SESSIONS *ABSTRACTS* **CHAIR**

Chair01 **Hagit Shatkay**

BioCreative II: Gene Mention Task

W. John Wilbur

[1.] CNIO, Madrid, Spain
[2.] CIPF, Valencia, Spain
[3.] Bioinformatics Unit,CNIO
[4.] Genetics and Pathology, Hospital Virgin de Salud, Toledo
[5.] Heamtology, Universty Hospital Puerta de Hierro

Although B cells, T cells, histiocytes and dendritic cells all derive from a common stem cell, it has generally been believed that once lineage commitment takes place, reversion to another lineage does not occur. Recent studies in murine systems have suggested that modulation of transcription factors in vitro can lead to reprogramming of B-cells into macrophages. (Xie, H., M. Ye, et al. (2004). “Stepwise reprogramming of B cells into macrophages.” Cell 117 (5): 663-76.) However, it is not known whether similar events take place in vivo, and under what circumstances they might occur. We have recently reported reprogramming of precursor B-cells and T-cells into histiocytes and Langerhans cell respectively. In both situations, patients with precursor B-cell or T-cell lymphoblastic lymphoma/leukemia (LBL) developed a Tumour lacking phenotypic evidence of B-cells or T-cells, but exhibiting markers of histiocytes and Langerhans cells. The histiocytic and dendritic cell neoplasms were clonally related to their B-cell and T-cell counterparts, and demonstrated identical clonal IgH and TCR gene rearrangements. In one case, a Langerhans cell Tumour arose simultaneously with precursor-T cell LBL, indicating that therapy did not play a role in driving the reprogramming of the T-cell neoplasm. In the above instances, programming occurred in a neoplastic cell with an immature, LBL phenotype. Reprogramming of a mature lymphoid malignancy seems less likely to occur. We have recently identified six cases of histiocytic sarcoma in patients with follicular lymphoma (FL). All cases of FL were positive for t(14;18). Four histiocytic Tumours were metachronous, following FL by <1 to 12 y. Two patients had synchronous FL and histiocytic sarcoma. As demonstrated by PCR or FISH for BCL2/JH, in 5/6 cases the histiocytic Tumour was clonally related to the FL and also carried the t(14;18).

Identifying Gene Mentions by Case-Based Classification

Mariana Lara Neves

Facultad
de Informática,
Universidad
Complutense
de Madrid,
Madrid, Spain

The work presented here proposes a case-based classification for the gene mention task in the BioCreAtIvE 2 challenge. The classification performed by the system for each word in an article is based on the selection of the best or more similar case in a base of known and unknown cases. The procedure showed good results, precision of 71.68 and recall of 62.33.

Keywords: text mining, gene mention, case-based reasoning.

Chair01 **Hagit Shatkay**

IASL Systems in the Gene Mention Tagging Task and Protein Interaction Article Sub-task

Hong-Jie Dai, Hsi-Chuan Hung, **Richard Tzong-Han Tsai,** Wen-Lian Hsu

Intelligent Agent Systems Lab (IASL), Institute of Information Science, Academia Sinica, Taipei, Taiwan

This named entity recognition (NER) task is a crucial step for information extraction of relationships between genes and gene products. BioCreAtIvE II Gene Mention (GM) tagging task is concerned with this problem. The first part of this work employs: 1) Conditional random fields (CRF) as underlying machine learning model, 2) A set of features which are selected by sequential forward search algorithm, 3) Numerical normalization, and 4) Pattern-based post processing to resolve the GM task.

For GM task, we collect train/testing/development dataset from BioCreAtIvE I and II to form a totally 15,443 sentences training set. In order to make use of this training set, we build a rule-based tokenizer bases on the dataset from BioCreAtIvE I Task 1A. This tokenizer is also used to tokenize the train/testing set in our BioCreAtIvE II GM task and Protein Interaction Article Sub-task 1 (IAS).

The second part of this paper is about identifying protein-protein interaction (PPI) related biomedical abstracts. We propose a novel feature representation scheme, contextual-bag-of-words, to exploit named entity information. We further improve the performance by extracting reliable and informative instances from unlabeled and likely positive data to provide additional training data.

Combined Conditional Random Fields and n-Gram Language Models for Gene Mention Recognition

Craig A. Struble, Richard J. Povinelli, Michael T. Johnson, Dina Berchanskiy, Jidong Tao, Marek Trawicki

Department of Mathematics, Statistics, and Computer Science, Department of Electrical and Computer Engineering, Marquette University, Milwaukee, USA

We propose the use of character n-gram and multiple conditional random field (CRF) models for BioCreAtIvE 2 Task 1, gene/protein name recognition. We investigated different state transition weighting schemes for CRFs and discovered that models provided independent non-overlapping mentions. To improve recall, the results of multiple CRF models are combined. To improve precision, character n-gram models classify gene/protein mention containing sentences.

Our best approach achieved a precision of 84.35%, recall of 81.39% and F-measure of 82.85%.

References

[1] alias i. LingPipe. http://www.alias-i.com/lingpipe/index.html, 2006. Version 2.3.0.

[2] BioCreAtIvE II: Critical assessment for information extraction in biology challenge (2006–2007). http://biocreative.sourceforge.net/biocreative_2.html.

[3] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics, 6 Suppl 1:S1, 2005.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. 18th International Conf. on Machine Learning, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[5] A. K. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[6] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics, 6 Suppl 1:S6, 2005.

[7] A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput, pages 451–462, 2003.

[8] B. Settles. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004, pages 107–110, 2004.

[9] R. Talreja. GeneTaggerCRF. http://www.cis.upenn.edu/datamining/software_dist/biosfier/.

Keywords: conditional random field, named entity recognition, n-gram models

Chair01 **Hagit Shatkay**

Tackling the BioCreative2 Gene Mention task with Conditional Random Fields and Syntactic Parsing

Andreas Vlachos

Computer Laboratory, University of Cambridge, Cambridge, UK

This paper presents an approach to Gene Mention tagging using Conditional Random Fields (CRFs) and syntactic parsing, by taking advantage of the flexibility of the former in order to add features from the output of the latter. We did not use any material or information other than the training data provided in order to maintain the domain independence of the system. Nevertheless, the resulting system achieved 82.84% F-score, which places it in the second performance quartile of the competition.

Keywords: CRFs, syntactic parsing, gene mention tagging.

Named Entity Recognition with Combinations of Conditional Random Fields

Roman Klinger, Christoph M. Friedrich, Juliane Fluck

Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

The Gene Mention task is a Named Entity Recognition (NER) task for labeling gene and gene product names in biomedical text. To deal with acceptable alternatives additionally to the gold standard, we use combinations of Conditional Random Fields (CRF) together with a normalizing tagger. This process is followed by a postprocessing step including an acronym disambiguation based on Latent Semantic Analysis (LSA). For robust model selection we apply 50-fold Bootstrapping to obtain an average F-Score of 84.58 % on the trainingset and 86.33 % on the test set.

Keywords: named entity recognition, text mining, data mining, conditional random fields, multi model approach.

Using Semi-Supervised Techniques to Detect Gene Mentions

Sophia Katrenko, Pieter Adriaans

Human-Computer Studies Lab, Institute of Informatics, University of Amsterdam, Amsterdam, The Netherlands

To find gene mentions in a corpus, we investigate the semi-supervised learning techniques. In particular, we consider co-training (with orthographic/contextual features split) and self-training using a subset of Genia corpus as a pool of unlabeled data.

