CISTI 2006
ISBN: 978-989-20-0271-2
Volume I, pág. 697

Extracción y normalización de entidades genómicas en textos biomédicos: una propuesta basada en transductores gráficos

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 falta de sistemas homologados para denominar a los genes es un problema para la identificación de información en la literatura biomédica y hace muy difícil un proceso esencial en el campo de la biología molecular: encontrar y descubrir relaciones biológicas, entre genes, en aquellos documentos que tratan la misma entidad genómica pero que usan símbolos distintos. Nosotros proponemos un procedimiento adoptado del procesamiento de lenguaje natural (PLN) basado en la aplicación de transductores de estado-finito que permite el reconocimiento de los diversos nombres de un gen y los relaciona con una forma unificada. El proceso de normalización requiere como *input* una lista de sinónimos, y como *output* un identificador único para ese gen. La base de datos genómica *FlyBase* nos ha aportado los recursos necesarios para exponer nuestra propuesta.

Palabras-clave: Extracción de entidades genómicas; Normalización de genes; Procesamiento del Lenguaje Natural; Transductores de estado-finito.

1. Introducción.

El campo de la biología molecular ha experimentado una auténtica revolución científica. La cantidad de información sobre el genoma ha aumentado exponencialmente en muy poco tiempo. Los biólogos moleculares, ingenieros genéticos y biotecnólogos descubren constantemente nuevos genes y proteínas que hay que nombrar. De forma paralela, se ha incrementado el desarrollo de sistemas automáticos para identificar los datos sobre el genoma en la literatura biomédica. El reconocimiento de los nombres de genes y proteínas en textos biomédicos ha desencadenado la necesidad de adoptar técnicas del procesamiento de lenguaje

natural (PLN) y de la recuperación de información (RI) para filtrar y extraer la inmensa cantidad de información generada sobre el genoma. La biología es ahora tanto una ciencia de laboratorio como una ciencia de la información (Morgan *et al*., 2004).

La extracción de información (EI) es una disciplina perteneciente al PLN, que se define como el conjunto de técnicas usadas para obtener datos estructurados y no-ambiguos del lenguaje natural con diferentes propósitos, tales como la construcción de bases de datos, o aplicaciones relacionadas con la RI (Cunningham, 2005). La EI es esencial para analizar y extraer información útil de los textos biomédicos, imposible de realizar de forma manual, y donde la tecnología de RI convencional resulta inadecuada debido a la complejidad y falta de terminología normalizada. Pero, lo más importante, la EI es crucial en el campo de la biología molecular debido a la necesidad urgente de este ámbito científico por el descubrimiento automático de rutas o *pathways* moleculares y relaciones biológicas entre genes en la literatura especializada. Fundamentalmente por esta última razón, son muchos los trabajos dedicados a la investigación sobre el empleo de técnicas de EI a los textos biomédicos (Proux *et al*., 1998; Ng & Wong, 1999; Thomas *et al*., 2000, Friedman *et al*., 2001; Hirschman *et al*., 2002; Seki & Mostafa, 2005).

Uno de los mayores obstáculos para la EI, y para los biólogos, lo constituye la denominación 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. Paul Smaglik (1998) en un trabajo publicado en *The Scientist* cita a un miembro de *HUGO (Human Genome Organization)-Gene Nomenclature Committee*, Julia A. White, que indicaba que aunque el Comité pretende eliminar el caos lingüístico se queda detrás como resultado de la velocidad del *Human Genome Project*, con cientos de miles de genes todavía por bautizar. 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 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).

