

**Actas da 1ª Conferência Ibérica de Sistemas e Tecnologias de Informação
Ofir, Portugal, 21 a 23 de Junho de 2006
ISBN 978-989-20-0271-2**

Editores:

Maria Manuela Cunha,
Instituto Politécnico do Cávado e do Ave
Tel. 351 253 802 205, Fax 351 253 802 269, Email mcunha@ipca.pt
Álvaro Rocha
Universidade Fernando Pessoa
Tel 351 225074630, Fax 225074637, Email amrocha@ufp.pt

Capa/Composição:

Cláudio Ferreira
João Carlos Silva
Paulo Teixeira
Ricardo João Simões

Comissão Coordenadora:

Álvaro Rocha, Universidade Fernando Pessoa, Portugal
Carlos Ferrás Sexto, Universidad de Santiago de Compostela, España
Manuel Pérez Cota, Universidad de Vigo, España
Maria Manuela Cunha, Instituto Politécnico do Cávado e do Ave, Portugal

Comissão Organizadora:

António José Tavares
Bruno da Conceição Cortes
Eva Oliveira
João Carlos Silva
Joaquim Pereira Silva
Luís Ferreira
Maria Manuela Cunha
Patrícia Isabel Leite
Patrícia Sofia Gonçalves
Paulo Adriano Teixeira
Ricardo João Simões

Escola Superior de Tecnologia
Instituto Politécnico do Cávado e do Ave
Barcelos – Portugal

Volume 1

Modelos organizacionais

Informática para a saúde

Comissão Científica Portuguesa

Álvaro Rocha, *Universidade Fernando Pessoa*
Adriano Moreira, *Universidade do Minho*
Alberto Sampaio, *ISEP, Instituto Politécnico do Porto*
Alcínia Sampaio, *IST, Universidade Técnica de Lisboa*
Altamiro da Costa Pereira, *Fac. de Medicina da Univ. Porto, Pres. Ass. Port. Informática Médica*
Américo Afonso, *Fac. de Medicina Dentária da Un. do Porto*
Américo Azevedo, *FEUP, Universidade do Porto*
Anabela Mesquita, *ISCAP, Instituto Politécnico do Porto*
André Ribeiro, *Universidade de Calgary (Canada)*
António Godinho, *Inst. Sup. de Línguas e Administração (Gaia)*
António Lucas Soares, *Fac. de Engenharia da UP e INESC Porto*
António Serrano, *Universidade de Évora, Presidente do CA do Hospital de Évora*
Armando Jorge Sousa, *FEUP, Universidade do Porto*
Ascensão Braga, *Instituto Politécnico da Guarda*
Carlos Costa, *Universidade de Aveiro*
Carlos Ferreira, *Universidade de Aveiro*
Cláudia Viana, *Instituto Politécnico do Cávado e do Ave*
David Lamas, *Universidade Jean Piaget de Cabo Verde*
Eurico Carrapatoso, *Fac. de Engenharia da UP e INESC Porto*
Feliz Ribeiro Gouveia, *Universidade Fernando Pessoa*
Fernando Bandeira, *Universidade Fernando Pessoa*
Fernando Moreira, *Universidade Portucalense*
Filipe Pinto, *Instituto Politécnico de Leiria*
Filipe Sousa, *INESC Porto*
Filomena Lopes, *Universidade Portucalense*
Flávio Ferreira, *ESEIG, Instituto Politécnico do Porto*
Guilhermina Miranda, *FPCE, Universidade de Lisboa*
Goran Putnik, *Universidade do Minho*
Henrique Santos, *Universidade do Minho*
Isabel Braga Sampaio, *ISEP, Instituto Politécnico do Porto*
Isabel Ramos, *Universidade do Minho*
Isaiás Barreto da Rosa, *Universidade Jean Piaget de Cabo Verde*
Isaura Ribeiro, *Universidade dos Açores*
Ivo Antão, *Administrador da Espírito Santo Saúde*
Ivo Dias de Sousa, *Universidade Aberta*
João Álvaro Carvalho, *Universidade do Minho, Presidente da APSI*
João Barroso, *Universidade de Trás-os-Montes e Alto-Douro*
João Manuel Brisson Lopes, *IST, Universidade Técnica de Lisboa*
João Pedro Couto, *Universidade dos Açores*
João Sarmiento, *Universidade do Minho*