In the self-training setting we carried out several experiments by varying the number of iterations, the size of the training set and the size of the unlabeled pool. The run we submitted had the following settings: number of iterations is equal to 5, number of instances added in each iteration is 100, and 1,000 Medline sentences from Genia corpus [8] are used as source of the unlabeled data.

In each ietarion, only the most confident predictions are added (top N). In this setting precision is much higher than recall (82,28% versus 71,08%) and F-score equals 76.27%.

In the co-training setting, the number of iterations was set to 6. Surprisingly, self-training outperformed co-training (F-score dropped to 71,74%). Co-training nevertheless provides better results than applying contextual and orthographic models separately.

Keywords: co-training, self-training.

BioCreative II Gene Mention Tagging System at IBMWatson

Rie Kubota Ando

IBM T.J. Watson
Research Center, ,
Hawthorne,
New York, USA

This paper describes our system developed for the BioCreative II gene mention tagging task. The goal of this task is to annotate mentions of genes or gene products in the given Medline sentences. Our focus was to experiment with a semi-supervised learning method, Alternating Structure Optimization (ASO) [1], by which we exploited a large amount of unlabeled data in addition to the labeled training data provided by the organizer. The system is also equipped with automatic induction of high-order features, gene name lexicon lookup, classifier combination, and simple postprocessing.

Our system appears to be competitive. All of our three official runs belong to the Quartile 1.

Chair02John Wilbur

Rich Feature Set, Unification of Bidirectional Parsing and Dictionary Filtering for High F-Score Gene Mention Tagging

Cheng-Ju Kuo, Yu-Ming Chang, Han-Shen Huang, Kuan-Ting Lin, Bo-Hou Yang, Yu-Shi Lin, Chun-Nan Hsu, I-Fang Chung

Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan
Institute of Information Science, Academia Sinica, Taipei, Taiwan
Department of Electrical Engineering, Chang-Gung University, TaoYuan, Taiwan

In the first BioCreative (2004), conditional random fields (CRF) were employed in tagging gene and protein mentioned in the biomedical text with high performance. Therefore, we chose CRF as our starting point and carefully selected a rich set of 5,059,368 predictes as the features. To further improve its performance, we combined the solutions of forward and backward parsing, a trick commonly used by biologists in sequencing. We tried different combination methods, including set operations and Co-Training. However, we found that Co-Training performed poorly. Instead, we selected the best solutions from the “adjacent” ten candidates of bidirectional parsing and then applied dictionary filtering to obtain the best F-score result.

High-Recall Gene Mention Recognition by Unification of Multiple Backward Parsing Models

Han-Shen Huang, Yu-Shi Lin, Kuan-Ting Lin, Cheng-Ju Kuo, Yu-Ming Chang, Bo-Hou Yang, I-Fang Chung, **Chun-Nan Hsu**

Institute of Information Science, Academia Sinica, Taiwan
Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan
Department of Electrical Engineering, Chang-Gung University, TaoYuan, Taiwan

We considered the gene mention tagging task as a classification problem and applied support vector machines (SVM) to solve it. We selected a large set of features as the input and trained two

SVM models with different multiclass extension methods. We found that backward parsing onstantly outperformed forward parsing regardless of the multiclass extension methods and obtained high precision rates, but recall rates were not as satisfactory. To enhance recall rates, our approach is to construct divergent but high performance models to cover different aspects of the feature space, and then combine them into an ensemble. We applied union and intersection to combine the outputs of SVM models with that of a CRF model, which was trained with the same

feature set, and successfully enhanced recall rates without degrading too much precision.

Attribute Analysis in Biomedical Text Classification

Francisco Carrero García[1], Enrique Puertas[1], José María Gómez Hidalgo[1], Manuel Maña[2]

[1] Universidad Europea de Madrid, Villaviciosa de Odón, madrid, Spain
[1] Universidad de Huelva, Huelva, Spain

Text Classification tasks are becoming increasingly popular in the field of Information Access. Being approached as Machine Learning problems, the definition of suitable attributes for each task is approached in an ad-hoc way. We believe that a more principled framework is requiered, and we present initial insights on attribute engineering for Text Classification systems. The library is currently being used and evaluated in our Information Access projects in the biomedical domain. In this paper we describe how we have used it in the Gene mention and the Protein-Protein Interaction (Protein Interaction Article) tasks in the Biocreative II Challenge.

Keywords: text classification, machine learning, attribute engineering.

Penn/UMass/CHOP Biocreative II systems

Kuzman Ganchev[1], Koby Crammer[1], Fernando Pereira[1], Gideon Mann[2], Kedar Bellare[2], Andrew McCallum[2], Steven Carroll[3], Yang Jin[3],Peter White[3]

[1.] Department of Computer and Information Science, University of Pennsylvania, PhiladelphiA, USA
[2.] Department of Computer Science, University of Massachusetts, Amherst, USA
[3.] Division of Oncology, The Children's Hospital of Philadelphia,

Our team participated in the entity tagging and normalization tasks of Biocreative II. For the entity tagging task, we used a k-best MIRA learning algorithm with lexicons and automatically derived word clusters. MIRA accommodates different training loss functions, which allowed us to exploit gene alternatives in training. We also performed a greedy search over feature templates and the development data, achieving a final F-measure of 86.28%. For the normalization task, we proposed a new specialized on-line learning algorithm and applied it for filtering out false positives from a high recall list of candidates. For normalization we received an F-measure of 69.8%.

Keywords: entity tagging, entity normalization, linear sequence models

Chair02John Wilbur

Corpus Annotation and Its Use in BioNLP

Tsujii Junichi

University of Tokyo and University of Manchester, Tokyo/Manchester, Japan/UK

The corpus that we have developed, GENIA, is richly annotated from the view points of both linguistics and biology. The multi-layered annotation of GENIA consists of annotation layers of token, POS, syntactic tree, semantic structure (PAS), co-reference, and named entity.

We are now adding another annotation layer, e.g. events. While event annotation of GENIA has not finished yet, 1000 abstracts have already been annotated. It has been used for an event recognition program, which shows promising performance.

The usefulness of a corpus such as GENIA as gold standard for evaluation and training data for domain adaptation has increasingly been recognized by the Bio-NLP community. In my talk, I will introduce the current state of the GENIA corpus with several use cases of adaptation in our group. We recognize that adaptation is one of the keys for the success of NLP and Text Mining. In particular, since language used in biology shows peculiar characteristics which are not frequently observed in general language such as English in newswire articles. Off-the-shelf NLP tools have to be adapted to the special sublanguage (or sublanguages) in biology.

For example, our CRF-based POS tagger has achieved the highest score (97.18) among the best performers in published papers, when trained by WSJ and applied to WSJ. However, without domain adaptation, the performance of the WSJ-trained tagger dropped significantly to 94.5 when applied to GENIA. When the WSJ-trained model was re-trained by using GENIA (20,000 sentences), the final performance improved to 98.58%, which is even higher than that of the WSJ-trained model when applied to WSJ.

Special care has also been taken in our group to development of low-cost adaptation for NLP tools, and we use the GENIA and PTB corpora for experiments which show the effectiveness of adaptation techniques. For example, an active learning technique used in the POS tagger was shown to require only 15% of the whole GENIA corpus to achieve performance of 98.40 %, which is very close to the final performance (98.58).

Our deep parser (Enju) also requires adaptation. The adaptation technique we developed uses an exiting model as reference model and learn a new model by annotated corpus from the target domain. The technique has been shown very effective by experiment where we use a WSJ-trained model as reference and the GENIA tree annotation as a corpus from the target domain.

