Software

Open Access

Introducing meta-services for biomedical information extraction

Florian Leitner[1], Martin Krallinger[1], Carlos Rodriguez-Penagos[1], Jörg Hakenberg[2,3], Conrad Plake[2], Cheng-Ju Kuo[4,5], Chun-Nan Hsu[5], Richard Tzong-Han Tsai[6], Hsi-Chuan Hung[5], William W Lau[7], Calvin A Johnson[7], Rune Sætre[8], Kazuhiro Yoshida[8], Yan Hua Chen[9], Sun Kim[10], Soo-Yong Shin[10], Byoung-Tak Zhang[10], William A Baumgartner Jr[11], Lawrence Hunter[11], Barry Haddow[12], Michael Matthews[12], Xinglong Wang[12], Patrick Ruch[13], Frédéric Ehrler[14], Arzucan Özgür[15], Güneş Erkan[15], Dragomir R Radev[15], Michael Krauthammer[16], ThaiBinh Luong[17], Robert Hoffmann[18], Chris Sander[19] and Alfonso Valencia[1]

Addresses: [1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor F. Almagro 3, 28029 Madrid, Spain. [2]Bioinformatics group, Biotechnological Centre, Technical University Dresden, Tatzberg 47-51, 01307 Dresden, Germany. [3]Humboldt Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. [4]Institute of Bioinformatics, National Yang-Ming University, No. 155, Sec. 2, Linong St., Beitou District, Taipei City 112, Taiwan. [5]Institute of Information Science, Academia Sinica, No.128, Sec. 2, Academia Rd., Nangang District, Taipei City 115, Taiwan. [6]Department of Computer Science and Engineering, Yuan Ze University, 135 Yuan-Tung Rd., Chung-Li, Taoyuan, R.O.C., 32003, Taiwan. [7]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, USA. [8]Department of Computer Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, 113-0033 Tokyo, Japan. [9]Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælands vei 7-9, NO-7491 Trondheim, Norway. [10]Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea. [11]Center for Computational Pharmacology, University of Colorado School of Medicine, P.O. Box 6511, Mail Stop 8303, Aurora, CO 80045-0511, USA. [12]School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK. [13]Text Mining Group, Medical Informatics Service, University and Hospitals of Geneva, 24 Micheli du Crest, 1201 Geneva, Switzerland. [14]Artificial Intelligence Group, University of Geneva, 7 route de Drize, 1227 Carouge, Switzerland. [15]Department of Electrical Engineering and Computer Science, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA. [16]Department of Pathology, Yale University School of Medicine, 300 Cedar Street, TAC 309, New Haven, CT 06520-8023, USA. [17]Program for Computational Biology and Bioinformatics, Yale University, Suite 501, 300 George Street, New Haven, CT 06520-8084, USA. [18]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), The Stata Center Building 32, 32 Vassar Street, Cambridge, MA 02139, USA. [19]Computational Biology Center, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA.

Correspondence: Florian Leitner. Email: fleitner@cnio.es. Alfonso Valencia. Email: valencia@cnio.es

Published: 01 September 2008

Genome ***Biology*** 2008, **9(Suppl 2):**S6

doi: 10.1186/gb-2008-9-S2-S6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/S2/S6

© 2008 Leitner *et al*; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

We introduce the first meta-service for information extraction in molecular biology, the BioCreative MetaServer (BCMS; http://bcms.bioinfo.cnio.es/). This prototype platform is a joint effort of 13 research groups and provides automatically generated annotations for PubMed/Medline abstracts. Annotation types cover gene names, gene IDs, species, and protein-protein interactions. The annotations are distributed by the meta-server in both human and machine readable formats (HTML/XML). This service is intended to be used by

biomedical researchers and database annotators, and in biomedical language processing. The platform allows direct comparison, unified access, and result aggregation of the annotations.

Background

Information retrieval (IR), information extraction (IE), and text mining have become integral parts of computational biology over the past decade [1]. However, these services are dispersed, integrated in specific packages, and include proprietary software. Therefore, progress in the field requires offering better access to the tools, methods, and their results [2]. Other areas, such as sequence analysis, genome analysis, or protein structure prediction, have benefited greatly from enhanced access to services and tools for the community of biologists, bioinformaticians (through web servers and portals), and developers (by providing free, open source academic software) [3].

Web services, widely used throughout the internet to provide the functionality for distributed systems, are becoming a common part of bioinformatics tools; For example, one of the most used text mining applications, namely iHOP (Information Hyperlinked Over Proteins), provides such an infrastructure to access its data [4]. Meta-services, too, are a ubiquitous component of the world wide web, found as meta-search engines, in business-to-buisness and business-to-consumer transactions (for example, for flight booking systems), and are used in scientific research (for example, for protein structure prediction) [5]. Another example of a distributed meta-service is BioDAS (Distributed Annotation System), a platform to exchange biologic sequence annotations between independent resources [6].

This publication describes the development of the BioCreative MetaServer (BCMS) prototype. The Results section (below) provides an overview of the system design and introduces the basic components, followed by short descriptions of the IE systems currently available through the platform prototype. The Discussion section (below) reviews what problems are solved and what issues need further investigation. The Conclusions section (below) closes with current and future utilities of this platform for the biomedical community. Technical details on the platform and implementation aspects can be found in the Materials and methods section (below).

Results

The fundamental aim of the BCMS platform is to provide users with annotations on biomedical texts from different systems. At the current prototype level, the dataset is restricted to a fixed number of approximately 22,800 PubMed/Medline abstracts. The available annotations consist of marking passages that are detected as gene or protein name mentions, annotating the articles with the gene/protein and taxonomic IDs (providing hyperlinks to the corresponding database entries), and a confidence score for whether the text contains protein-protein interaction information. Expanding on stand alone IE systems, this platform gathers the results of several systems developed by various research groups, unifies them, and allows the user to access abstracts and annotations in a combined view. It is conceivable that collating classification results will often enhance performance, simply because multiple equal classifications for a given annotation are more likely to be correct. The gathered data are accessible to the user both as human-readable hypertext and as machine processable XML in the form of XML-RPC requests.

System design

The platform is to be regarded as a distributed system requesting, retrieving and unifying textual annotations, and delivering these data to the user at different levels of granularity. The BCMS can be divided into three main units.

• A static collection of text (a set of approximately 22,800 PubMed abstracts used in the BioCreative II challenge [7]).

• A set of active servers providing annotations for text (see Table 1 for participating servers) upon request; these annotation servers (AS) only interact with the meta-server and not directly with each user.