La demanda de información normalizada es crítica, asimismo, para un área de investigación de bioinformática, denominada *genómica comparativa*, que consiste básicamente en analizar cualquier aspecto biológico de los genomas de organismos distintos, mediante la comparación de los genomas animales con el genoma humano, para determinar sus diferencias y similitudes. La secuenciación del genoma tiene como aplicación última la cura de enfermedades y la mejora de la salud. Para ello es necesario no sólo conocer la secuencia genética de los seres humanos sino la de otros seres vivos para averiguar qué función cumplen los genes y poder desvelar los secretos de la evolución y de las enfermedades. Esas comparaciones se realizan a través de las similitudes de la información almacenada en las distintas bases de datos de los organismos específicos. Por esta razón, el

establecimiento de denominaciones oficiales a los genes constituye un esfuerzo constante por parte de los científicos y es un desafío cada vez mayor debido a la creciente información biomédica. Las múltiples denominaciones de los genes amenazan a los beneficios que se pudieran derivar de la secuencia del genoma humano.

Ante la falta de denominaciones consensuadas, y en un esfuerzo para dirigir la necesidad de descripciones coherentes de los genes, el Consorcio de Ontología Genética, *Gene Ontology (GO) Project*[1], ha desarrollado vocabularios controlados y estructurados 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: a) *Función molecular*; b) *Procesos biológicos*; y c) *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.

No obstante, aunque finalmente se implanten los términos GO para la anotación de los genes, 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, que deben ser continuamente actualizadas. Nuestro objetivo en este trabajo es proponer un modelo, todavía incipiente en este ámbito científico, que facilite el proceso de reconocimiento y la interacción de los genes en los textos biomédicos. Con esta finalidad, vamos a presentar un método adoptado del PLN, basado en la aplicación de transductores de estado-finito o *finite-state transducers (FST)*, que permita la identificación de las entidades genómicas y las asocie con un término normalizado, definido en el sistema de nomenclatura estandardizado. A su vez, las bases de datos del genoma de los principales organismos nos van a proporcionar los recursos necesarios para poder realizar esta aplicación PLN, porque publican especificaciones fiables y actualizadas sobre las entidades biomoleculares. De entre estos recursos, los más valiosos son las bases de datos genómicas de organismos específicos, tales como: GBD[2] (*Genoma Humano*), FlyBase[3] (*Drosophila melanogaster*), WormBase[4] (*Caenorhabditis elegans*), o Mouse Genome Informatics[5] (*Mus musculus*).

[1] Disponible en: <http://www.geneontology.org/>
[2] Disponible en: <http://gdbwww.gdb.org/>
[3] Disponible en: <http://www.flybase.org>
[4] Disponible en: <http://www.wormbase.org/>
[5] Disponible en: <http://www.informatics.jax.org>

2. El problema de la denominación de los genes.

El primer paso para poder unificar las diferentes denominaciones de los genes es su identificación. Como en cualquier otro proceso de reconocimiento y extracción de información se pueden emplear básicamente dos aproximaciones. La primera consiste en la aplicación de reglas heurísticas para identificar los nombres de los genes, o de bases de conocimiento, tales como diccionarios. La segunda consiste en la aplicación de procesos de aprendizaje automático o *machine learning methods*, para crear las reglas y derivar las entidades etiquetadas de acuerdo a la información que se pretende extraer. Sin embargo, el primer obstáculo, independientemente del método utilizado, para la identificación de genes en los textos biomédicos es la falta de consenso sobre las denominaciones genéticas. Esta limitación hace que surjan algunas de las siguientes dificultades:

- *Problemas de homonimia*: 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*.
- *Problemas de sinonimia*: un único nombre de gen puede tener un gran número de sinónimos, tales como el gen *Acf1*, con 14 alias (*CG1966*, *ACF*, *ATP*, *CAF*, *acf1*, *p170/p185*, *CHRAC*, *Chromatin Accessibility Complex*, *dACF*, *dCHRAC*, *ACF1*, *Acf-1*, *Acf*, y *CHRAC-175*).
- *Problemas de normalización*: diferentes denominaciones de un mismo gen, que aparecen dispersas en los textos biomédicos, pueden ser asociadas con una forma unificada, o con un identificador único.