The WSJ-trained Enju shows performance of 89.81 in terms of labeled F-value of predicate-argument relations. This performance is among the highest of deep parsers. While the performance dropped to 86.39% when applied to GENIA without adaptation, a GENIA-trained Enju with the WSJ model as reference has achieved 90.15 %, which is again higher than that of the WSJ trained model when applied to WSJ.

These examples show that properly annotated corpora with proper adaptation techniques improve the performance of NLP tools significantly.

Chair03Luis Rocha

Overview of BioCreative II Gene Normalization

Lynette Hirschman

[1.] CNIO, Madrid, Spain
[2.] CIPF, Valencia, Spain
[3.] Bioinformatics Unit,CNIO
[4.] Genetics and Pathology, Hospital Virgin de Salud, Toledo
[5.] Heamtology, Universty Hospital Puerta de Hierro

Although B cells, T cells, histiocytes and dendritic cells all derive from a common stem cell, it has generally been believed that once lineage commitment takes place, reversion to another lineage does not occur. Recent studies in murine systems have suggested that modulation of transcription factors in vitro can lead to reprogramming of B-cells into macrophages. (Xie, H., M. Ye, et al. (2004). “Stepwise reprogramming of B cells into macrophages.” Cell 117 (5): 663-76.) However, it is not known whether similar events take place in vivo, and under what circumstances they might occur. We have recently reported reprogramming of precursor B-cells and T-cells into histiocytes and Langerhans cell respectively. In both situations, patients with precursor B-cell or T-cell lymphoblastic lymphoma/leukemia (LBL) developed a Tumour lacking phenotypic evidence of B-cells or T-cells, but exhibiting markers of histiocytes and Langerhans cells. The histiocytic and dendritic cell neoplasms were clonally related to their B-cell and T-cell counterparts, and demonstrated identical clonal IgH and TCR gene rearrangements. In one case, a Langerhans cell Tumour arose simultaneously with precursor-T cell LBL, indicating that therapy did not play a role in driving the reprogramming of the T-cell neoplasm. In the above instances, programming occurred in a neoplastic cell with an immature, LBL phenotype. Reprogramming of a mature lymphoid malignancy seems less likely to occur. We have recently identified six cases of histiocytic sarcoma in patients with follicular lymphoma (FL). All cases of FL were positive for t(14;18). Four histiocytic Tumours were metachronous, following FL by <1 to 12 y. Two patients had synchronous FL and histiocytic sarcoma. As demonstrated by PCR or FISH for BCL2/JH, in 5/6 cases the histiocytic Tumour was clonally related to the FL and also carried the t(14;18).

Text Detective: Alma Bioinformatics' gene/proteins annotation tool

Rafael Torres, Pablo D. Sánchez, **Christian Blaschke**

Alma Bioinformatics, SL, Tres Cantos, Spain

Text Detective is a system based on carefully constructed rules and lexicons that detects genes and proteins mentioned in biomedical abstracts. Text Detective was used in BioCreative 2007 in the Gene Mention (GM) and Gene Normalization (GN) tasks. The results have been: P=84.3; R=68.6; F=75.6 (Q3,Q4) for GM and P=74.3; R=80.7; F=77.4 (Q1) for GN.

Keywords: name entity recognition, BioCreative, gene name disambiguation.

Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup

Martijn Schuemie, Rob Jelier, Jan Kors

Biosemantics Group, Medical Informatics Department, ErasmusMC University Medical Center Rotterdam, Rotterdam, The Netherlands

To achieve high speed with minimal effort, we created a system dubbed Peregrine that performs gene name normalization by simple dictionary lookup followed by several post-processing steps.

The system was tested with two different dictionaries: the dictionary that was provided by the Biocreative 2 organisation, and a dictionary that was constructed by combining different gene and protein databases.

Peregrine tokenises both the terms in the dictionary, and the text that is to be analysed. Gene and protein names are found by matching the sequences of tokens constituting dictionary terms in the text.

We investigated several steps for improving the performance of Peregrine: (1) Manual curation of the 250 terms most frequently found in Medline, (2) automatic generation of common spelling variations, (3) automatic filtering of highly ambiguous terms, (3) automatics removal of family names, (4) simple rule based homonym disambiguation, and (5) keyword detection for improving disambiguation performance.

The resulting system is fast and can analyse 100,000 Medline records in 212 seconds on a single computer. When tested on the Biocreative 2 test set, the system achieves a maximum precision of 0.75, at a recall of 0.76. Each of the post-processing steps improved the performance of the system.

Keywords: gene name normalization, dictionary.

Gene Mention and Gene Normalization Based on Machine Learning and Online Resources

Hongfang Liu[1], Manabu Torii[1], Zhang-Zhi Hu[2], Cathy Wu[2]

[1.] Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, 4000 Reservoir Road NW, Washington, DC, USA
[2.] Protein Information Resource, Georgetown University Medical Center, , Washington DC, , USA

The identification of phrases in text representing genes/proteins and the mapping of those phrases with entries in databases are critical for literature mining applications in the biomedical domain. In this paper, we report the participation of BioTagger, an automated gene/protein name identification and normalization system, in both the gene mention and the gene normalization tasks of BioCreAtIvE, a competition for automated gene/protein name identification and mapping. For the gene mention task (i.e., gene/protein name identification), we used BioThesaurus, a collection of synonyms for all protein records in UniProtKB, and Metathesaurus, a collection of synonyms for medical concepts available at the Unified Medical Language System (UMLS). The machine learning task for gene mention was defined by i) transforming each word into a feature vector consisting of various types of features, and ii) training a classification system using Conditional Random Field (CRF) to classify each word to three categories: *B* word (beginning of a gene mention phrase), I word (inside of a gene mention phrase), and *O* word (outside of a gene mention phrase). For the gene normalization task, we assembled a dictionary consisting of synonyms for each gene record from online resources such as BioThesaurus and HUGO, conducted flexible dictionary lookup, and obtained a list of mapping pairs (Phrase, EGID), where Phrase is a term in text and EGID is one of the associated Entrez gene identifiers. We then defined a machine learning task to classify each mapping pair as Positive or Negative. Features were derived based on the mapping information related to Phrase and EGID in the corresponding document. We experimented with various machine learning algorithms available in Weka, a machine learning software package written in JAVA, and chose the one with the best performance (i.e., Bagging on Decision Tree). Our system achieved F-measures of over 85% for the gene mention task and around 78% for the gene normalization task.

Keywords: gene mention, gene normalization, machine learning, online resources, literature mining.

Chair03 **Luis Rocha**

Me and my Friends: Gene Mention Normalization with Background Knowledge

Jorg Hakenberg[1], Loic Royer[1], Conrad Plake[2], Hendrik Strobelt[1], Michael Schroeder[1]

[1.] Biotechnological Centre, Technical University of Dresden, Dresden, Germany
[2.] Transinsight GmbH, Dresden, Germany

"Tell me who your friends are, and I will tell you who you are" – this proverb best illustrates our approach to the normalization of gene names. In this approach, we rely on background knowledge that describes various aspects of a gene: it is localized on a chromosomal band, it belongs to an operon structure, it is a member of a gene family, its products take part in biological processes, they fulfil molecular functions, they occur at dedicated cellular locations, mutations of the gene ultimately cause diseases, its proteins contain domains and form secondary, tertiary and quaternary structures. Whenever a gene (or one of its products) is discussed, some of these aspects –the gene's friends– will be mentioned as well. The paradigm we follow with this approach demands not only the presence of a gene's name, but also of some of its friends.