• A meta-server providing the combined data, namely both the annotations and the corresponding text. Therefore, users indirectly communicate with the annotation servers, using the meta-server as proxy.

For all communication purposes, the system utilizes the XML-RPC protocol [8]. The meta-server sends requests to annotate a PubMed/Medline abstract to all known annotation servers. Once the ASs have finished processing the text, the annotation data are returned to the meta-server, which stores all annotations in its central repository. Whenever a user requests annotations for an abstract, the meta-server checks whether the annotation data already exist. If not, then it triggers a remote procedure call (RPC) with the PubMed ID to the ASs; otherwise, the server immediately returns the stored results to the client (Figure 1). There are two principal ways to access the meta-server: via web browser or by using the XML-RPC web service. In the former case, the results are asynchronously returned to the user via AJAX, whereas in the latter case - the web service - the response is sent once all results have been gathered. The system is intended to work at a maximum latency of about 10 seconds, after which the annotation servers are expected to have returned their anno-

Table 1

Annotation servers

Team/Group	GM	GN	TX	PPI	Conf	State	Web Page
Hakenberg	+	+	+	+	True	Dynamic	http://alibaba.informatik.hu-berlin.de/bcms/
Kuo	+				False	Dynamic	http://aiia.iis.sinica.edu.tw/biocreative2.htm
Tsai	+			+	True	Dynamic	http://asqa.iis.sinica.edu.tw/biocreative2/
Lau		+			True	Dynamic	http://giant.cit.nih.gov/
Sætre	+	+		+	True	Static	
Kim				+	True	Dynamic	http://bi.snu.ac.kr/pie
Baumgartner	+	+		+	False	Dynamic	
Haddow	+	+	+	+	True	Static	
Ruch		+		+	True	Dynamic	http://129.194.97.165/EAGLtools/
Özgür				+	True	Static	
Luong		+			True	Dynamic	
Hoffmann	[a]	+	+	+	True	Dynamic	http://www.ihop-net.org/
Totals	6	8	3	9	10	12 teams	

List of annotation servers used by the meta-server, plus a Boolean flag determining whether the classifiers use a confidence score (Conf) and the system state (State): dynamic = online system, already capable of delivering content for any PubMed abstract; static = offline system, server in development. The webpage columns provide a link to an online site for a team's annotation system. [a]iHOP (Hoffmann) delivers GMs, but because of data compatibility issues this is a feature to be added in later versions of the meta-server. GM, gene/protein mention detections; GN, gene/protein normalizations; PPI, protein-protein interaction classification; TX, taxon classifications.

tation results to the meta-server. Obviously, these response times will increase under heavy load when many requests for non-annotated citations are made and will need constant monitoring. If the annotations for the PubMed ID have been generated already, the stored results are returned instantaneously for both (browser and RPC) scenarios.

The data can be provided by three different means, which also correlate with the three main components.

• Via web browser [9]. The main intention of this access method is to allow end-users (biomedical researchers) to search for a specific piece of information, e.g., to identify or confirm interaction partners for a given gene or protein. This view correlates with the meta-server unit (the third BCMS unit described above) and offers the user a graphical interface to explore the text and annotations (Figure 2).

• The second option is to use the XML-RPC protocol. This method is intended to provide developers with a means to integrate the platform data with their own applications, for example to use in combination with other annotation pipelines. Therefore, this is the direct interface to the ASs (the second BCMS unit described above), because the meta-server only acts as a proxy in this scenario. The API of the XML-RPC service can be found online [10].

• The third option is to contact the authors for a database snapshot of the current state of the meta-server data. This option is of interest for IE and text mining applications that make heavy use of the data, where online RPC would not be an option. This roughly correlates with the static content of the platform (the first BCMS unit described above). Because this is a rather crude access method, it might be improved (for example, a daily updated FTP download service) once the prototype stage is fully completed and enough interest is signaled.

Annotation systems

Annotating biomedical abstracts can be done at various levels of granularity. Currently, the service provides four types of annotations.

• Gene/protein mention (GM): locate positions in the text that are detected as gene or protein names.

• Gene/protein normalization (GN): detect which genes or proteins are mentioned, assigning sequence database identifiers to the text.

• Taxon classification: identification of the organisms to which the text pertains, together with a confidence score, providing an ID for the National Center for Biotechnology Information (NCBI) taxonomic database.

• Protein-protein interaction (PPI): classifies whether the text contains PPI information and assigns a confidence score to the classification.

GM and GN may also provide confidence scores, depending on the annotation system. All confidence scores are normalized to the 0, 1 range to render them directly comparable. For any given identical annotation between two or more annotation servers, the mean (to compensate for outliers) of the con-

fidence scores is calculated. If an AS returns no confidence scores, then the result is not accounted for in the calculation of the mean. Note that the annotation systems employed here result from a recent challenge evaluation of the state of the art for such text-mining tasks [11]. According to the evaluation, gene mentions can be recognized with an F measure of more than 87% [12]. The gene name normalization has been shown to yield a top performance of more than 81% [13]. Classifications of whether a text discusses one or more protein-protein interactions can reach F measures above 78% [14]. Currently, there are 12 teams providing annotations for the meta-server.

