolos de los alfabetos de *input* y *output*. Formalmente, un FST [9] *se define como una 6-tupla* $(Q, \Sigma_1, \Sigma_2, q_0, F, \delta)$, *donde*

- Q *es un conjunto finito de estados*
- Σ_1 *es un alfabeto finito de input*
- Σ_2 *es un alfabeto finito de output*
- q_0 *es el estado inicial*
- F *es un conjunto de estados finales*
- δ *es una función total que mapea* $Q \times \Sigma_1 \cup \{\varepsilon\} \times \Sigma_2 \cup \{\varepsilon\}$ *en* 2^Q

Partiendo de la premisa de que los transductores se utilizan para representar relaciones de equivalencia entre lenguajes, consideramos que el problema de la normalización de términos se podría resolver si se planteara como una relación que equipara variantes de términos a formas normalizadas. Para establecer la mencionada relación proponemos la utilización de gráficos parametrizados, definidos como gráficos de estado-finito compilados en FSTs, cuyo alfabeto de *input* y *output* contiene parámetros con los valores almacenados previamente en una *matriz binaria*.

4 RESULTADOS Y DISCUSIÓN

Sobre la muestra recopilada aplicamos un gráfico parametrizado a partir de los datos almacenados en una matriz binaria (Tabla 1). Usando una aplicación gráfica, *FSGraph* [10], trazamos una red de transición compilada en un transductor con parámetros (Fig. 2), que se enlazan al contenido de la matriz por medio de *variables* (@A, @B, @C,..). El proceso de unificación se realizaría, entonces, por *equiparación parametrizada de variables*: teniendo como *input* las variables que se refieren a los sinónimos de los genes, y como *output* las diferentes denominaciones asociadas al identificador de código único asignado por FlyBase.

Tabla 1. Matriz binaria

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Official Symbol	Full name	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	Other Aliases	FlyBase ID
aact	anti-activin		anti-activin	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0082611
Abi	Abelson Interacting Protein	CG9749	CG9749	Abiphilin			-	-	-	-	-	-	-	-	-	-	FBgn0020510
Abl	Abl tyrosine kinase		l(3)c-abl	D-ash	abl	Dabl	cAbl	D-abl	l(3)04674	abelson	DROTKABL3	Am ABL	DAbl	Dash	4674	c-abl	FBgn0000017
ac	achaete	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0000022
Ac3	Ac3	CG1506	DAC3	DAC39E	AC 3	-	-	-	-	-	-	-	-	-	-	-	FBgn0023416
acc	accilnal wing	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0000023
AcCoAS	Acetyl Coenzyme A synthase	ligase	acetyl-CoA synthetase	acetyl CoA synthase	Acetyl CoA synthase	ACS	Acetyl-CoA synthase	Acetyl CoA synthetase	BEST:GH28401	-	-	-	-	-	-	-	FBgn0012034
Ace	Acetylcholine esterase	CG17907	AChE	AcChE	ACE	CHE	ache	dAChE	ace	AchE	dmAChE	DmAChE	acetylcholinesterase	l(3)26	l(3)67Ed	-	FBgn0000024
Ace::Nrt	Ace::Nrt	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0020234
achi	achintya	CG8819	zaa	zuker-aly-like a	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0033749
Ack	Ack	CG14982	DACK	ACK	ACK2	dACK	p145	DAck	BcDNA:GH10777	-	-	-	-	-	-	-	FBgn0028484
Aco	Accordion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0024502
Acon	Aconitase	CG9244	ACON-1	AconM	mAc	mitochondrial aconitase	aconitase	m-Aconitase	l59	l204	m-acon	l(2)07054	BEST:GH10550	-	-		FBgn0010100
Acox57D-d	acyl-Coenzyme A oxidase at 57D distal	CG9709	acyl co-enzyme A oxidase	D-acox	acox	Acox57Dd	-	-	-	-	-	-	-	-	-	-	FBgn0034629
Acox57D-p	acyl-Coenzyme A oxidase at 57D proximal	CG9707	acyl co-enzyme A oxidase	D-acox	acox	Acox57Dp	-	-	-	-	-	-	-	-	-	-	FBgn0034628
Acp24A4	Acp24A4	CG31779	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0051779
Acp26Aa	Accessory gland-specific peptide 26Aa	CG8982	mst355a	mst 355a	Mst26A	Acp26A	Mst26Aa	ovulin	Acp26Aab	Ovulin	msP355a	-	-	-	-	-	FBgn0002855
Acp26Ab	Accessory gland-specific peptide 26Ab	CG9024	mst355b	mst 355b	Mst26A	Acp26A	Mst26Ab	Acp26Aab	msP355b	-	-	-	-	-	-	-	FBgn0002856
Acp29AB	Accessory gland-specific peptide 29AB	CG17797	Acp29B	Accessory gland protein 29AB	AC 004423B	AC004423b	accessory gland protein	Acp29AB-II	-	-	-	-	-	-	-	-	FBgn0015583
Acp31F	Accessory gland protein 31F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0024501
Acp32CD	Accessory gland-specific peptide 32CD	CG4605	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0023415
Acp33A	Accessory gland-specific peptide 33A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0023414
Acp36DE	Accessory gland peptide 36DE	CG7157	Accessory gland protein	36DE accessory gland protein	CT22109	anon-WO0140519.55	-	-	-	-	-	-	-	-	-	-	FBgn0011559
Acp53C14a	Acp53C14a	CG8626	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0034152
Acp53C14c	Acp53C14c	CG33530	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0053530
Acp53Ea	Accessory gland-specific peptide 53Ea	CG8622	Acp53E	Accessory gland protein 53Ea	BcDNA:AT09943	-	-	-	-	-	-	-	-	-	-	-	FBgn0015584
Acp53Eb	Accessory gland protein 53Eb	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0024500
Acp54A1	Acp54A1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0083936
Acp63F	Accessory gland peptide 63F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	FBgn0015585
Acp65Aa	CG10297	Acp65A	DMACP65A	BcDNA:RH61147	DCP5	-	-	-	-	-	-	-	-	-	-	-	FBgn0020765