The system we propose for identification of gene names in texts consists of four major components. The basic step provides an initial recognition of candidate terms, which also assigns all potential EntrezGene IDs to each candidate. From there on, the next components deal with refining these candidate hits: removal of false positives and disambiguation of polysemous names. Thus, the second component masquerades text parts that never contain a gene name but might account for errors of the recognition step. The third component filters false positives by looking at term frequencies, and reduces the candidate IDs by comparing new to known texts. The final component disambiguates remaining terms and identifiers using each gene's typical context. On the BioCreative2 GN test set, our system achieves an F1-measure of 81% (our highest recall: 87.5%, highest precision: 79%).

Context-Aware Mapping of Gene Names using Trigrams

ThaiBinh Luong[1,2], Nam Tran[1], Michael Krauthammer[1,2]

[1.] Department of Pathology, Yale University, New Haven, USA
[2.] Program for Computational Biology and Bioinformatics, Yale University, New Haven, USA

We present a method for the mapping of gene names to Entrez Gene identifiers. We first resolve lexical variation by transforming domain terms into their unique trigrams, and use this representation for a preliminary term mapping. We then perform fine-mapping via contextual analysis of the abstract that contains the domain term. We have formalized our method as a sequence of matrix manipulations, allowing for a fast and coherent implementation of the algorithm. We pair our method with existing approaches for entity recognition, and achieve an F-score of 0.761 in the BioCreative 2 Gene Normalization Task.

ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries

Juliane Fluck, Heinz Theodor Mevissen, Holger Dach, Marius Oster, Martin Hofmann-Apitius

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department of Bioinformatics, Schloss Birlinghoven, Sankt Augustin, Germany

For the recognition of gene and protein names and their normalization to gene and protein centered databases (Entrez Gene and UniProt) regularly updated dictionaries generated from these sources are used by the ProMiner system to search gene and protein names in scientific publications. A multistep curation process and inclusion of different biomedical dictionaries in the curation process leads to an increase of precision and recall. The recognition of names containing special parenthesis expressions augments the recall further. Human gene and protein names in the test corpus provided in BioCreAtIvE II could be recognized with the adapted ProMiner system and a regularly updated dictionary with a final F-measure of 80 %.

Keywords: named entity recognition, text-mining, gene normalization.

Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation

Katrin Fundel, Ralf Zimmer

Teaching and Research Unit Practical Informatics and Bioinformatics, LMU Munich, Amalienstr, München, Germany

We present an integrated system for named entity identification and the results of its application for human gene name normalization. The system builds on extensively curated synonym dictionaries and expands on exact text matching and ProMiner by implementing new modules for abbreviation resolution and disambiguation. The system achieved encouraging results in the BioCreAtIvE challenge.

Chair04 **Lynette Hirschman**

A Hybrid Gene Normalization Approach with Capability of Disambiguation

Jung-Hsien Chiang, **Heng-Hui Liu**

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

Gene normalization is critical for precise biomedical information extraction. We have developed an automatic gene normalization process that takes the output of named entities recognition (NER) systems designed to identify gene mentions and normalizes them to Entrez Gene IDs. Most of gene mentions referred to a unique definition would be normalized by a thesaurus-based procedure using morphological normalization rules. For the rest mentions associated with more than one definition, we propose a hybrid information fusion framework to deal with the ambiguities. An acceptable performance (precision 0.8 and Recall 0.74) was evaluated on 261 articles that BioCreative 2006 provided for training.

Keywords: fuzzy aggregation, gene normalization, maximum entropy classifier, disambiguation.

Exploring Match Scores to Boost Precision of Gene Normalization Precision

Cheng-Ju Kuo, Yu-Ming Chang, Han-Shen Huang, Kuan-Ting Lin, **Bo-Hou Yang**, Yu-Shi Lin, Hun-Nan Hsu, I-Fang Chung

Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan
Institute of Information Science, Academia Sinica, Taipei, Taiwan
Department of Electrical Engineering, Chang-Gang University, Tao Yuan, Taiwan

Gene normalization task is to identify EntrezGene IDs corresponding to the human genes and direct gene products appearing in a given MEDLINE abstract. Given a dictionary that maps gene and protein synonyms to EntrezGene IDs, a naive approach to the problem is to apply a gene mention tagger to identify all potential name entities of genes and then look them up in the dictionary. However, mostly due to the difficulty to compile a complete yet noise-free

dictionary for gene synonyms, the results are far from satisfactory. In our experiments using a gene mention tagger based on a conditional random field (CRF) model and a string matcher based on softTFIDF to look up the dictionary, the F-score is below 0.5. To improve the performance, previous works proposed many methods to clean up dictionaries. These methods may help case by case but may not applicable in general. In this paper, we focus on the problem of whether there exists a systematic method that always improves the result of dictionary lookup. We propose to train an ensemble of classifiers using AdaBoost to recognize true positives from false ones based on the match scores,

which are readily available when anyone applies an approximate string matching function to look up the dictionary. Experimental results show that applying boosting can successfully increase the F-score from about 0.56 to 0.69 with our best F-score reaching 0.75. These results were obtained without modifying the dictionary.

Rule-based Gene Normalization with Statistical and Heuristic Confidence Measure

William Lau, Calvin Johnson

National Institutes of Health, Center for Information Technology, Bethesda, USA

In the gene normalization task, a rule-based approach has certain advantages including the fact that no gold standard is likely to contain all the genes that need to be considered. We have developed a rule-based algorithm that includes pattern matching for gene symbols and an approximate term searching technique for gene names. The algorithm performs confidence estimation by appropriately weighting measures of uniqueness, inverse distance, and coverage. An F-measure of 0.753 has been achieved, using nominal confidence-measure weights.

Keywords: gene normalization, rule-based, approximate term search, confidence measure.

Annotating molecular interactions in the MINT database

Gianni Cesareni

University Rome Tor Vergara, Rome, Italy

The Molecular INTeraction Database (MINT, http://mint.bio.uniroma2.it/mint/) is a relational database storing protein-protein interactions. I will report on the database model and on the stored information. I will highlight aspects of the curation procedure that are relevant for the evaluation of the Biocreative competition results.

Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L. and Cesareni, G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res*, **35**, D557-560.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res*, **35**, D572-574.

Kiemer, L., Costa, S., Ueffing, M. and Cesareni, G. (2007) WI-PHI: A weighted yeast interactome enriched for direct physical interactions. *Proteomics*, **7**, 932-943.

Chair05Alfonso Valencia

Enhancing access to the bibliome for genomics with evaluation tasks derived from user information needs: The TREC Genomics Track

Aaron M. Cohen

Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Oregon, USA

The goal of the Text Retrieval Conference (TREC) Genomics Track, sponsered by the National Institute for Standards and Technology (NIST) is to provide a forum for evaluation of information retrieval systems in the genomics domain. The track has been running since 2003 and has run yearly challenge tasks based on shared biomedical text corpora, with world-wide participation. This year the challenge task is a question answering task based on extracting passages that answer entity-type list information needs. The 2007 year will be the last for the genomics track within TREC, but the organizers are investigating other venues due to the high interest and participation.

Keywords: TREC, NIST, genomics, information retrieval, text mining, NLP.

The application of ontologies in the biological realm

Suzanna Lewis

Lawrence Berkeley National Laboratory, Berkeley, USA

The representation of biological knowledge in databases is a necessity for modern biomedical research. Historically, there has been very little collaboration or coordination between different database providers. Although many grandiose schemes for the “integration” of biological data-bases have been proposed over the years, none have been practical to the point of implementation. Yet the need for integration remains, as many biologists, both those at the bench and those who analyse data computationally, wish to integrate data from a diversity of sources. The Gene Ontology Consortium (GOC) began, some several years ago, to develop a resource that could be used by both the model organism databases (e.g. FlyBase, WormBase, Mouse Genome Da-tabase, The Arabidopsis Information Resource) and the large “horizontal” databases (e.g. Uni- Prot, GeneDB, TIGR Gene Index) as a standard for the annotation of gene products. Out of this experience arose the Open Biomedical Ontologies (OBO) Foundry collaboration whose goal is to produce well- structured vocabularies for shared use across different biological and medical domains. Those involved agree in advance to the adoption of a growing set of principles specifying best practices in ontology development. These principles are designed to foster interoperability to ensure a gradual improvement of quality and formal rigor in ontologies, in ways designed to meet the increasing needs of data and information integration in the biomedical domain.