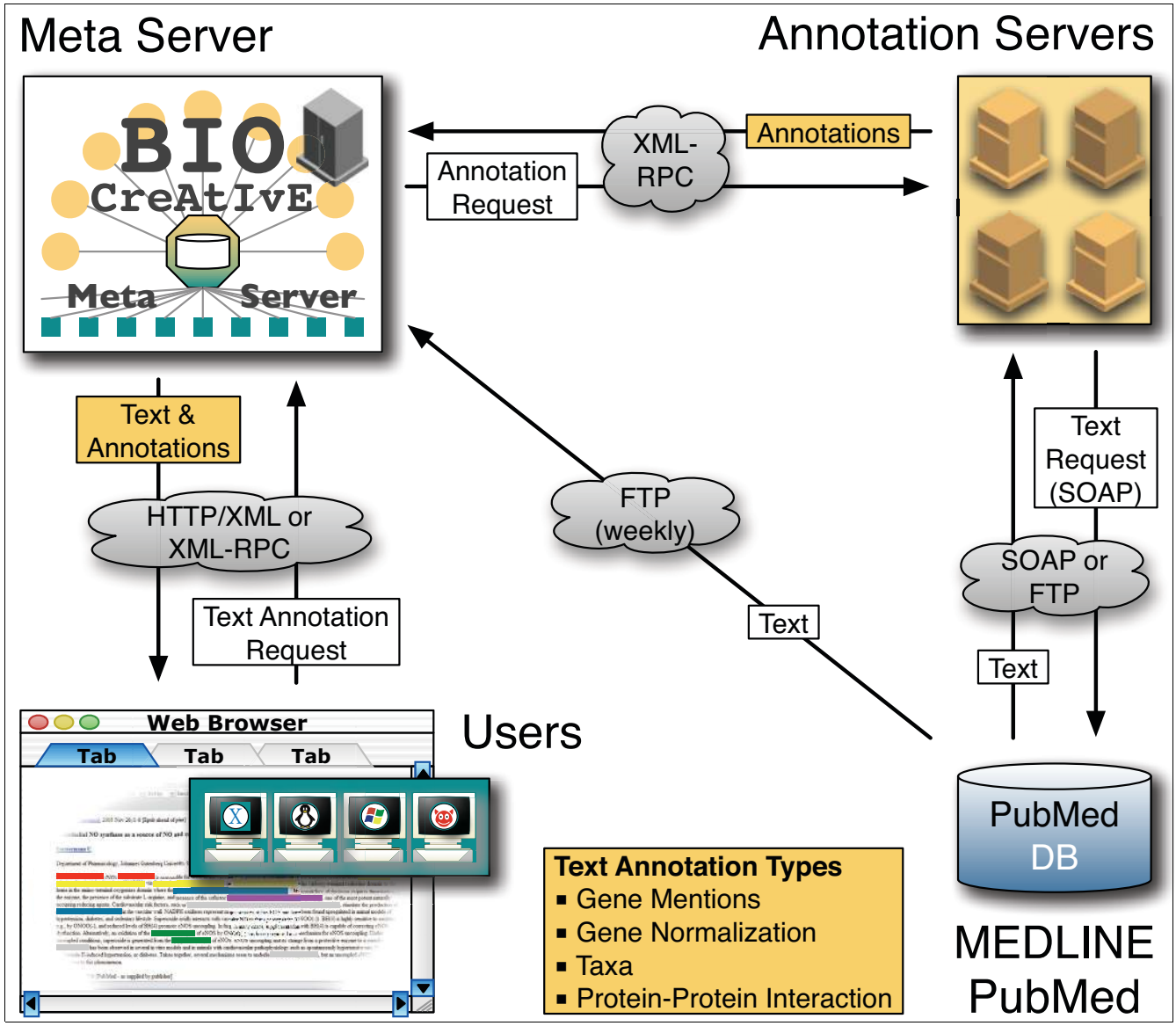

Figure 1
BioCreative MetaServer (BCMS) system design. The users contact the meta-server either via browser interface or remote procedure call (RPC), requesting annotations for a given PubMed ID. The meta-server checks whether the citation has already been annotated. If this is the case, then the data are immediately returned; otherwise the meta-server sends annotation requests to the annotation servers (AS), awaiting their response. Meanwhile, the Medline entry is presented to the user. Whenever an AS returns its results, the user web page is asynchronously updated to present the new information. Alternatively, if the user submitted a RPC request, then the request simply finishes when all ASs have returned their annotations and classifications. The ASs and the meta-server either use a local copy of Medline (downloaded via FTP) or fetch the relevant citation using service oriented architecture protocol (SOAP) and the National Center for Biotechnology Information (NCBI) eUtils.