Varios trabajos han aplicado técnicas de desambiguación para resolver el problema de la homonimia por medio de métodos de aprendizaje automático (Hatzivassiloglou, Duboue & Rzhetsky, 2001; Liu, Lussier & Friedman, 2001; Liu, Johnson & Friedman, 2002). Entre los estudios dedicados a los problemas de sinonimia se encuentran procedimientos automáticos para reconocer sinónimos usando tesauros (Schijvenaars *et al*., 2005). En general, los problemas de las palabras con múltiple sentido y la ambigüedad se han tratado extensamente por Tuason *et al.* (2004).

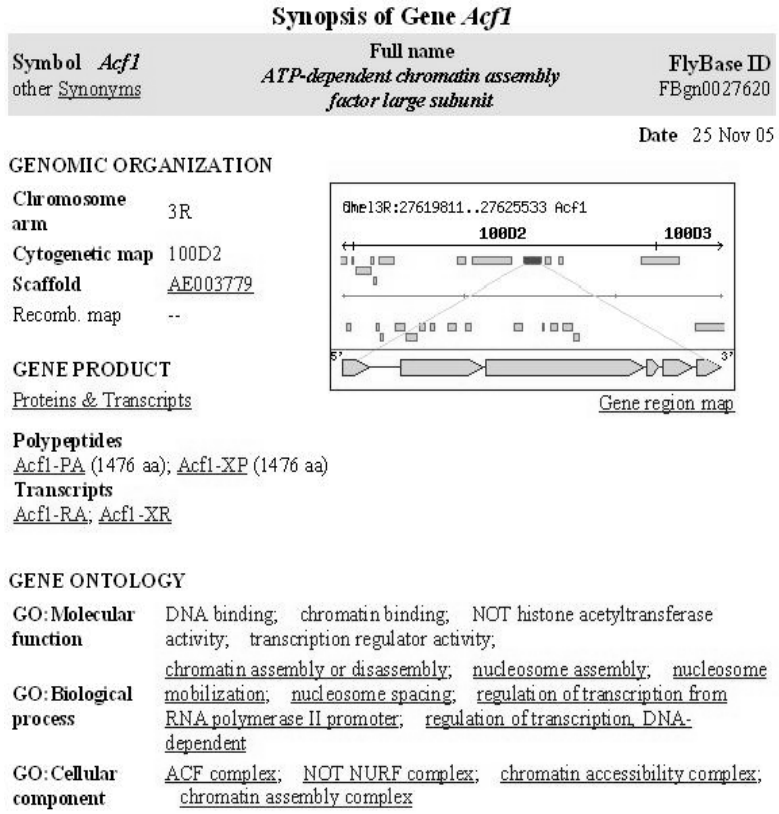
Frente a estas investigaciones, el problema de la normalización de genes es un campo relativamente nuevo e inexplorado (Crim, McDonald & Pereira, 2005). Para enfrentarnos a esta cuestión tendríamos que seguir básicamente dos etapas, según Morgan *et al*. (2004): primera, anotación de los documentos biomédicos con las listas de los identificadores de los genes mencionados en los documentos; y segunda, equiparación de las distintas denominaciones de los genes reconocidos con un identificador único del gen, dentro del organismo específico. En este estudio

vamos a proponer un procedimiento semiautomático, con el objetivo de normalizar las diferentes denominaciones de los genes, basado en técnicas de equiparación de patrones y gráficos de estado-finito. Fuera de este trabajo quedarían los problemas de ambigüedad producidos por homonimia, en los que el nombre de un gen puede referirse a múltiples genes.

3. Normalización de nombres de genes usando transductores gráficos.

El proceso de unificación de genes en nuestro proyecto requiere dos etapas: (*i*) obtención de una lista de sinónimos, en la que cada entrada de la lista representaría un gen específico, que contendría tanto el identificador único para ese gen, denominado la forma estandardizada, y un conjunto de formas diferentes por medio de las cuales el gen puede ser mencionado; y (*ii*) equiparación de las diferentes denominaciones de los genes con un identificador de gen único.