Keywords: GO, OBO, Ontology, Annotation, PATO, Function.

The Interaction-Article Sub-Task (IAS) evaluation

Martin Krallinger

Structural Computational Biology Group Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

To extract biological annotations from the literature it is crucial to detect first the articles which are relevant for further manual curation. Although this aspect is important for subsequent information extraction steps, it has often been neglected by previously published protein-protein interaction (PPI) extraction systems. Thus the aim of the Interaction Article Subtask (IAS) was to evaluate the automatic detection and ranking of articles relevant to extract protein interaction information, according to the curation criteria followed by interaction annotation databases. Participants were provided with a labeled training collection of relevant and non-relevant PubMed abstracts. For the collection of test set articles, participants where asked to classify and rank them whether they are relevant to extract protein interactions. A total of 19 teams submitted 51 runs for the IAS. Many participating strategies adapted traditional supervised learning techniques to address this problem. The top scoring systems reached an f-score of over 0.78, an area under the ROC curve

(AUC) of around 0.85 and an accuracy of over 0.75.

Automatically Expanded Dictionaries with Exclusion Rules and Support Vector Machine Text Classifiers: Approaches to the BioCreAtIve 2 GN and PPI-IAS Tasks

Aaron Michael Cohen

Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

For BioCreAtIvE 2 we participated in the gene normalization task (GN) and the protein-protein interaction article subtask (PPI-IAS). Our GN submission used automatically extracted and expanded symbol dictionaries, along with manually generated exclusion rules to filter out likely false positives. Our best submission achieved an F1 of 0.724, which placed it in the second quartile. Our best PPI-IAS submission was a "bag of words" linear SVM system with chi-square based feature selection. This system achieved an AUC of 0.8284, which was greater than one standard deviation above the mean. We were able to improve these results slightly by including all features instead of performing the feature selection step. While our submissions performed well, it is likely that these results can be improved with further study. One particularly interesting question is why cross-validation on the PPI-IAS training set grossly overestimates the results achieved on the test collection.

Keywords: named entity recognition, text classification, machine learning, bioinformatics.

A Semi-Supervised Approach to Learning Relevant Protein-Protein Interaction Articles

Mark Greenwood, **Mark Stevenson**

Natural Language Processing Group, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

This paper describes an Information Extraction system that can be used to identify articles containing protein-protein interactions. The approach relies on the automatic acquisition of dependency tree based patterns which can be used to identify these interactions and consequently select relevant documents. Evaluation shows an F-Score performance of approximately 64%.

Keywords: semi-supervised learning, dependency trees, relation extraction, linked chains.

ProtIR Prototype: Abstract Relevance for Protein-Protein Interaction in BioCreAtIvE2 Challenge, PPI-IAS Subtask

Yan Hua Chen[1], Heri Ramampiaro[1], Astrid Lægreid[2], Rune Sætre[3]

[1.] Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway
[2.] Norwegian University of Science and Technology, Trondheim, Norway
[3.] Department of Computer Science, University of Tokyo, Bunkyo-ku, Tokyo, Japan

ProtIR is a prototype developed for the IAS subtask of the BioCreAtIvE2 task, protein-protein interaction (PPI) extraction. This poster gives a description of the skeleton of the system and demonstrates the motivation for the current solution.

Our idea is to adapt information retrieval (IR) techniques for this task to classify and rank a set of abstracts that may contain protein interaction. By using a list of well-known protein interaction related keywords, and a list of protein and gene symbols and names collected from the GeneTools' annotation database of NTNU, we experimented the bag-of-words approach to explorer its advantages and limitations in such a task of biomedical domain. In the phase of recognizing a protein mention, we introduced a name evidence scoring scheme, that use the inverse document frequency (idf) as a weighted factor. By including this factor, the system can easier discriminate between a term in the protein name that are specific to the protein and a term that are not.

The preliminary result was evaluated by BioCreAtIvE, and attained a f-score of 68,2% on the test corpus. We believe that extending the system to integrate sentence context information will improve the performance. Engaging protein information and protein interaction information from other public biomedical databases will most likely enhance the performance.

We compared and analyzed the ROC and precision-recall graphs for abstract relevance prediction using weighted and un-weighted scoring based on protein and connection keyword occurrence, and scoring solely based on connection keyword occurrence. The un-weighted scoring of PPI and the prediction using connection keywords occurrence alone, had the best results. For this, further suggestion for improvement will be to normalize the name evidence score.

A Term Investigation and Majority Voting for Protein Interaction Article Sub-task 1 (IAS)

Man LAN[1], Chew Lim TAN[1], Jian SU[2]

[1.] School of Computing, National University of Singapore, Singapore
[2.] Institute for Infocomm Research, Singapore

The BioCreAtIvE II PPI Interaction Article Sub-task 1 (IAS) is a biological text classification task which concerns whether a given abstract contains protein interaction information. In order to improve the performance of text classification, we examined ways to represent text from the term type and term weighting aspects. In addition, we also combined different classifiers by majority voting technique.

Keywords: biological text classification, text representation, named entity, term weighting.

Identifying Protein-Protein Interaction Sentences Using Boosting and Kernel Methods

Soo-Yong Shin, **Sun Kim**, Jae-Hong Eom, Byoung-Tak Zhang, Ram Sriram

National Institute of Standards and Technology, USA
Biointelligence Laboratory, School of Comp. Sci. & Eng. Seoul National University, Seoul, Republic of Korea

As the amount of biological research literature increases, finding information is becoming a aunting task. Since machine learning techniques could alleviate this problem, we propose a machine learning framework to identify protein-protein interaction sentences from research apers. This machine learning technique is one of the basic components needed to automatically extract biological information from texts. Since the protein-protein interaction (PPI) sentences have their own patterns at article and sentence levels, these patterns are mined by using boosting and kernel methods. Both approaches have good characteristics for the PPI extraction asks, and naturally can handle heuristic information for future extensions.

Keywords: Protein-Protein Interaction Identi¯cation, Boosting Methods, Tree Kernels, Support, Vector Machines.

IntAct - Serving the text-mining community with high quality molecular interaction data

Samuel Kerrien

European Bioinformatics Institute (EBI), European Molecular Biology Laboratory (EMBL), Cambridge, UK

IntAct provides an open source database and toolkit for the storage, presentation and analysis of molecular interactions. High-quality manually annotation of the literature is a time consuming process and coverage of the available interaction data is far from complete. The use of text-mining procedures to highlight appropriate publications and make an initial extraction of interaction data could help to improve both the efficiency of the curation process and the reporting of the data available in the literature. The 2006 BioCreative competition was aimed at evaluating the success of such procedures in comparison to manual annotation.

Keywords: Molecular Interactions, Manual Curation, PSI-MI.