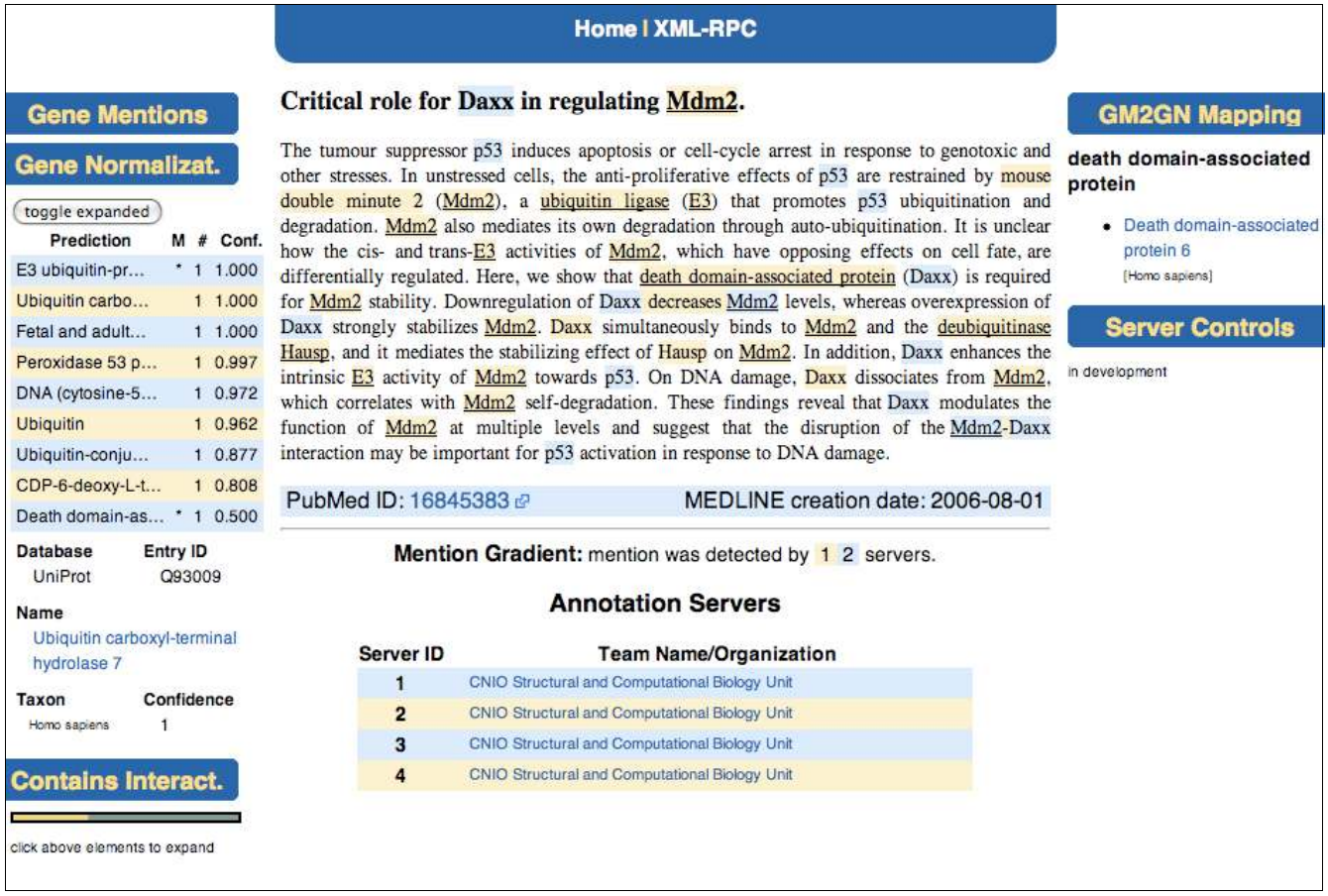

Figure 2
BioCreative MetaServer (BCMS) annotation view screenshot. This screenshot of the annotation view of the meta-server shows the main annotations for the given Medline abstracts (PMID 16458891). Central view: gene mentions (GMs) are marked in the text, ranging from gray (single annotation server [AS] detecting the particular mention) to yellow (all five ASs that have analyzed the text detect the highlighted text snippet as a GM), as a gradient that is shown below the text. At the bottom, the list of servers providing the annotations for this abstract can be found (only four of all thirteen visible). Left column: all raw annotation results can be viewed here. Gene mentions (GMs) results are expanded and sorted first by the number of servers predicting that mention and then by the median confidence for it. On the bottom left, a quick bar indicates protein-protein interaction (PPI) results. The bar is split in two, where the left and right bar lengths indicate the number of servers classifying this abstract as negative or positive in relation to mentioning PPIs. The color of the bars indicate the mean confidence of all classifications of one type: the negative (left) bar ranges from blue (low) to yellow (high confidence), and the positive (right) bar from yellow (low) to blue (high confidence). The bar also provides some interactivity: shortened names (indicated py an elipsis at the end) can be seen in their full form by mouse over, mousing over the gene mention highlights its position in the text, and individual gene normalization results can be clicked to see the exact database identifier, name, organism and a link to the DB record. Right column: by clicking on an italic mention in the central view, all possible mappings of GMs to GNs are shown: in bold the GM, and then the list of GNs (together with the species) and their official names (here for the text span "interferon-inducible p200 family"). This simple mapping is based on case-insensitive substring matching of GMs and the GN names and synonyms extracted from the DB records.

Here, a short overview of each system explains how results are generated (see Table 1 for an overview of the classifiers that each annotation server provides). The section titles (below) consist of the team locations and author initials in parenthesis, followed by the team name identifier used in Table 1 in square brackets.

Biotec TU Dresden and Humboldt-Universität zu Berlin (JH, CP) [Hakenberg]

The annotations we currently provide are gene mention normalization (32,795 human genes from EntrezGene), protein mention tagging (about 200,000 proteins from UniProt/SwissProt), NCBI taxonomy IDs for species mentioned in texts, and classifications of whether the text discussed one or more protein-protein interactions. Gene mention normalization was evaluated on the BioCreative II GM data and yields a precision of 79% at 83% recall [12]. Entity mention normalization is based on large lexicon of known names and synonyms, which are kept in main memory at all times for efficiency. Once a potential named entity has been found, we further identify it using context profiles in case multiple entities share the same name [15]; these profiles contain knowledge about each candidate entity, such as GO terms, chromosomal locations, or tissue specificity. We rank genes according to pieces of profiles also recognized in the current text. Annotations of proteins, species, and protein-protein

interactions are based on the Ali Baba system [16]. Protein names are not normalized to a single UniProt ID, but potentially multiple IDs are returned for polysemous names. Recognition of species is currently based on about 200,000 names from the NCBI taxonomy. Ali Baba matches consensus patterns identified by multiple sentence alignment to recognize relationships between entities in texts. We decided to split the annotation services into two different servers: one for gene mention normalization and one for the other tasks. Please refer to the BCMS and the website mentioned in Table 1 for more detailed information on how to contact the services.

Institute of Information Science Academia Sinica (CK and CH) [Kuo]
Kuo and coworkers' system [17] based on conditional random fields (CRFs) for gene mention tagging, is among the best performing systems in this challenge evaluation. The key features of the system include a rich feature set, unification of bidirectional parsing models, a dictionary-based filtering, postprocessing, and its high performance. We carefully selected several feature types for CRF tagging, including character *n*-grams (window size 2 to 4) and morphological as well as orthographic features. In addition, we picked up several domain specific features (for example, biochemical terms such as cDNA, mRNA, tyrosine, and so on). On the other hand, some more commonly used features, such as stop words, prefix, and suffix, were not labeled. We utilized -2 to 2 as the offsets to generate contextual predicates. Then, we trained both forward and backward parsing models and combined them to obtain the final tagging results. This release is different from the version that we used to produce runs for BioCreative II. We still used MALLET [18] to perform CRF training and testing and Genia Tagger [19] for POS tagging, but we rewrote the feature extractors in Java and optimized the implementation to enhance greatly the efficiency. We also tuned the feature set to remove redundancies and other minor issues in the original feature set. As a result, this release can achieve a slightly higher F score than the original version with better efficiency.

Institute of Information Science (RT and HH) [Tsai]
Our annotation system [20] supports GM recognition and PPI text classification. For named entity recognition, we employ CRF as the underlying machine learning (ML) model; a set of features selected by a sequential forward search algorithm; numerical normalization; and pattern-based postprocessing [21] to help ML-based GM to deal with extremely difficult cases that need longer context windows. For PPI, we use a support vector machine (SVM) with a novel feature representation scheme, contextual-bag-of-words [22], to exploit named entity information. We further improve the performance by extracting likely positive and likely negative data from unlabeled data to provide additional training data. The performance of our GM and PPI text classification system is in the first quartile of the BioCreative II GM task (see the review by Smith and coworkers [12] included in this supplement). Our services support high-throughput online data processing and can be accessed online (see Table 1) and as an XML-RPC service at [23].

Division of Computational Bioscience, CIT, NIH (WL and CJ) [Lau]
GIANT (Gene Identification and Normalization Tool) is a rule-based system that uniquely identifies human gene mentions in free text [24]. The process is divided into two major steps. The goal of the first step is to extract all the potential gene mentions from the input text. Using a set of regular expression rules, gene symbols are detected using pattern matching. An approximate term searching technique is employed for gene names to account for typical morphological variations, such as word ordering. In the second step a set of statistical and heuristic features is used to estimate the level of confidence for each mention extracted. The confidence score is essentially a weighted linear combination of individual feature scores. The feature weights are optimized using the Nelder-Mead simplex method [25]. Precision of the result is improved by filtering out mentions with low confidence scores. The system has an F measure of 0.7622 from evaluations against the BioCreative II datasets. GIANT is implemented in Java and can be accessed either through a web interface or by remote procedure calls. The system stores a local copy of the Medline collection in a relational database.