Para conseguir la lista de sinónimos utilizamos, en este caso, los recursos proporcionados por la base de datos FlyBase, especializada en el genoma de la mosca del vinagre *Drosophila melanogaster*. Esta base de datos aporta listas de sinónimos de cada gen, junto con su correspondiente identificador único en FlyBase. La Fig. 1 muestra una parte de la entrada FlyBase para el gen *Acf1*, en la que se distinguen, entre otros datos: un enlace a los sinónimos del gen, el identificador único *FBgn0027620* asignado por la base de datos, y los términos GO que describen el gen, según la estructura *función molecular*, *proceso biológico* y *componente molecular*. Con esta información, nosotros proponemos realizar una equiparación de las diferentes denominaciones de un gen, a partir de la lista de sinónimos, con una forma unificada, obtenida del identificador único del gen, por medio de *transductores de estado-finito*.



Synopsis of Gene *Acf1*

Symbol *Acf1* other Synonyms	Full name *ATP-dependent chromatin assembly factor large subunit*	FlyBase ID FBgn0027620

Date 25 Nov 05

GENOMIC ORGANIZATION

Chromosome arm	3R
Cytogenetic map	100D2
Scaffold	AE003779
Recomb. map	--

Gene region map

GENE PRODUCT

Proteins & Transcripts

Polypeptides
Acf1-PA (1476 aa); Acf1-XP (1476 aa)
Transcripts
Acf1-RA; Acf1-XR

GENE ONTOLOGY

GO: Molecular function	DNA binding; chromatin binding; NOT histone acetyltransferase activity; transcription regulator activity;
GO: Biological process	chromatin assembly or disassembly; nucleosome assembly; nucleosome mobilization; nucleosome spacing; regulation of transcription from RNA polymerase II promoter; regulation of transcription, DNA-dependent
GO: Cellular component	ACF complex; NOT NURF complex; chromatin accessibility complex; chromatin assembly complex

Figura 1 – Entrada en la base de datos FlyBase del gen *Acf1*.

Los transductores de estado-finito son modelos matemáticos de un sistema con *input* y *output*, se definen como un conjunto de estados y conjunto de transiciones de un estado a otro (Hopcroft & Ullman, 1979). 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 transductor de estado-finito se caracteriza por una tupla de cinco elementos, $T = (Q, \Sigma, q_0, F, \delta)$, donde, Q es el conjunto de estados, Σ es el alfabeto de *input* y *output*, q_0 es es estado inicial, F es el conjunto de estados finales, y δ es el conjunto de transiciones de un estado a otro (Roche & Schabes, 1995). Los transductores se pueden representar como gráficos dirigidos, cuyos vértices denotan los estados, mientras que las transiciones constituyen los arcos que llevan de un estado inicial a un estado final.

Usando una interfaz gráfica, *FSGraph* desarrollada por Silberztein (2000), nosotros dibujamos gráficos de estado-finito que tendrían como *input* la lista de sinónimos de los genes y como *output* el identificador de código único asignado por FlyBase. La Fig. 2 muestra el transductor gráfico encargado de fusionar las distintas denominaciones, o sinónimos, del gen *Acf1* con el identificador *FBgn0027620.*

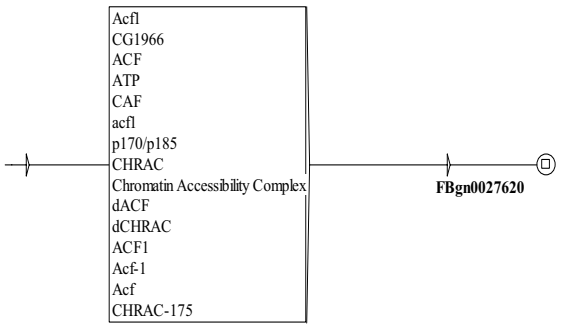


Figura 2 – Transductor gráfico que normaliza las diferentes denominaciones del gen *Acf1*.