The Interaction-Pair and Interaction Method Sub-Task evaluation

Martin Krallinger

Structural Computational Biology Group Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

To provide useful tools which assist biologists in extracting biological annotations from the literature, several aspects are of importance. A crucial point is the correct identification and association of mentioned interactor proteins to their corresponding database entries (e.g. SwissProt record IDs). Not only the individual interactors, but also the correct binary interaction pair needs to be extracted. Biological annotations of protein interactions are associated to qualitative information in the sense of the interaction detection experiments which have been carried out to characterize the given interaction. Finally, for human interpretation, textual passages which summarize the mentioned interaction are relevant for efficient curation. All these aspects have been addressed in the Protein-Protein Interaction (PPI) task, in the form of several sub-tasks, each focusing in one of the above-mentioned points, namely the Interaction Pair Sub-task (IPS), the Interaction Method Sub-task (IMS) and the Interaction Sentence Sub-Task (ISS).

Teams which extracted normalized protein interaction pairs from full text articles reached an f-score of 0.3. The highest precision obtained for the IPS was of 0.39. When considering the detection of the normalized individual interactor proteins, the highest f-score was of

0.48 with a precision of 0.56. In case of the correct association of full text articles to an ontology of controlled vocabulary terms for interaction detection methods (MI-ontology), the best participant achieved a precision of 0.67. As for the retrieval of the best interaction-summarizing passages, 19% of the passages submitted by one of the teams could be mapped to the previously manually extracted best interaction-describing text passages.

The PPI task covers all the relevant steps for the extraction of protein interaction annotations from full text articles. It shows the main potentials as well as difficulties encountered by participating text mining systems in extracting biological annotations when compared to manual human curation. Especially the protein interactor normalization (without any restriction of the associated organism source), the retrieval of interaction text descriptions which span multiple sentences, as well as implicit difficulties when processing full text articles affected the performance of the participating strategies.

OntoGene in Biocreative II

Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Michael Hess, Jean-Marc von Allmen, Martin Romacker, Therese Vachon

Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
Computational Knowledge Management and Text Mining Unit at Novartis Pharma AG, Basel, Switzerland

We describe in detail the approach that we adopted within the 2nd Biocreative Competition for the PPI-IPS and PPI-IMS tasks. Our approach for PPI-IPS is based on a high-recall protein annotation step, followed by two sharp disambiguation steps. The remaining proteins are then pairwise combined in sentences that are considered “curatable” by a machine learning algorithm. Those sentences are analyzed by a pipeline of NLP tools, including a dependency parser. The results of the pipeline are then used by a number of lexico-syntactic filters, in order to select only the linguistically meaningful interactions. The goal of the approach is to deliver high-precision results, while maintaining a reasonable recall.

The approach adopted for PPI-IMS is based on a pattern matching approach looking for clues of the experimental methods adopted by the experimenters. While some patterns are automatically derived from existing resources, others are manually built, whereby the focus has been put on the methods most frequently used.

In the talk we will describe how the approach was developed on the basis of results obtained on the training data, and we will discuss an analysis of the perfomance on the test data.

GeneTeam Site Report for BioCreative II: Customizing a Simple Toolkit for Text Mining in Molecular Biology

Patrick Ruch

Text Mining, University and Hospitals of Geneva, Geneva, Switzerland

In this technical report, we describe our participation in two of the three BioCreative II tasks:

gene normalization, article selection for protein-protein interaction and protein-protein interactions.

We report on the customization of a simple modular toolkit, which can be applied several text mining applications in molecular biology. The toolkit comprises an automatic generic text categorizer, a retrieval engine and an argumentative classifier, trained to differentiate between PURPOSE, METHODS, RESULTS and CONCLUSION in MEDLINE abstracts. The automatic text categorizer requires a very limited tuning set, and the system keeps most of its effectiveness when tuning data are sparse. We use the categorizer for several subtask: Gene Normalization of ENTREZ-Gene entries, selection/ranking of relevant articles, recognition of Swiss-Prot protein identifier. This last task assumes that we are able to: recognize species, select appropriate sentences,

and finally be able to automatic assign interaction detection methods. The overall results, although still partial at the time of writing this report, show that our toolkit can achieve competitive performances with minimal task customization efforts.

Keywords: text mining, text categorization, protein-protein interaction, database curation, machine learning, information retrieval.

AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask

Rune Sætre[1], Kazuhiro Yoshida[1], Akane Yakushiji[2], Yusuke Miyao[1], Yuichiro Matsubayashi[1], Tomoko Ohta[1]

[1.] Department of Computer Science, University of Tokyo, Tokyo, Japan
[2.] FUJITSU LABORATORIES LTD., Kanagawa, Japan

This report summarizes the participation of the Tsujii-lab group in the 2006 BioCreative2 text mining challenge. It describes the systems used, the results attained, and the lessons learned. The basic idea was to see how well the AKANE system could perform on a full-text Protein- Protein Interaction (PPI) Information Extraction (IE) task. AKANE system is a recently developed, sentence-level PPI system that achieved a 57.3 F-score on the Almed corpus. In order to use the AKANE system for the BioCreative task, the given training data had to be preprocessed. The BioCreative training data contained just a list of interacting protein pair identifiers for each given full-text article, while the expected input for the AKANE system is annotated sentences like in the Almed corpus. In order to transform the full-text articles into Almed sentence-level annotations, the text was first stripped of all HTML coding to get a plain text representation. Then, each mention of protein names were tagged by a Named Entity Recognizer (NER), and all interacting and co-occurring pairs in single sentences were used for training. A pipeline architecture was made to deal with each of these challenges. Some postprocessing was also necessary, in order to transform the results from the AKANE system into the expected format for the BioCreative2 challenge. The postprocessing included filtering and ranking the results, and balancing precision and recall to maximize the F-score.

The poster will describe the AKANE system in more detail, and give some analysis of the errors made.

References: Almed corpus, ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/interactions

Keywords: bionlp, protein-protein interaction, natural language processing.

Consensus Pattern Alignment to find Protein-Protein Interactions in Text

Jörg Hakenberg, Michael Schroeder, Ulf Leser

Technical University Dresden, Humboldt-Universität zu Berlin, Dresden, Germany

"Don't I know you from somewhere?" - comparing new to known texts plays a key role in the system we propose for searching protein-protein interactions. Our system builds on an inexact pattern matching strategy, where patterns reflected the compositional structure of known occurrences of protein-protein interactions in text. To describe this structure, part-of-speech tags (verbs) and entity classes (proteins), words, and word stems were used. From comparable systems proposed before, it became clear that collecting a suitable set of patterns is of major importance, and this step formed the main component of our system. From the IntAct database, we extracted all pairs of proteins known to interact. We scanned PubMed for textual evidences for each such interaction, and retain all single sentences that described them. Using pairwise sentence alignment as a similarity scoring function, we performed a clustering on the resulting set of sentences. Within each cluster, multiple sentence alignment (MSA) identified commonalities and variable positions across all sentences, expressed in a consensus pattern. We could now align such consensus patterns against arbitrary text to extract new protein-protein interactions. Our approach did not need any pre-annotated corpora nor manually pre-defined patterns. The system yields a maximum recall of 69% --which was the best reported among all participating systems--, precision of 45% and F1-measure of 41% on the BioCreative test set.

Chair07 **Carlos Rodriguez**

Identifying Protein-Protein interactions in Biomedical publications

Alejandro Figueroa, GÄunter Neumann

DFKI - LT Lab,
SaarbrÄucken,
Germany

The paper describes the approaches and the results of our participation in the protein-protein interaction (PPI) extraction task (sub-tasks 1 to 3) of the BioCreative II challenge.1 The core of our approach is to analyse the logical forms of those sentences which contain the mentioning of relevant protein names, and to rank the sentences from which the relations where extracted using the class descriptors computed in the sub-task 1 and interaction sentences from the Christine Brun corpus.