Department of Computer Science, University of Tokyo (RS, KY and YC) [Sætre]
The AKANE++ system is a recently developed sentence-level PPI system. In order to use the AKANE system for the BioCreative tasks, the output format had to be simplified, because BioCreative just considers whether the abstract level contains interacting protein pairs or not. The original format of the AKANE system used annotated sentences like those in the AImed corpus [26]. In the new system, the abstracts are sent through a processing pipeline, containing modules for sentence splitting, tokenization and parsing, and then each mention of protein names are tagged by a named entity recognizer and normalized to their UniProt Identifiers. Finally, co-occurring pairs in single sentences are used as candidates for the PPI classification system. Some simple postprocessing is done in order to transform the sentence-level results from the AKANE system into the expected format for the BioCreative II challenge. The postprocessing included filtering and ranking of the sentence-level results, and then deciding whether the collective PPI confidence was high enough to assume that the abstract contains PPI interactions [27]. A separate system developed by our team is the ProtIR, filtering and ranking articles by their PPI relevance, based on a bag-of-words IR approach. Further details can be found in [28].

Biointelligence Laboratory, Seoul National University (SK, SS and BZ) [Kim]
Our PIE (Protein Interaction Extraction) system was developed to identify the PPI information from biomedical literature. The system consists of two modules for PPI article

filtering and PPI sentence filtering. Each module uses ML techniques, and performs the filtering tasks based on the idea that the PPI descriptions have their own patterns at the article and sentence levels [29]. For the meta-services, the PPI article filter is utilized to support PPI classifications from PubMed abstracts and full text. The article filter uses a cost-sensitive learning algorithm, AdaCost [30], combined with the naïve Bayes classifiers. Unlike other ML-based classifiers minimizing the number of incorrect classifications, AdaCost provides the flexibility of controlling the precision and recall rates by means of a cost factor. In addition, naïve Bayes classifiers can easily take into account heuristic knowledge in a probabilistic form. For the AdaCost algorithm, a document is preprocessed by stemming and stopword removal. We use a modified stopword list (available at [31]), where the PPI-related words are omitted from common stopwords. Then, the remaining sentences are converted into the bag-of-words representation to discover the specific words or combinations of the words that best capture the PPI relevance at the article level.

Center for Computational Pharmacology, University of Colorado (WB and LH) [Baumgartner]

The Center for Computational Pharmacology's Annotation Server provides gene mention and gene normalization annotation, and protein interaction classification functionality on both full-text and PubMed abstracts. Annotation output is generated using an integrated approach to concept recognition. Gene mentions are detected using a stochastic tagging system built and trained for the inaugural BioCreative challenge [32]. Gene normalization is achieved by matching gene mention text to a lexicon of gene names constructed from human Entrez Gene records. Features of the normalization system include use of multiple gene taggers as input, simple conjunction resolution, a heuristic regularization procedure for processing gene names, exact matching of gene names to the lexicon, and a disambiguation step for gene names that match to multiple Entrez Gene records. Protein interactions are classified using our BioCreative II IPS (interaction pair subtask) system, which uses a concept recognition system developed by our group, OpenDMAP, and a series of manually generated patterns to classify PPIs in text [33]. Future development of the annotation server will involve streamlining the various systems to facilitate faster processing as well as incorporation of our BioCreative II ISS (interaction sentence subtask) system and extending our GN system's ability to normalize to more than just human genes.

School of Informatics, University of Edinburgh (BH, MM and XW) [Haddow]

In the system from the University of Edinburgh, the gene mentions are found using a CRF-based named entity tagger trained on the BioCreative training data. The tagger employes contextual, shallow grammatical, and morphological features tailored to the biomedical domain, as well as a gazetteer of protein names derived from RefSeq. For gene normalization, each of these gene mentions is mapped to a set of possible UniProt identifiers selected from the lexicon using a modified version of the Jaro-Winkler string similarity function [34]. To choose the most likely identifier from the set, a ML-based disambiguator (trained on BioCreative data) and a species tagger (trained on in-house data) are employed. Taxonomy annotations are also provided by the species tagger. The articles containing PPIs are selected using a SVM classifier trained on the BioCreative training set, and using conventional bag-of-words features, as well as features derived from the output of our PPI pipeline [35-39].

Medical Informatics Service, University and Hospitals of Geneva (PR, FE) [Ruch]

Our approach is based on the combination of basic pattern matching methods, the use of specialized heuristics and database resources, and a generic text categorization engine. The first step consists of extracting protein names from the abstracts together with a targeted list of the interaction verbs. The second step consists of deciding which proteins should be selected in order to build the appropriate interaction pairs. Because we have to provide a UniProt ID (accession number) for every protein, during a third step we identify the different species that appear in the documents. Indeed, protein names are usually associated with several species and therefore they are highly ambiguous regarding the sequences that they refer to. In the following step, all of the information is combined to obtain a unique UniProt ID. Finally, the interactions are ranked based on a combined model that takes into account the following features: protein names, interaction verbs, species, and word distances between these different entities. Species categories are identified using an automatic text categorization framework [40,41]. Because it is often difficult to find specific NEWT species in texts, a mapping table is manually maintained to associate MeSH-based species with their equivalent in NEWT. When no species are found, the system assumes that the protein is a human protein. The GPSDB resource [42] is used to help identify protein names in textual contents. Optionally, the system can also provide information concerning the interaction methods using the PSI-MI controlled vocabulary and the previously mentioned generic categorization system [41].

Department of Electrical Engineering and Computer Science, University of Michigan (AÖ, GE and DR) [Özgür]

We provide annotations for identifying interaction relevant articles. Our approach is based on extracting interacting protein pairs and evidence sentences from the articles by using dependency parsing and SVMs. After segmenting a given article into sentences and tagging the protein names with Genia Tagger [43], we build the dependency parse trees of the sentences by using Stanford Parser [44]. From the dependency parse trees of the sentences, we extract the shortest paths between each protein pair. We define a kernel function based on the edit distance based similarity among the extracted dependency paths. We use this kernel function with SVM to

classify each sentence as being an evidence for the interaction of a protein pair or not. We annotate an article as an interaction relevant article if it contains an evidence sentence for the interaction of a protein pair. Detailed information about our annotation system can be found in [45,46].