Sistemas e Tecnologias de Informação no Espaço Ibérico

João Manuel R. S. Tavares, *FEUP, Universidade do Porto*
João Varajão, *Universidade de Trás-os-Montes e Alto-Douro*
Joaquim Madeira, *Universidade de Aveiro*
Joaquim Reis, *Inst. Sup. Ciências do Trabalho e da Empresa*
Joaquim Sousa Pinto, *Universidade de Aveiro*
Jorge Batista, *Administrador da Primavera Software*
Jorge Mamede, *Instituto Politécnico do Porto e INESC Porto*
José Adriano Pires, *ESTIG, Instituto Politécnico de Bragança*
José Braga de Vasconcelos, *Universidade Fernando Pessoa*
José Carlos Nascimento, *Universidade do Minho*
José Manuel Oliveira, *Fac. de Economia da UP e INESC Porto*
José Rascão, *Instituto Politécnico de Setúbal*
José Ribeiro, *Universidade Aberta*
José Torres, *Universidade Fernando Pessoa*
Leonel Santos, *Universidade do Minho*
Leonilde Reis, *Instituto Politécnico de Setúbal*
Lídia Oliveira Silva, *Universidade de Aveiro*
Luciana Ferreira, *Universidade de Aveiro*
Luís Amaral, *Universidade do Minho*
Luís Bernardo, *FCT, Universidade Nova de Lisboa*
Luís Borges Gouveia, *Universidade Fernando Pessoa*
Luís Carlos Bruno, *Instituto Politécnico de Beja*
Luís Paulo Reis, *FEUP, Universidade do Porto*
Luís Velez Lapão, *Coordenador de Sistemas de Saúde, INA - Instituto Nacional de Administração*
Luís Vieira, *ISEG, Universidade Técnica de Lisboa*
Manuel Filipe Santos, *Universidade do Minho*
Manuel João Pereira, *Universidade Católica, Vice-Presidente do INA*
Manuela Natário, *Instituto Politécnico da Guarda*
Maria João Castro, *ISCAP, Instituto Politécnico do Porto*
Maria João Ferreira, *Universidade Portucalense*
Maria José Fernandes, *Instituto Politécnico do Cávado e do Ave*
Maria Manuela Cunha, *Instituto Politécnico do Cávado e do Ave*
Maria do Rosário Oliveira Martins, *ISEGI, Universidade Nova de Lisboa*
Maribel Santos, *Universidade do Minho*
Mário Caldeira, *ISEG, Universidade Técnica de Lisboa*
Mário David, *LIP Lisboa*
Mário Freire, *Universidade da Beira Interior*
Mário Rui Gomes, *IST, Universidade Técnica de Lisboa*
Miguel Mira da Silva, *IST, Universidade Técnica de Lisboa*
Nuno Lau, *Universidade de Aveiro*
Nuno Magalhães Ribeiro, *Universidade Fernando Pessoa*
Paula Morais, *Universidade Portucalense*

Paulo Costa, *FEUP, Universidade do Porto*
Paulo Gomes, *Universidade de Coimbra*
Paulo Gomes, *Instituto Politécnico de Portalegre*
Paulo Martins, *Universidade de Trás-os-Montes e Alto-Douro*
Paulo Pinto, *FCT, Universidade Nova de Lisboa*
Paulo Rupino, *Universidade de Coimbra*
Paulo dos Santos Silva, *Universidade Jean Piaget de Cabo Verde*
Pedro Antunes, *FC, Universidade de Lisboa*
Pedro Anunciação, *Instituto Politécnico de Setúbal*
Pedro Araújo, *Universidade da Beira Interior*
Pedro Nunes, *Instituto Politécnico do Cávado e do Ave*
Pedro Rangel Henriques, *Universidade do Minho*
Pedro Sobral, *Universidade Fernando Pessoa*
Ramiro Gonçalves, *Universidade de Trás-os-Montes e Alto-Douro*
Ricardo Machado, *Universidade do Minho*
Ricardo Morla, *INESC Porto*
Ricardo Simões, *Instituto Politécnico do Cávado e do Ave*
Rui Dinis Sousa, *Universidade do Minho*
Rui Gomes, *Instituto Politécnico de Viana do Castelo*
Rui Moreira, *Universidade Fernando Pessoa*
Sónia Sousa, *Universidade Jean Piaget de Cabo Verde*
Thomas Panagopoulos, *Universidade do Algarve*
Vítor Santos, *Microsoft Portugal*