El gráfico de estado-finito se compila en un transductor de estado-finito, además la propia aplicación permite transformar el gráfico en una tabla o matriz de transición (Tabla 1), en la cual se especifican los siguientes componentes:

- Número de estados, $Q = 25$.
- Número de símbolos del alfabeto, $\Sigma = 24$.
- Estado inicial, $q_0 = 0$.
- Estado final, $F = 1$.
- Número de transiciones entre estados, $\delta = 38$.

Tabla 1 – Transformación del gráfico de estado-finito en una tabla de transición

Characteristics of the FST

```
#define NUMBER_OF_STATES 25 // states are numbered from 1 to NUMBER_OF_STATES

#define NUMBER_OF_SYMBOLS 24 // symbols are numbered from 0 to NUMBER_OF_SYMBOLS-1

#define NUMBER_OF_TRANSITIONS 38



// FST Alphabet/Vocabulary

static const char *symbols[NUMBER_OF_SYMBOLS]={"p","Acf","CG","ACF","acf","ATP","CAF","Chromatin","Complex","\/","1","9","6","7","0","8","5","-","Accessibility","CHRAC","dACF","<E>/FBgn0027620","<E>","dCHRAC"};



// FST terminal states: 0=non terminal; 1=terminal

static const char terminal_state[NUMBER_OF_STATES+1]={'\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\0','\1','\0','\0','\0'};



// FST transitions: each transition is a 3-tuple (current state,symbol,outgoing state)

static const int transitions[NUMBER_OF_TRANSITIONS*3]={

1,0,2,1,1,3,1,2,4,1,3,5,1,4,6,1,5,7,1,6,7,1,7,8,1,19,9,1,20,7,1,23,7,

2,10,13,

3,10,7,3,17,21,3,21,22,

4,10,10,

5,10,7,5,21,22,

6,10,7,

7,21,22,

8,18,19,
```

```
9,17,23,9,21,22,

10,11,11,

11,12,12,

12,12,7,

13,13,14,

14,14,15,

15,9,16,

16,0,17,

17,10,18,

18,15,20,

19,8,7,

20,16,7,

21,10,7,

23,10,24,

24,13,25,

25,16,7

};
```

El transductor de estado-finito obtenido es capaz de generar y reconocer 15 denominaciones del gen *Acf1*, que pertenecen a la misma clase se equivalencia, caracterizada por un identificador único, definido como la forma estandardizada:

Acf=> FBgn0027620
ACF=> FBgn0027620
ATP=> FBgn0027620
CAF=> FBgn0027620
CHRAC=> FBgn0027620
dACF=> FBgn0027620
dCHRAC=> FBgn0027620
acf 1=> FBgn0027620
ACF 1=> FBgn0027620
Acf 1=> FBgn0027620
Acf - 1=> FBgn0027620

Chromatin Accessibility Complex=> FBgn0027620
CG 1 9 6 6=> FBgn0027620
CHRAC - 1 7 5=> FBgn0027620
p 1 7 0 / p 1 8 5=>FBgn0027620

4. Aplicación de los transductores gráficos.

La identificación y normalización de los nombres de los genes se ha verificado en resúmenes obtenidos de la base de datos MEDLINE a partir de una consulta con el nombre del gen *Acf1* en el campo *Abstract* (AB) en la que se ha obtenido una pequeña muestra de 20 registros (Tabla 2). Esta verificación se podría haber realizado igualmente en el texto completo de los artículos.