Yale University School of Medicine and Yale University (MK and TL) [Luong]

We believe that gene name identification is a modular process that involves term recognition, classification, and mapping [47]. Here, we focus on gene name mapping, and use an existing program (ABNER [48]) for gene name recognition and classification (entity recognition). We use a combination of two methods to map recognized entities to their appropriate gene identifiers (Entrez GeneIDs): the trigram method and the network method. Both methods require preprocessing, using resources from Entrez Gene, to construct a set of method-specific matrices. We first address lexical variation by transforming gene names into their unique trigrams (groups of three alphanumeric characters) and perform trigram matching against the preprocessed gene dictionary. For ambiguous gene names we additionally perform a contextual analysis of the abstract that contains the recognized entity. We have formalized our method as a sequence of matrix manipulations, allowing for a fast and coherent implementation of the algorithm [49].

iHOP: information hyperlinked over proteins (RH, AV and CS) [Hoffmann]

The iHOP information resource [50,51] selectively retrieves information that is specific to genes and proteins and summarizes their interactions and functions. The system supports filtering and ranking of extracted sentences according to significance, impact factor, date of publication, and syntactical properties. Entity recognition and annotation processes (GN) in iHOP are based on a dictionary approach to screen for synonyms of genes and proteins, MeSH terms, and chemical compounds. Synonym dictionaries are regularly compiled and updated from various resources (for example, NCBI and UniProt) and extended programmatically to account for orthographic variations specific to the type of entity or organism. Draft annotations from entity-specific annotator processes are integrated into a final annotation, where individual finding sites are evaluated for uniqueness (relative to the entire synonym space), quality (based on properties of the synonym and the immediate context), and confidence (based on context information in the complete document and meta information). All annotations are mapped to corresponding external databases.

Discussion

The BCMS platform unites and standardizes access to textual information extracted by various IE systems, presenting the annotations and classifications in a consistent structure. It aims to provide a public protocol to annotate biomedical text at the most basic level. At this stage, the platform provides an interface to explore and extract some of the annotation data created during the BioCreative II challenge [7], namely the four annotation types described in the Annotation systems section (above), for all of the official training and test set abstracts (a total of 22,804 Medline citations, minus 44 expired records at the time of this writing). A basic web interface and a web service API have been created. The communications layer (the XML-RPC transactions) is fully developed. The system can be synchronized with the complete PubMed/Medline database. It may be stated that the initial setup has been done, allowing us to advance to a fully featured version of the platform once the current state is accepted by the community.

Although such a distributed IE system seems fairly simple, numerous obstacles needed to be solved, such as the following.

• One of the most obvious problems is data consistency. The PubMed database is a dynamic resource in which citations are not only added but also changed and deleted (see the annual 'Medline/PubMed update charts' [52]; this affects several tens of thousands of records per year and is occurring on a daily basis.

• A less obvious difficulty is string encoding. As with biological sequences, when talking about positions and offsets in the sequences, using different encoding schemas would produce different and ultimately erroneous data. Therefore, continuous use of Unicode is enforced.

• Special attention had to be paid to the communications layer, specifically between the meta-server and the annotation servers. This component is virtually separated from the meta-server and multithreaded to ensure consistent and unimpeded usage.

At this stage, the system provides a limited compilation of the data generated during BioCreative II, offering integrated access to the systems produced by some of the groups participating in the second BioCreative challenge. The platform at its current state is confined to the approximately 22,800 abstracts used during BioCreative II. The intent is to open the system up (most likely stepwise, to avoid massive overload of the annotation queues) to the complete set of Medline records, and we are considering allowing annotation of user-provided full text after the prototype stage has been completed. The development of a platform that can operate freely on the complete set of Medline abstracts will be of great advantage to the biomedical community. Therefore, the next step is to go from the prototype state with a limited set of abstracts to the open system, where users can obtain classifications for any PubMed citation.

Conclusion

This prototype is the first meta-service for biomedical information extraction. The platform is based on design principles of simplicity and expandability. Future initiatives to expand the system, such as adding annotation types or opening the system for user-provided texts, are likely to be possible with little effort. This implies that other research groups can join the platform, providing their own annotations, including the expansion of the system for new annotation types, for example, for protein-interaction detection methods. The three main units of the system (the various annotation systems, the annotated data, and the access methods) as well as their components (data, communications, and application layer) are independent of each other, so that one of the parts can be manipulated or completely exchanged without affecting the platform as a whole.

Furthermore - and similar to the development of the meta-servers in the field of protein structure prediction [53], in which a central server collects the results of several structure prediction algorithms and unifies these to create a jury prediction on the sequence, delivering the result to the client - we foresee that this platform and possibly others would evolve to compare the various annotation systems. This is because a single dataset is returned for the complete annotations from all systems on a given abstract by using the web service interface, with calibrated annotations and systematic consensus annotations. It will be interesting to use these consensus annotations as a baseline for future BioCreative challenges.

Materials and methods

Research groups interested in contributing to the BCMS platform are requested to contact the corresponding author for further specifications and sample implementation of an AS in Java.

Data layer

The most critical element is the textual data *per se* (the PubMed/Medline records, which are represented using UTF-8 encoding). Both the meta-server and the ASs are required to have access to the same Medline records (data persistency and consistency). ASs are required to keep their local copy of Medline up to date on at least a weekly basis. There are various means by which this can be achieved, for instance by maintaining a local copy of the Medline database (FTP download) or using the eUtils service oriented architecture protocol (SOAP) API [54] to retrieve the record for the current request. The latter has the advantage of being up to date with the very latest state (data consistency) from Medline, but it also makes the AS dependent on the availability of the NCBI service. This online solution is the default provided by the sample AS implementation.

Communications layer

For solving inter-server communications, as well as client to meta-server communications, the three most common web protocols were considered: representational state transfer (REST), XML-RPC, and SOAP. XML-RPC was chosen for the following reasons.

• General: it is the oldest protocol available, which means that it is the most widespread (many web libraries include this protocol by default) and best known, and can therefore be assumed to be robust.

• Versus REST: it has the advantage that message length is not limited, as opposed to REST, which uses the URL to transmit parameters. The general World Wide Web Consortium recommendation for web development is not to use URLs of more than 2,000 characters. These URL specifications would limit free text annotations.

• Versus SOAP: XML-RPC is much less complex, and the specification paper is about one-sixth of the SOAP specification. All functionality required for the platform is provided by the XML-RPC protocol already. Therefore, XML-RPC guarantees rapid and simple implementation and provides a straightforward means of debugging and maintaining the system.

In the current setup, teams providing annotations can either create their own XML-RPC server implementation (following the platform guidelines) or they can simply use the default AS system, which is a standard Java implementation based on the Apache XML-RPC library (because Java is the most common programming language and many sites use the Apache framework), which can be requested from the corresponding authors.