Comissão Científica Espanhola

Carlos Ferrás Sexto, *Universidad de Santiago de Compostela*
Antonio Guevara Plaza, *Universidad de Málaga*
Antonio Moreno Jiménez, *Universidad Autónoma de Madrid*
Antonio Rodríguez del Corral, *Director de Desarrollo de la Operadora R*
Antonio Vaquero Sánchez, *Universidad de Complutense de Madrid*
Arturo Mendez Penín, *Universidad de Vigo*
Artur Serra i Hurtado, *Universidad Politécnica de Cataluña*
Belén Prieto, *Universidad de Oviedo*
Bernardo Bernal González, *Centro de Innovación y Tecnologías Avanzadas*
Carlos Macía Arce, *Universidad de Santiago de Compostela*
Carmen Fernández Morante, *Universidad de Santiago de Compostela (Lugo)*
Eduardo Paz Lloveras, *Comercio Electrónico Global, Presidente de la AEDED*
Emili Girault i Guarro, *Generalitat de Cataluña-Innovad*
F. Xosé Armas Quintá, *Universidad de Santiago de Compostela*
Fernando Garrido, *Observatorio Cibersociedad*
Francisco Carreras Riudavets, *Universidad de Las Palmas de Gran Canaria*
Francisco Sanchis Marco, *Universidad Politécnica de Madrid*
Horacio Capel Sáez, *Universidad de Barcelona*
Isabel Ramos Ramón, *Universidad de Sevilla*
Jacinto González Dacosta, *Universidad de Vigo*
Javier Franco Tubío, *Universidad de Santiago de Compostela, Secretario Ejecutivo CRUE-TIC, Miembro de la Junta de Gobierno ISOC-ES*
Javier García Tobío, *Director CESGA, Centro de Supercomputación de Galicia*
Jesús González Pérez, *Universidad de les Illes Balears*
Joana María Seguí Pons, *Universidad de les Illes Balears*
Joaquín Pinto Escribano, *Director CITA (Madrid), Centro de Innovación y Tecnologías Avanzadas*
Jörg Thomaschewski, *Fachhochschule Oldenburg, Ostfriesland/Wilhemshaven*
José Ayude Vázquez, *Universidad de Vigo*
José Baltazar García Pérez-Shofield, *Universidad de Vigo*
José Bravo, *Universidad de Castilla-La Mancha*
José Carlos Millán Calenti, *Universidad da Coruña*
José L.García Cuesta, *Universidad de Valladolid*
José Luis Gurria Gascón, *Universidad de Extremadura*
José Manuel Molina, *Universidad Carlos III de Madrid*
José María Cavanillas, *Atos Origin, Director - Atos Research & Innovation, Vice-Chairman - NESSI Steering Committee*
José María Feria Toribio, *Universidad Pablo Olavide (Sevilla)*
José Martín Brocos Fernández, *Universidad San Pablo-CEU, Madrid*
José Pérez Aguiar, *Universidad de Las Palmas de Gran Canaria*
Josefina Domínguez Mujica, *Universidad de Las Palmas de Gran Canaria*
Juan Manuel Cueva Lovelle, *Universidad de Oviedo*