Keywords: Protein-Protein interactions identification, Predicate Analysis.

Integrating Knowledge Extracted from Biomedical Literature: Normalization and Evidence Statements for Interactions

Graciela González, Luis Tari, Anthony Gitter, Robert Leaman, Shawn Nikkila, Ryan Wendt, Amanda Zeigler, Chitta Baral

Department of Biomedical Informatics and Department of Computer Science and Eng, School of Computing and Informatics, Fulton School of Engineering, Arizona State University, Tempe, USA

The presentation will report our approach to two specific tasks of the BioCreAtIvE II challenge: protein interaction sentences (PPI-ISS) and protein interaction pairs (PPI-IPS). Our approach to software engineering and implementation decisions was based on addressing first and foremost the core problem of integrating knowledge extracted from the literature: thus, we saw PPI-ISS as pairing statements of certain characteristics to core facts extracted elsewhere in the document and GN as mapping extracted entities to some standard names. This allows us to focus on generic solutions that can then be gradually refined to solving specific problems. In this same spirit, we developed a text-extraction XML format, a query language for the extraction of parse tree constructs, a prototype extraction system, and a prototype web-based generic evaluation system around the BioCreAtIvE challenge that are optimized for broader applications in biomedical text processing.

Keywords: normalization, protein-protein extraction, NLP, ranking, evaluation, data mining.

Chair07 **Carlos Rodriguez**

Mining Physical Protein-Protein Interactions by Exploiting Abundant Features

Minlie Huang, Shilin Ding, Hongning, Wang, Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems (LITS), Department of Computer Science and Technology, Tsinghua University, Beijing, China

It is a great challenge to mine protein-protein interactions from bioscience literature. From a general perspective, there are three sub-tasks to mine biologically meaningful knowledge: first, classify documents containing interactions or not and filter irrelevant ones; second, extract protein-protein interactions (or interacting protein pairs) from the documents; finally, extract detailed information about the interactions, such as experimental detection methods of interactions, and summarization sentences describing them. Particularly, it is the knowledge from the third sub-task that is really meaningful for biologists. In this paper, we present a method of mining physical protein-protein interactions by exploiting abundant features during our participation in the PPI task of BioCreAtIvE Challenge 2006. Several machine learning algorithms for classification and ranking, including SVM and probabilistic model, and abundant features, including strings, unigrams, ontology features, template-like features, and profile-features, have been proposed. Compared with the averaged performance released up to now, our method has shown very promising results.

Keywords: protein-protein interaction, relation extraction, named entity recognition, SVM, kernel.

Uncovering Protein-Protein Interactions in the Bibliome: BioCreative II

Alaa Abi-Haidar, Jasleen Kaur, Ana Maguitman, Predrag Radivojac, Andreas Retchsteiner, Karin Verspoor, Zhiping Wang, **Luis M. Rocha**

School of Informatics, Indiana University, USA
Dep. de Ciencias e Ing. de la Computación, Universidad Nacional del Sur, Argentina
Center for Genomics and Bioinformatics, Indiana University, USA
Modeling, Algorithms and Informatics Group, Los Alamos National Laboratory, USA
Biostatistics, School of Medicine, Indiana University, USA
FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciencia, Portugal

We participated in three of the Protein-Protein Interaction (PPI)subtasks: Protein Interaction Article Subtask 1 (IAS), Protein Interaction Pairs Sub-task 2 (IPS), and Protein Interaction Sentences Sub-task 3 (ISS). Our approach includes a feature detection method based on a Spam-detection algorithm, which identified feature words, bi-grams, and word pairs. For IAS we submitted three runs from distinct classification methods: a novel spam-detection-inspired method (Variable Threshold Protein Mention Model), Support Vector Machines, and an integration method based on the Singular Value Decomposition and measures of uncertainty. For IPS and ISS we used the features discovered from IAS abstracts as well as features from training IPS and ISS data to predict appropriate passages and pairs. We also used the number of protein mentions in a passage as a feature.

Keywords: Protein interaction, text mining, bibliome informatics, support vector machines, singular value decomposition, spam detection, uncertainty measures, proximity graphs, complex networks.

The Interaction-Sentence Sub-Task evaluation

Martin Krallinger

Structural Computational Biology Group Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

To provide useful tools which assist biologists in extracting biological annotations from the literature, several aspects are of importance. A crucial point is the correct identification and association of mentioned interactor proteins to their corresponding database entries (e.g. SwissProt record IDs). Not only the individual interactors, but also the correct binary interaction pair needs to be extracted. Biological annotations of protein interactions are associated to qualitative information in the sense of the interaction detection experiments which have been carried out to characterize the given interaction. Finally, for human interpretation, textual passages which summarize the mentioned interaction are relevant for efficient curation. All these aspects have been addressed in the Protein-Protein Interaction (PPI) task, in the form of several sub-tasks, each focusing in one of the above-mentioned points, namely the Interaction Pair Sub-task (IPS), the Interaction Method Sub-task (IMS) and the Interaction Sentence Sub-Task (ISS).

Teams which extracted normalized protein interaction pairs from full text articles reached an f-score of 0.3. The highest precision obtained for the IPS was of 0.39. When considering the detection of the normalized individual interactor proteins, the highest f-score was of

0.48 with a precision of 0.56. In case of the correct association of full text articles to an ontology of controlled vocabulary terms for interaction detection methods (MI-ontology), the best participant achieved a precision of 0.67. As for the retrieval of the best interaction-summarizing passages, 19% of the passages submitted by one of the teams could be mapped to the previously manually extracted best interaction-describing text passages.

The PPI task covers all the relevant steps for the extraction of protein interaction annotations from full text articles. It shows the main potentials as well as difficulties encountered by participating text mining systems in extracting biological annotations when compared to manual human curation. Especially the protein interactor normalization (without any restriction of the associated organism source), the retrieval of interaction text descriptions which span multiple sentences, as well as implicit difficulties when processing full text articles affected the performance of the participating strategies.

An Integrated Approach to Concept Recognition in Biomedical Text

Lawrence Hunter, William A. Baumgartner, Jr., J. Gregory Caporaso, Helen L. Johnson, Anna Lindemann, Zhiyong Lu, Olga Medvedeva, Jesse Paquette, Elizabeth K. White, K. Bretonnel Cohen

Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, USA

We participated in all three of the BioCreative 2006 tasks. Our approach was characterized by three things: (1) Use of an architecture that allowed us to apply a single, integrated framework to all three tasks; (2) Extensive use of a semantic analysis engine; and (3) use of rule-based approaches to handling coordination of protein names.

We made extensive use of the UIMA (Unstructured Information Management Architecture) framework for integrating almost every component that we used in any BioCreative 2007 task. Three benefits accrued from this strategy: (a) The complete integration of all processing steps allowed us to quickly and easily experiment with different approaches to the many subtasks involved. (b) It made it easy for us to quickly evaluate the results of these experiments against the official data sets. (c) It provided us with a clean interface for incorporating tools from other groups, including LingPipe, ABNER, Schwartz and Hearst's abbreviation detection algorithm, and the Stanford Parser.

Keywords: semantic parsing, conceptual language processing, knowledge-based language processing, direct memory access parsing (DMAP).