Application layer

The following tools, libraries, and applications have been used to develop the platform.

• PostgreSQL 8.1: database for both Medline and the annotation data [55].

• Django 0.96/stable: WDF, embedded in Apache 1.3 [56].

• jQuery 1.2.1: AJAX and interactive webpage elements [57].

• LingPipe 3.2.0: download of and synchronization with the Medline database [58] (on a daily basis).

• Python 2.5: in-house implementation of the XML-RPC communications layer using the standard library [59].

The versions correspond to the latest used versions as of the time of this writing and are subject to change.

Abbreviations

AS, annotation server; BCMS, BioCreative MetaServer; CRF, conditional random field; GIANT, Gene Identification and Normalization Tool; GM, gene mention; GN, gene normalization; IE, information extraction; iHOP, Information Hyperlinked Over Proteins; IR, information retrieval; ML, machine learning; NCBI, National Center for Biotechnology Information; PPI, protein-protein interaction; REST, representational state transfer; RPC, remote procedure call; SOAP, service oriented architecture protocol; SVM, support vector machine.

Competing interests

The work of BH, MM, and XW was funded by ITI Life Sciences, Scotland, whose mission is to explore commercialization of promising technologies in the life sciences. All other authors declare that they have no competing interests.

Authors' contributions

AV and MK: BioCreative II workshop. AV, CRP and MK contributed to meta-server system design. FL: meta-server system design and implementation. AV: supervisor at CNIO. All others: annotation servers and systems. Author names were sorted by institution/university names except for CNIO authors. All authors have read and approved the final manuscript.

Acknowledgements

We should like to thank a number of bioinformaticians who have discussed with us the need for this type of system for their own research, in particular Jaak Vilo and Ewan Birney. Also, we should like especially to thank all participants of the second BioCreative challenge and their vital contributions during the discussions at the workshop and in the mailing lists. Without this feedback, this project would not have been possible. FE and PR would like to thank Julien Gobeill and Anne-Lise Veuthey for their assistance during the campaign. AÖ, GE and DR would like to thank David J States for his helpful comments. RH, AV and CS: the iHOP resource is hosted by the Computational Biology Department of the Memorial Sloan Kettering Institute. The CNIO authors would like to appreciate all the participants who were willing to join this initiative at such an early stage for their contributions. FL would like to thank Jörg Hakenberg for his contributions to the XML-RPC server development, Bob Carpenter for his help with the PubMed/MEDLINE importer provided by his LingPipe library, and José Manuel Rodríguez Carrasco for interfacing with the iHOP system.

• CP and JH: we kindly acknowledge funding by the EC Sixth Framework Programme SEALIFE project, number IST-2006-027269, and by the German BMBF, grant contract 0312705B.

• CJ and WL: this work was supported by the Intramural Research Program of the NIH, CIT.

• SK and BZ: this work was supported by KOSEF through the NRL Program (number M10400000349-06J0000-34910).

• SS: this work was supported by the Korea Research Foundation Grant funded by Korean Government (MOEHRD; KRF-2006-214-D00140).

• BH, MM and XW: this work was carried out as part of an ITI Life Sciences Scotland [60] research programme with Cognia EU [61] and the University of Edinburgh.

• AÖ, DR and GE: this work was supported in part by grants R01-LM008106 and U54-DA021519 from the US National Institutes of Health.

• FE and PR: our participation in the BioCreative competition would not have been possible without the support of the Swiss National Science Foundation (EAGL: Engine for Question Answering in Genomic Literature; SNF 3252B0-105755).

• AV, CRP, FL and MK: BioCreative is supported by donations from the ENFIN European Comission FP6 Programme NoE LSHG-CT-2005-518254, the workshop by a grant from the ESF Programme 'Frontiers of Functional Genomics'. The research group at the Spanish National Cancer Research Centre (CNIO) is funded by the DIAMONDS European Comission grant LSHG-CT-2004-512143 and by the National Institute for Bioinformatics [62], a platform of 'Genoma España'.

This article has been published as part of *Genome Biology* Volume 9 Supplement 2, 2008: The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge. The full contents of the supplement are available online at http://genomebiology.com/supplements/9/S2.

References

1. Krallinger M, Valencia A: **Text-mining and information-retrieval services for molecular biology.** *Genome Biol* 2005, **6:**224.
2. Cohen A, Hersh W: **A survey of current work in biomedical text mining.** *Brief Bioinform* 2005, **6:**57-71.
3. Labarga A, Valentin F, Anderson M, Lopez R: **Web Services at the European Bioinformatics Institute.** *Nucleic Acids Res* 2007:W6-W11.
4. Fernández J, Hoffmann R, Valencia A: **iHOP web services.** *Nucleic Acids Res* 2007:W21-W26.
5. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **Structure prediction meta server.** *Bioinformatics* 2001, **17:**750-751.
6. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2:**7.
7. **BioCreative Homepage** [http://biocreative.sourceforge.net/]
8. **XML-RPC Specification** [http://www.xmlrpc.com/]
9. **BioCreative MetaServer** [http://bcms.bioinfo.cnio.es/]
10. **BioCreative XML-RPC MetaService** [http://bcms.bioinfo.cnio.es/xmlrpc/]
11. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, Hirschman L, Valencia A: **Evaluation of text mining systems for biology: overview of the Second BioCreative community challenge.** *Genome Biol* 2008, **9(Suppl 2):**S1.
12. Smith L, Tanabe LK, Johnson nee Ando R, Kuo C-J, Chung I-F, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K, Torii M, Liu H, Haddow B, Struble CA, Povinelli RJ, Vlachos A, Baumgartner WA Jr, Hunter L, Carpenter B, Tsai RT-H, Dai H-J, Liu F, Chen Y, Sun C, Katrenko S, Adriaans P, Blaschke C, Torres R, Neves M, Nakov P, *et al.*: **Overview of BioCreative II gene mention recognition.** *Genome Biology* 2008, **9(Suppl 2):**S2.
13. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu H-h, Torres R, Krauthammer M, Lau WW, Liu H, Hsu C-N, Schuemie M, Cohen KB, Hirschman L: **Overview of BioCreative II gene normalization.** *Genome Biol* 2008, **9(Suppl 2):**S3.
14. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biology* 2008, **9(Suppl 2):**S4.
15. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M: **Gene mention normalization and interaction extraction with context models and sentence motifs.** *Genome Biol* 2008, **9(Suppl 2):**S14.
16. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph.** *Bioinformatics* 2006, **22:**2444-2445.
17. Kuo CJ, Chang YM, Huang HS, Lin KT, Yang BH, Lin YS, Hsu CN, Chung IF: **Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007.
18. **Mallet: A machine learning for language toolkit** [http://mallet.cs.umass.edu]
19. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J: **Developing a robust part-of-speech tagger for biomedical text.** In *Advances in Informatics, 10th Panhellenic Conference*