Juan Manuel López Zafra, *Universidad Complutense de Madrid*
Leopoldo Seijas Candelas, *Universidad San Pablo-CEU, Madrid*
Luciano Boquete Vázquez, *Universidad de Alcalá*
Manuel J. Maña López, *Universidad de Huelva*
Manuel Ortega Cantero, *Universidad de Castilla-La Mancha, Presidente de ADIE*
Manuel Pérez Cota, *Universidad de Vigo*
Manuel Rodríguez Victoriano, *Universidad de Valencia*
María Lado José Touriño, *Universidad de Vigo*
María Rosario González Rodríguez, *Universidad de Sevilla*
Mercedes Ruiz Carreira, *Universidad de Cádiz*
Montserrat Sebastià i Salat, *Universidad de Barcelona*
Octavio Santana Suarez, *Universidad de Las Palmas de Gran Canaria*
Pedro Pérez Celis, *Universidad Pontificia Comillas*
Pedro Requés Velasco, *Universidad de Cantabria*
Rafael Crecente-Maseda, *Universidad de Santiago de Compostela (Lugo)*
Sergio Gálvez Rojas, *Universidad de Málaga*
Vicente Rodríguez Rodríguez, *Consejo Superior de Investigaciones Científicas, IEG (Madrid)*
Xosé López García, *Universidad de Santiago de Compostela*
Yolanda García Vázquez, *Universidad de Santiago de Compostela*

Prefácio

Sentimo-nos particularmente honrados, na Microsoft Portugal, pela oportunidade de apoiar uma iniciativa como a 1ª Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI).

Sendo a Microsoft uma empresa que, de há 30 anos a esta parte, tem vindo a trabalhar na concepção e na democratização da Inovação, traduzindo-a em tecnologia que impacte positivamente a vida dos utilizadores, quer em contextos pessoais quer profissionais, é por demais evidente que o tema desta 1ª Conferência nos diz muito.

O interesse e a vitalidade que a Comunidade Académica Portuguesa demonstrou possuir com a organização desta 1ª CISTI constitui um sinal claro do potencial da contribuição que os nossos académicos podem oferecer para resolvermos com sucesso um dos maiores desafios que se colocam a Portugal: o desafio da Inovação!

Ultimamente temos ouvido falar muito em Inovação. No entanto, nunca é demais sublinhar a importância preponderante que a diferenciação qualitativa, inaugurada por esta, poderá trazer para Portugal, em diversos sectores de actividade, desde os sectores da economia mais tradicional, aos serviços mais contemporâneos.

A Inovação em Portugal terá que acontecer em diversas áreas – a começar pelas mentalidades e formas de fazer – e é aqui que os Sistemas de Informação e a indústria de Software em particular, podem dar um contributo nada desprezável. Um país que não inova tecnologicamente é um país morto, porque será facilmente asfixiado nos palcos da feroz competição global. E, de entre todas as formas simbióticas que a inovação tecnológica pode assumir, uma há que gostaria de salientar, porque pode ser facilmente transposta para outras indústrias: a importância da valorização da propriedade intelectual, que é o mesmo que dizer: da assinatura, da marca.

A Tecnologia é um sector habituado a lidar com a frequência da obsolescência e por isso o respeito pela propriedade intelectual de quem inova é chave para o reconhecimento e muitas vezes para a sobrevivência comercial dos autores da Inovação. Por esta razão, todas as iniciativas que se destinem a passar a mensagem de que Portugal precisa de mais autoria só podem ter da Microsoft todo o apoio.

Estamos firmes na convicção de que as conclusões desta 1ª Conferência constituem mais um passo para o desenvolvimento da investigação em Sistemas de Informação em Portugal.