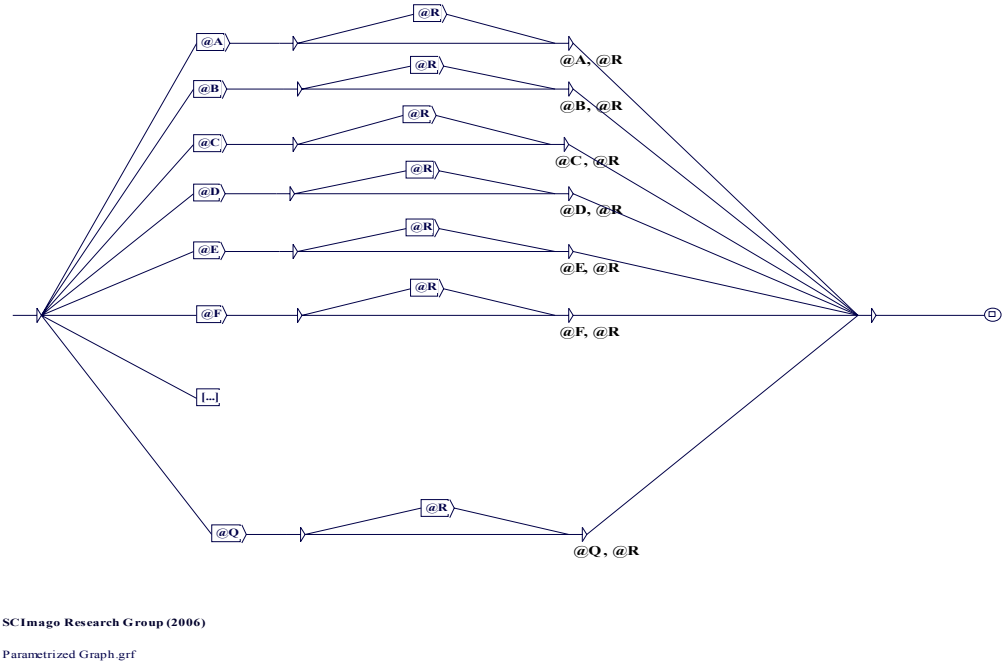

Fig. 2. Red de transición parametrizada

Un total de 554 nombres de genes son identificados, que corresponderían a 26 nombres de genes únicos, como los siguientes ítems:

Acetylcholine esterase, **FBgn0000024**
acetylcholinesterase, **FBgn0000024**
AChE, **FBgn0000024**
ache, **FBgn0000024**

Aconitase, **FBgn0010100**
aconitase, **FBgn0010100**
m-Aconitase, **FBgn0010100**
mitochondrial aconitase, **FBgn0010100**

Sin embargo, con el procedimiento propuesto surgen problemas de ambigüedad cuando un mismo nombre de gen puede estar relacionado con dos identificadores diferentes, como: "*acyl co-enzyme A oxidase*, **FBgn0034629**" y "*acyl co-enzyme A oxidase*, **FBgn0034628**". En estos casos, el sistema se limita a realizar la equiparación por la primera forma que encuentra en la matriz binaria, es decir, la variante de gen "*acyl co-enzyme A oxidase*" se correspondería únicamente con el identificador "**FBgn0034629**". Para evitar este grave error, con consecuencias fatales en posteriores aplicaciones, se podrían aplicar procedimientos capaces de distinguir el contexto donde los nombres de genes aparecen, tales como: a) métodos cuantitativos basados en redes que utilizan probabilidades de transición, como *Modelos de Markov* o *transductores probabilísticos* que combinan la idea de la probabilidad condicionada por el contexto (los *n-grams*) con la noción de sucesos encadenados; o b) métodos basados en relaciones de similaridad, como el *modelo de espacio vectorial*, el *análisis de co-ocurrencia,* o las *técnicas de clustering* fundamentadas en la similaridad contextual.

5 Conclusión

La ambigüedad es una de las mayores dificultades en la identificación y normalización de nombres de genes. Si hay una relación unívoca, de *uno-a-uno* entre *nombre de gen-identificador*, el modelo de equiparación parametrizada de variables es adecuado, resolviendo un porcentaje elevado de casos. Sin embargo, cuando el sistema tiene que escoger entre varias conexiones posibles, la probabilidad de transición y las medidas de similaridad podrían ayudar a establecer la correspondencia exacta. Los métodos cuantitativos aportarían así soluciones al gran obstáculo de los modelos cualitativos. Es necesario, por tanto, la combinación de diferentes procedimientos, que consigan superar las limitaciones inherentes a cada uno, para resolver este persistente problema.

NOTAS

[2] Disponible en: <http://www.flybase.org>
[3] Disponible en: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed>
[4] Disponible en: <http://www.geneontology.org/>

Referencias

[1] Stapley, B.J., Benoit, G. "Biobibliometrics: Information Retrieval and Visualization from Co-Occurrence of Gene Names in Medline Abstracts." Proc. of Pacific Symposium on Biocomputing, pp. 529-540, 2000.
[2] Jenssen, T.-K., Laegreid, A., Komorowski, J., and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 28(1), pp. 21-28, 2001.
[3] Boyack, K., Mane, K., and Börner, K. "Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research." Proc. of Eight International Conference on Information Visualization, pp. 965-971, 2004.
[4] Galvez, C., Moya-Anegón, F. "Extracción y Normalización de Entidades Genómicas en Textos Biomédicos: una Propuesta Basada en Transductores Gráficos." Proc. of I Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI), 2006.
[5] Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. Using BLAST for identifying gene and protein names in journal articles. Gene, 259, pp. 245-252, 2000.
[6] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. "Toward Information Extraction: Identifying Protein Names From Biological Papers." Proc. of the Pacific Symposium on Biocomputing, pp. 705-716, 1998.
[7] Olsson, F., Eriksson, G., Franzén, K., Asker, L., and Liden, P. "Notions of Correctness when Evaluating Protein Name Taggers." Proc. of the 19th International Conference on Computational Linguistics, pp. 765-771, 2002.
[8] Collier, N., Nobata, C., and Tsujii, J. "Extracting the Names of Genes and Gene Products with a Hidden Markov Model." Proc. of the 18th International Conference on Computational Linguistics, pp. 201-207, 2000.
[9] Roche, E., Schabes, Y. Finite State Language Processing. Cambridge, Massachusetts: MIT Press, 1997.
[10] Silberztein, M. "The Lexical Analysis of Natural Language." In: Roche, E., Schabes, Y. (Eds.), Finite-State Processing. Cambridge, Massachusetts: MIT Press, 1997.