on Informatics; 11-13 November 2005 Volos, Greece. Springer; 2005:382-392.
20. Dai HJ, Hung HC, Tsai RTH, Hsu WL: **IASL systems in the gene mention tagging task and protein interaction article subtask.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007.
21. Tsai RTH, Sung CL, Dai HJ, Hung HC, Sung TY, Hsu WL: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition.** *BMC Bioinformatics* 2006, **7(suppl 5):**S11.
22. Tsai RTH, Hung HC, Dai HJ, Hsu WL: **Exploiting likely-positive and unlabeled data to improve the identification of protein-protein interaction articles.** *Proceedings of the 6th International Conference on Bioinformatics; HongKong-Hanoi-Nansha; 27-31 August 2007* .
23. **Sinica Annotation Server - Web Service** [http://asqa.iis.sinica.edu.tw:8081/XmlRpcServlet]
24. Lau WW, Johnson CA: **Rule-based human gene normalization in biomedical text with confidence estimation.** *Comput Syst Bioinformatics Conf* 2007, **6:**371-379.
25. Nelder J, Mead R: **A simplex method for function minimization.** *Computer J* 1965, **7:**308-313.
26. Sætre R, Sagae K, Tsujii J: **Syntactic features for protein-protein interaction extraction.** *Short Paper Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM-2007); 6-7 December 2007; Singapore* .
27. Sætre R, Yoshida K, Yakushiji A, Miyao Y, Matsubyashi Y, Ohta T: **AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007:209-212.
28. Chen YH, Ramampiaro H, Lægreid A, Sætre R: **ProtIR prototype: abstract relevance for protein-protein interaction in BioCreAtIvE2 challenge, PPI-IAS subtask.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007:179-182.
29. Jang H, Lim J, Lim JH, Park SJ, Lee KC, Park SH: **Finding the evidence for protein-protein interactions from PubMed abstracts.** *Bioinformatics* 2006, **22:**e220-e226.
30. Fan W, Stolfo S, Zhang J, Chan P: **AdaCost: misclassification cost-sensitive boosting.** *Proceedings of the 16th International Conference on Machine Learning; 27-30 1999. Bled, Slovenia* 1999:97-105.
31. **PIE: Protein Interaction Information Extraction** [http://bi.snu.ac.kr/pie]
32. Kinoshita S, Cohen KB, Ogren PV, Hunter L: **BioCreAtIvE Task1A: entity identification with a stochastic tagger.** *BMC Bioinformatics* 2005, **6(suppl 1):**S4.
33. Baumgartner WA Jr, Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A, White EK, Medvedeva O, Cohen KB, Hunter L: **Concept recognition for extracting protein interaction relations from biomedical text.** *Genome Biology* 2008, **9(Suppl 2):**S9.
34. Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, Tobin R, Wang X: **Automating curation using a natural language processing pipeline.** *Genome Biol* 2008, **9(Suppl 2):**S10.
35. Grover C, Haddow B, Klein E, Matthews M, Nielsen LA, Tobin R, Wang X: **Adapting a relation extraction pipeline for the BioCreAtIvE II task.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007.
36. Alex B, Haddow B, Grover C: **Recognising nested named entities in biomedical text.** *Proceedings of BioNLP; June 2007; Prague, Czech Republic* 2007:65-72.
37. Wang X: **Rule-based protein term identification with help from automatic species tagging.** *Proceedings of CICLING; Mexico City, Mexico* 2007:288-298.
38. Nielsen LA: **Extracting protein-protein interactions using simple contextual features.** *Proceedings of BioNLP; New York* 2006:120-121.
39. Matthews M: **Improving biomedical text categorization with nlp.** *Proceedings of the SIGs, The Joint BioLINK-Bio-Ontologies Meeting* 2006:93-96.
40. Ehrler F, Geissbuhler A, Jimeno A, Ruch P: **Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot.** *BMC Bioinformatics* 2005, **6(suppl 1):**S23.
41. Ruch P: **Automatic assignment of biomedical categories: toward a generic approach.** *Bioinformatics* 2006, **22:**658-664.
42. Pillet V, Zehnder M, Seewald AK, Veuthey AL, Petrak J: **GPSDB: a new database for synonyms expansion of gene and protein names.** *Bioinformatics* 2005, **21:**1743-1744.
43. **Genia Tagger** [http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/]
44. de Marneffe MC, MacCartney B, Manning CD: **Generating typed dependency parses from phrase structure parses.** *5th International Conference on Language Resources and Evaluation (LREC 2006); 2006; Genoa, Italy* .
45. Erkan G, Özgür A, Radev DR: **Semi-supervised classification for extracting protein interaction sentences using dependency parsing.** *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL); Prague, Czech Republic* 2007, **1:**228-237.
46. Erkan G, Özgür A, Radev DR: **Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007.
47. Krauthammer M, Nenadic G: **Term identification in the biomedical literature.** *J Biomed Inform* 2004, **37:**512-526.
48. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.** *Bioinformatics* 2005, **21:**3191-3192.
49. Luong T, Tran N, Krauthammer M: **Context-aware mapping of gene names using trigrams.** In *Proceedings of the Second BioCreative Challenge Workshop* Madrid, Spain. CNIO; 2007:145-148.
50. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004, **36:**664.
51. Hoffmann R, Valencia A: **Implementing the iHOP concept for navigation of biomedical literature.** *Bioinformatics* 2005, **21(suppl 2):**ii252-22258.
52. **MEDLINE/PubMed update charts** [http://www.nlm.nih.gov/bsd/licensee/table_rev.html]
53. Valencia A: **Meta, Meta(N) and cyber servers.** *Bioinformatics* 2003, **19:**795.
54. **eUtils SOAP API** [http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html]
55. **PostgreSQL Open Source Database** [http://www.postgresql.org/]
56. **Django Web Development Framework** [http://www.djangoproject.com/]
57. **jQuery JavaScript and AJAX library** [http://jquery.com/]
58. **LingPipe - Java Text Mining Library and Medline Importer** [http://www.alias-i.com/lingpipe/]
59. **Python Programming Language** [http://www.python.org/]
60. **ITI Life Sciences Homepage** [http://www.itilifesciences.com]
61. **Cognia EU Homepage** [http://www.cognia.com]
62. **Instituto Nacional de Bioinformática** [www.inab.org]