Lisboa, 30 de Maio de 2006
Nuno Duarte

Índice

Nota Introdutória.....	1
Post-Project Effort Analysis Method	3
Medidas de Acompanhamento e Controle Aplicadas ao Desenvolvimento de Software	21
Um sistema de apoio à decisão para projecto e gestão de rega por gravidade, baseado numa aplicação cliente-servidor	37
Sistema Integral para la Formación Online (SIFO): Aplicación al ámbito universitario	47
Aplicação das teorias de Bloom e Kolb no estudo da análise de sistemas	63
Um Sistema de Informação Académico para a Universidade Jean Piaget de Cabo Verde	81
Factores Determinantes na Adopção da Internet em Portugal.....	97
Uma Abordagem sobre a Utilização de Tecnologia da Informação nas Empresas de Serviços Contábeis em Recife.....	115
Sociedade da Informação na Região de Alto Trás-os-Montes	127
Barriers to FLOSS in SMEs: the lack of knowledge and skills.....	143
GASP: Definición de una Plataforma para la Gestión y el Análisis Estadístico de Proyectos Software para PYMES y Emprendedores.....	157
Emerging Technologies and Standards on Business Process Management in Collaborative Networks Environments	167
Proposta de uma plataforma de integração para a administração pública	179
Aplicación del Modelado Workflow a la Reingeniería de Sistemas de Información Basándose en Interfaces de Usuario.	195
Modelo Explicativo das Iniciativas de Comércio Electrónico em Organizações Portuguesas	211
Integração de Empresas Virtuais: Um caso de aplicação á Indústria Têxtil.....	221
Qual o Futuro do Teletrabalho? Um Estudo Prospectivo com base no Método dos Cenários e no Método Delphi	239
Monitorização remota do desempenho da empresa.....	257

Metodologia de gerenciamento de projetos aplicado à produção de conteúdo digital para e-learning	265
Novas Tecnologias para uma Nova Aprendizagem nas Instituições de Ensino Superior	281
Um Sistema de Informação Criativo baseado na técnica de criatividade “whiteboard”	297
La Formación a Distancia y las Organizaciones Portuguesas	305
Information Web Audit. Estudio de caso comparado entre Galicia e Irlanda	319
La eficacia del formato y la animación en la publicidad en Internet ¿Se ha desgastado el banner?	333
Las tecnologías web 2.0 y los sistemas de información automatizados de los SGE.	351
Solución Metodológica Genérica para la Evaluación de la Calidad de Sitios Web	365
Governo Eletrónico: uma avaliação dos <i>websites</i> das Secretarias de Receita Estaduais da Fazenda da região Nordeste do Brasil	383
Avaliação de Soluções de <i>e-Urbanismo</i>	401
O Governo Electrónico e os Sistemas de Informação Públicos em Portugal	421
Estratégias de <i>e-procurement</i> na Administração Pública: Uma revisão de literatura	439
An innovative approach in supporting the operation of complex equipment machinery: the KoBaS Project case	455
Eficiência na Construção de Equipas Colaborativas <i>Online</i>	471
Urbanismo organizacional – Instrumento para a gestão num contexto relacional. O Caso Internet Banking	489
Activos Intangíveis dos Sistemas de Informação	507
Sistema Dispensador de Píldoras Controlado por Telefonía Móvil.	523
TeleGerontología: Un nuevo recurso de Apoyo Gerontológico A Domicilio	533
Valoración de la calidad e-Health: El modelo @racne	547
Informática Dentária: Sistemas de informação de suporte à prática clínica	557
Intelligent Mechanisms for Early Detection of Cervix Cancer	571
Uma Abordagem à Gestão do Risco para os Sistemas de Informação das Unidades de Saúde	589

Sistemas e Tecnologias de Informação no Espaço Ibérico

MAID – Multi Agent for the Integration of Data.....	603
Descoberta de Padrões de Proteínas Utilizando um Algoritmo Genético	615
Web.care – Gestão de acessos e recursos para estudos clínicos multicêntricos on-line	631
O <i>Data Mining</i> na Compreensão do Fenómeno da Dor: Uma Proposta de Aplicação	641
Implementação de um Sistema de Monitorização de Sistemas de Informação Clínicos indirecto	657
Sistemas de Saúde e processos empresariais.....	671
Automatic detection of patient data inconsistencies on integrated Health Information Systems	689
Extracción y normalización de entidades genómicas en textos biomédicos: una propuesta basada en transductores gráficos.....	697
Caso de Estudo: Modelação e Verificação Formal de um Serviço de Urgência Hospitalar	711
O Tratamento de Dados na Saúde: Do enquadramento Jurídico à utilização de Modelos de Gestão.....	727
Control y Mejora en la Producción de Software Hospitalario, Nuevas Tecnologías	741
Estruturação automática de informação de relatórios médicos: Um estudo de caso	753

Nota Introdutória

No actual contexto económico e organizacional, caracterizado por uma crescente importância da informação e do conhecimento, os Sistemas e Tecnologias de Informação (STI) desempenham um papel cada vez mais preponderante enquanto factor de competitividade das organizações e sociedades modernas, facilitando a obtenção, transformação, armazenamento e disponibilização da informação.