Tabla 2 – *Abstract* de la base de datos MEDLINE

Binding of Acf1 to DNA involves a WAC motif and is important for ACF mediated chromatin assembly.
Fyodorov,-D-V; Kadonaga,-J-T
Mol-Cell-Biol. 2002 Sep; 22(18): 6344-53
ACF is a chromatin-remodeling complex that catalyzes the ATP-dependent assembly of periodic nucleosome arrays. This reaction utilizes the energy of ATP hydrolysis by ISWI, the smaller of the two subunits of ACF. Acf1, the large subunit of ACF, is essential for the full activity of the complex. We performed a systematic mutational analysis of Acf1 to elucidate the functions of specific subregions of the protein. These studies revealed DNA- and ISWI-binding regions that are important for the chromatin assembly and ATPase activities of ACF. The DNA-binding region of Acf1 includes a WAC motif, which is necessary for the efficient binding of ACF complex to DNA. The interaction of Acf1 with ISWI requires a DDT domain, which has been found in a variety of transcription and chromatin-remodeling factors. Chromatin assembly by ACF is also impaired upon mutation of an acidic region in Acf1, which may interact with histones during the deposition process. Lastly, we observed modest chromatin assembly defects on mutation of other conserved sequence motifs. Thus, Acf1 facilitates chromatin assembly via an N-terminal DNA-binding region with a WAC motif, a central ISWI-binding segment with a DDT domain, and a C-terminal region with an acidic stretch, a WAKZ motif, PHD fingers, and bromodomain

En los 20 registros encontramos 5 formas distintas del nombre de gen *Acf1* (*ACF*, *ATP*, *acf1*, *CHRAC*, *ACF1*) que se asocian con el identificador único *FBgn0027620* por medio de la aplicación del transductor gráfico desarrollado. Los resultados se muestran en la Tabla 3.

Tabla 3 – Resultado de la aplicación del transductor gráfico

Binding les Acf1 to DNA involves a WAC motif and les important for ACF mediated chromatin assembly.
Fyodorov,-D-V; Kadonaga,-J-T
Mol-Cell-Biol. 2002 Sep; 22(18): 6344-53
FBgn0027620 is a chromatin-remodeling complex that catalyzes the FBgn0027620-dependent assembly of periodic nucleosome arrays. This reaction utilizes the energy of FBgn0027620 hydrolysis by ISWI, the smaller of the two subunits of FBgn0027620. FBgn0027620, the large subunit of FBgn0027620, is essential for the full activity of the complex. We performed a systematic mutational analysis of FBgn0027620 to elucidate the functions of specific subregions of the protein. These studies revealed DNA- and ISWI-binding regions that are important for the chromatin assembly and ATPase activities of FBgn0027620. The DNA-binding region of FBgn0027620 includes a WAC motif, which is necessary for the efficient binding of FBgn0027620 complex to DNA. The interaction of FBgn0027620 with ISWI requires a DDT domain, which has been found in a variety of transcription and chromatin-remodeling factors. Chromatin assembly by FBgn0027620 is also impaired upon mutation of an acidic region in FBgn0027620, which may interact with histones during the deposition process. Lastly, we observed modest chromatin assembly defects on mutation of other conserved sequence motifs. Thus, FBgn0027620 facilitates chromatin assembly via an N-terminal DNA-binding region with a WAC motif, a central ISWI-binding segment with a DDT domain, and a C-terminal region with an acidic stretch, a WAKZ motif, PHD fingers, and bromodomain.

5. Conclusiones.

Debido a la falta de denominaciones oficiales para los genes, el desarrollo de sistemas que identifiquen este tipo de cadenas es decisivo en el campo de la biología molecular y genómica, fundamentalmente por la necesidad de descubrir de forma automática nexos entre genes dentro de la literatura biomédica. Las técnicas de extracción de información son procedimientos imprescindibles para llevar a cabo este proceso. Un problema relacionado con el reconocimiento y extracción de información lo constituye la normalización de las diferentes denominaciones de los genes, sin embargo los trabajos dedicados a este propósito son escasos. La unificación de los nombres de los genes requiere la identificación de las entidades genómicas y su vinculación a una forma controlada.