Não obstante as facilidades estarem ao dispor de todos, somente serão conseguidas se as novas tecnologias de informação forem devidamente exploradas e integradas nos sistemas de informação organizacionais e na sociedade em geral. Assim, a génese da CISTI assentou na necessidade sentida de estimular este desiderato e, simultânea e cumulativamente, na inexistência de um fórum anual que promovesse de forma abrangente a partilha e discussão de experiências, conhecimentos e inovações por pessoal das universidades, investigadores e profissionais de STI do espaço ibérico.

A CISTI é uma iniciativa promovida e coordenada pelo grupo disciplinar de Sistemas e Tecnologias de Informação da Escola Superior de Tecnologia do Instituto Politécnico do Cávado e do Ave, representado por Álvaro Rocha e Maria Manuela Cunha, e pelo GIMED (Grupo de I&D em Informática Médica) da Universidade Fernando Pessoa, representado por Álvaro Rocha. A iniciativa foi alargada ao Grupo de Investigación Socio-Territorial da Universidade de Santiago de Compostela, representado por Carlos Ferrás Sexto, e ao Departamento de Informática da Universidade de Vigo, representado por Manuel Pérez Cota.

Da Comissão Científica da CISTI faz parte um leque pluridisciplinar de peritos oriundos ou fortemente relacionados com a área dos STI do espaço ibérico, aos quais cabe a responsabilidade de avaliar, num processo de revisão cega, os trabalhos submetidos em cada uma das edições da conferência.

Tal como acontece nesta primeira edição, Junho será o mês preferencial para a realização anual da CISTI, uma vez que não pretendemos sobrepor-nos, colidir ou interferir com possíveis interesses de outros eventos, como, por exemplo, a Conferência da Associação Portuguesa de Sistemas de Informação (Outubro/Novembro) e as Jornadas Luso-Espanholas de Gestão Científica (Fevereiro/Março).

Anualmente serão envidados esforços para que os melhores artigos da CISTI possam integrar periódicos e livros ibéricos que abordem temáticas correlatas com os sistemas e tecnologias de informação. Serão igualmente envidados esforços para que todos os artigos publicados nas actas da CISTI venham a integrar bases de dados de publicações científicas, como, por exemplo, a DBLP. E serão ainda envidados esforços para o estabelecimento de um repositório *on-line* com todos os artigos igualmente publicados na CISTI.

Reportando-nos a esta primeira edição, foram recebidos cerca de 250 artigos, na sua maioria de elevadíssimo interesse e valor. Destes, cerca de 15% eram provenientes de fora da Península Ibérica (alguns países da Europa e América Latina). Contudo, mediante a impossibilidade de estender a duração da CISTI para além de dois dias e meio, houve necessidade de estabelecer em 40% o número de artigos a aceitar para apresentação na conferência e publicação nos respectivos livros de actas.

As presentes actas, constituídas por dois volumes, organizam-se em torno das grandes temáticas identificadas na chamada de trabalhos da CISTI. Assim, no Volume I encontram-se os artigos relacionados com Modelos Organizacionais e Sistemas de Informação e ainda Informática para a Saúde, e no Volume II encontram-se os artigos relacionados com Gestão de Conhecimento e Sistemas de Suporte à Decisão, Integração de Sistemas e Agentes de Computação e ainda Interação Homem-Máquina.

Mediante aquele que foi um desafio sério de criar um fórum ibérico nestes domínios da actividade científica e empresarial, que fez mover uma vasta equipa empenhada, estamos certos que os resultados justificaram plenamente o esforço e que desta forma o desafio foi feito vitória. Tudo aponta para que a CISTI venha a estar posicionada na rota dos eventos científicos de maior destaque no domínio dos STI.