El proceso de normalización que nosotros proponemos se ha basado en el establecimiento de relaciones de equivalencia y en la aplicación de transductores de estado-finito. Nuestro método, frente a otros trabajos previos, no precisa de etiquetadores, que son herramientas muy costosas de desarrollar para este tipo de secuencias, ni de diccionarios que exigirían una continua actualización. La

normalización por medio de transductores requiere como *input* una lista de sinónimos y como *output* un identificador único del gen. La base de datos genómica FlyBase nos ha proporcionado los recursos para poder realizar esta aplicación, que ha sido probada en resúmenes de la base de datos MEDLINE. En este trabajo hemos presentado un procedimiento descriptivo que suministra un modelo teórico para unificar las denominaciones de los genes de una forma sistemática. En el futuro pensamos evaluar la eficacia de este sistema para demostrar el auténtico alcance de nuestra propuesta.

Referencias.

Crim, J., McDonald, R. & Pereira, F. (2005). Automatically Annotating Documents With Normalized Gene Lists. *BMC Bioinformatics, 6*(1), 13-19.

Cunningham, H. (2005), Information Extraction, Automatic. *Enclyclopedia of Language and Linguistics*. Oxford: Elsevier.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GENIES: a Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics, 17*(1), 74-82.

Hatzivassiloglou, V., Duboue, P. A. & Rzhetsky, A. (2001). Disambiguating Proteins, Genes, and RNA in Text: a Machine Learning Approach. *Bioinformatics, 17*, 97-106.

Hirschman, L., Park, C., Tsujii, J., Wong, L. & Wu, C. H. (2002). Accomplishments and Challenges in Literature Data Mining for Biology. *Bioninformatics, 18*(12), 1553-1561.

Hopcroft, J. E. & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Liu, H.,Johnson, S. B. & Friedman, C. (2002). Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. *Journal of the American Medical Informatics Association Online, 9*, 621-636.

Liu, H., Lussier, Y. A. & Friedman, C. (2001). Disambiguating Ambiguos Biomedical Terms in Biomedical Narrative Text: an Unsupervised Method. *Jorunal of Biomedical Informatics, 34*, 249-261.

Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S. & Colombe, J. B. (2004). Gene Name Identification and Normalization Using a Model Organism Database. *Journal of Biomedical Informatics, 37*, 396-410.

Ng, S., Wong, M. (1999). Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. In *Proceedings of Genome Informatics*, 104-112.

Pearson, H. (2001). La Catarata de nuevos genes pone en evidencia la anarquía de sus nombres . *El País (España)*. Disponible en: <http://www.elpais.es/suplementos/futuro/20010711/24genes.html>.

Proux, D., Rechenmann, F. & Julliard, L. (1998). Detecting Gene Symbols and Names in Biological Texts: a First Step Toward Pertinent Information Extraction. In *Proceedings of Genome Informatics*, 78-80.

Roche, E. & Schabes, Y. (1995). Deterministic Part-Of-Speech Tagging With Finite State Transducers. *Computational Linguistics, 21*(2), 227-253.

Schijvenaars, B. J., Mons, B., Weeber, M., Shuemie, M. J., Van Mulligen, E. M.,Wain, H. M. & Kors, J. A. (2005). Thesaurus-Based Disambiguation of Gene Symbols. *BMC Bioinformatics, 6*(1), 149.

Seki, K., Mostafa, J. (2005). A Hybrid Approach to Protein Name Identification in Biomedical Texts. *Information Processing & Management, 41*(4), 723-743.

Silberztein, M. (2000). INTEX: an FST toolbox. *Theoretical Computer Science, 231*,33–46.

Smaglik, P. (1998). Creativity, Confusion for Genes. *The Scientist, 12*(7), 1. Disponible en: <http://www.the-scientist.com/article/display/17971/>.

Thomas, J., Milward, D., Ouzounis, Pulman, S. & Carroll, M. (2000). Automatic Extraction of Protein Interactions from Scientific Abstracts. In *Proceedings of the Pacific Symposium on Biocomputing*, 538-549.

Tuason, O., Chen, L., Liu, H., Blake, J. & Friedman, C. (2004). Biological Nomenclatures: a Source of Lexical Knowledge and Ambiguity. In *Proceedings of the Pacific Symposium on Biocomputing*, 238-249.