Lançada a CISTI, resta-nos agradecer a colaboração de todos os que directa e indirectamente tiveram (ou venham a ter) algum envolvimento com este fórum, deixando aqui o desafio de juntos contribuirmos para que venha efectivamente a ser uma referência no espaço ibérico, cada vez mais uno.

Até breve!

Os Editores

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

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

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 estandarizado. 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: GBD² (*Genoma Humano*), FlyBase³ (*Drosophila melanogaster*), WormBase⁴ (*Caenorhabditis elegans*), o Mouse Genome Informatics⁵ (*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 Arthritis*, *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 estandarizada, 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*.

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

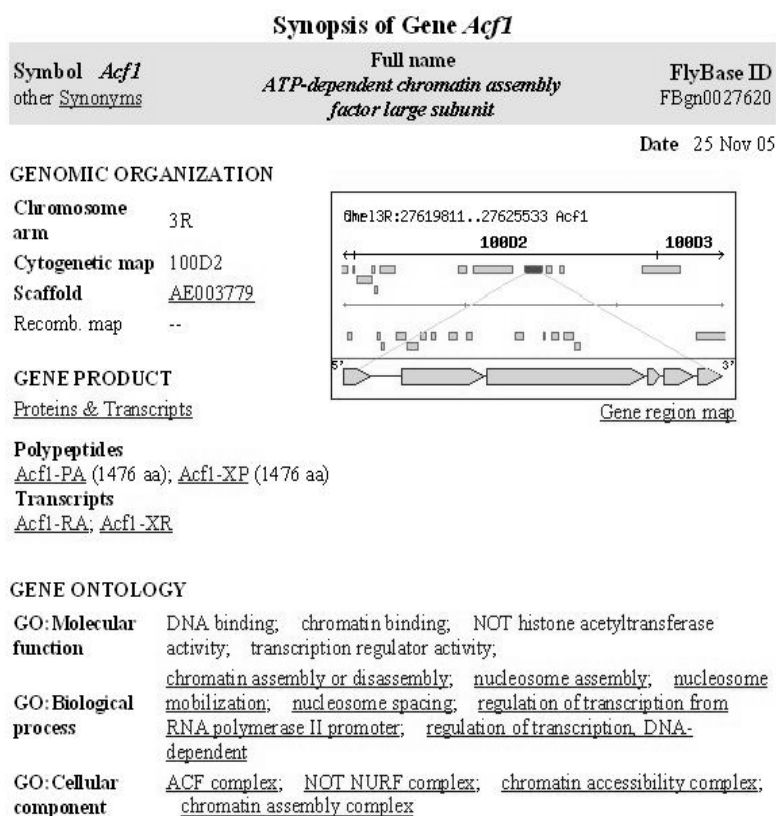
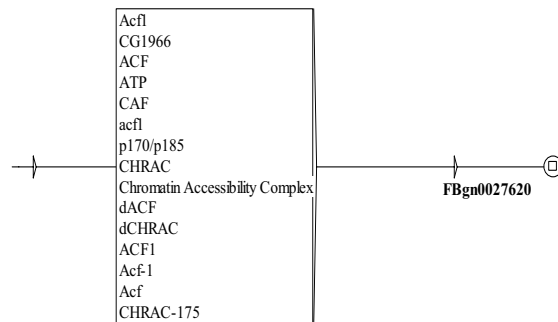


Figura 1 – Entrada en la base de datos FlyBase del gen *Acf1*.

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*.



Gene: Acf1.gcf
Mon Jan 30 11:42:08 2006

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, $\delta = 38$.


```
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 de equivalencia, caracterizada por un identificador único, definido como la forma estandarizada:

```
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
```

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

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

<p>Binding les Acf1 to DNA involves a WAC motif and les important for ACF mediated chromatin assembly.</p> <p>Fyodorov,-D-V; Kadonaga,-J-T</p> <p>Mol-Cell-Biol. 2002 Sep; 22(18): 6344-53</p> <p>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.</p>

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

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

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. *Encyclopedia 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. *Bioinformatics*, 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 Ambiguous Biomedical Terms in Biomedical Narrative Text: an Unsupervised Method. *Journal 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 Online 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.

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