Chair08**Patrick Ruch**

From Interaction Mentions to Curatable Interactions

Barry Haddow, Michael Matthews

Language Technology Group, University of Edinburgh, Edinburgh, Scotland

The IPS submission from team 6 made use of the first prototype release of a biomedical information extraction pipeline to identify mentions of protein-protein interactions, together with additional modules to normalise proteins to Uniprot and to identify the curatable interactions from amongst the interaction mentions. The pipeline is being developed to extract information on protein-protein interactions and tissue expression from research papers. The information extraction pipeline includes preprocessing, named entity recognition, term normalisation and relation extraction components. In the preprocessing stage, the text is tokenised, part-of-speech tagged and chunked , and then information about abbreviations and species words is added to the document. The named entity recogniser uses an extended version of the Curran and Clark Maximum Entropy Markov Model NER tagger, trained on data annotated with protein names. The term normaliser aims to map the protein names to identifiers drawn from a RefSeq derived lexicon, using a combination of fuzzy matching rules and machine learning based species tagging, but the output of the term normaliser was not used directly for IPS as the subtask employed a different lexicon. The relation extractor uses a maximum entropy model with a variety of shallow linguistic features, trained on annotated data, to extract mentions of protein-protein interactions from the text.

In order to extend the pipeline for the IPS submission, two extra components were implemented: a Uniprot term normaliser and a curation filter. The role of the Uniprot normaliser was to map the proteins in the interaction mentions output by the pipeline to identifiers drawn from Uniprot, and two different methods were used for normalisation: a string similarity based fuzzy matcher, and an exact matcher which just performed a case-free match. Where multiple matches were found for a given protein the species words occurring in the text were used to choose the most probable match. The role of the curation filter was to pick out the curatable interactions from amongst the interactions predicted by the pipeline. It was implemented using a support vector machine (SVM) model trained on the data provided for IPS, and using features based on how the mentions occurred in the text, the proteins involved in the interaction mentions, and their predicted species.

Keywords: biomedical NLP, relation extraction, named entity recognition, term identification.

Extracting Interacting Protein Pairs and Evidence Sentences by using Dependency Parsing and Machine Learning Techniques

Gunes Erkan, **Arzucan Ozgur**, Dragomir R. Radev

University of Michigan, Department of Electrical Engineering and Computer Science, Ann Arbor, Michigan, USA

We use dependency parsing and machine learning to identify interacting proteins (Protein Interaction Pairs Sub-task 2 (IPS)) and extracting the most relevant sentences that describe their interaction (Protein Interaction Sentences Sub-task 3 (ISS)). The description of our system is as follows. After converting html documents to text, sentence splitting and tokenization is done. Protein names are identified by using the provided release of the SwissProt database as a dictionary. Features are extracted from the dependency parse trees of the sentences, which not only capture syntactic properties, but also the semantic predicate-argument structures. The following features are extracted: binary features representing each interaction word; a binary feature describing whether an interaction word is the parent of a protein pair; each interaction word that is an ancestor of a protein at one or two levels above; and each word that is a parent of a protein. These features are used to train a linear SVM classifier to identify and rank sentences that describe an interaction. To map a protein name to its corresponding UniProtID, candidate organism names and synonyms in the article are matched and weighted according to their proximity to the protein. The frequencies of the organism name appearing just before the protein name, in the same sentence with the protein name, and in the same article with the protein name are considered in descending order of importance. Finally, extracted sentences are mapped back to their html counterparts in two steps by using an approximate string matching algorithm based on edit distance and an approximate token matching algorithm. In the improved version of our system, we extract the paths between a protein pair in the dependency parse tree of a sentence and define two new kernel functions for SVM based on the cosine and edit distance based similarities among these paths.

Keywords: protein-protein interaction, relation extraction, named entity recognition, SVM, kernel.

Protein Interaction Sentence Identification by Using Hierarchical Pattern-Based Approach

Jung-Hsien Chiang, Tsung-Lu Michael Lee, **Yong-Xi Liu**

Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, ROC

Our system is a pattern-based architecture which identifies protein interaction patterns from biomedical literatures. The framework contains protein name recognition step, automated pattern generating step, pattern matching step, and sentence ranking step. The automated pattern generating step scans the positive interaction sentences and automatically constructs patterns based on the results of shallow parsing (chunking). A pattern must consist of a least one interaction keyword and two protein name entities. Our interaction keyword list includes 308 words with different tenses such as binding, binds, bind, and bound.

Moreover, the patterns are built into three levels. From bottom up, the patterns go from specific to more general.

In the automated pattern generating step, hierarchical patterns are computed automatically with selected interaction key words and protein name entities. The sentence ranking procedure ranks each sentence according to the level of matched patterns and the confidence scores of interaction keywords. The hierarchical patterns provide different confidence levels (scores) that can be used to rank our sentences.

Keywords: text mining, information retrieval, protein-protein interactions, bioinformatics.

The RegCreative Jamboree: an experiment in text-mining assisted community annotation

Casey Bergman

University of Manchester, Manchester, UK

Understanding gene expression on a global scale will require computational methods that can decode the cis-regulatory sequences that control transcription. Despite rapid progress in the development of bioinformatic tools to predict cis-regulatory sequences, the field of regulatory bioinformatics is currently limited by a lack of freely accessible, large-scale datasets of functionally characterized regulatory sequences to evaluate such predictive techniques. An abundance of high quality transcriptional regulatory sequence information has been published over the last 25 years, and thus a current challenge in regulatory bioinformatics is how to unlock the vast store of data in the biomedical literature.

Recently, two open-access curation systems (www.oreganno.org, www.pazar.info) have been developed specifically to address this need in the regulatory bioinformatics community, however the problem of how to accelerate the extremely labor intensive task of curating regulatory sequence data from primary literature remains to be solved. Since the primary obstacle to regulatory annotation is fundamentally a text-mining problem, the RegCreative Jamboree was organized in November 2006 to explore the interface between regulatory bioinformatics and text mining. I will present the background to the RegCreative Jamboree and the outcomes of this workshop (including assessment of the size of the regulatory corpus and extracting DNA sequences from text), with an aim towards a future BioCreAtIvE text-mining challenge in the area of regulatory bioinformatics and protein-DNA interactions.

Keywords: Community annotation, regulatory bioinformatics, protein-DNA interactions, information extraction.

Proceedings of the Second BioCreative Challenge Evaluation Workshop

POSTERS *ABSTRACTS*

Feature Engineering and Quick Prototyping of PPI Classifiers

Francisco Carrero García[1], Jose Maria Gomez Hidalgo[1], Manuel Maña Lopez[2], Jacinto Mata Vazquez[2]

[1.] Departamento de Sistemas Informáticos Escuela Superior Politécnica Universidad Europea de Madrid, Madrid, Spain
[2.] Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática Escuela Politécnica Superior Universidad de Huelva Huelva, Spain

One of the most relevant steps in learning-based Text Classification tasks is the modeling of the task, which is the definition of a suitable set of attributes, amenable of being used by effective learning algorithms. In fact, the learning step is conveniently supported by a number of machine Learning libraries like WEKA and others. Our work is focused on the analysis of the most suitable attributes for a number of Text Classification tasks. We have developed a framework and software library, JTLib, which allows together the analysis, modeling and fast prototyping of classification systems, supporting both the experimentation phase and the development of functional system prototypes. The library provides the essentials of Text Classification currently not provided by WEKA, and in fact, it is a complement to it.

This library is being used in two R&D projects, Isis and Sinamed [1], whose objective is to enhance Information Access in the medical domain through the improvement and utilization of Text Classification tasks, like Text Categorization, Automated Text Summarization, and Biological Entity Recognition.

1. Document indexing. After processing the training data, a representation based on the selected attributes is obtained and configured into the WEKA ARFF format. Our set of attributes consists of the most relevant words (unigrams), as well as the most relevant pairs (bigrams) and trios (trigrams) of words. Each n-gram becomes an binary attribute.

2. Dimensionality reduction. During the iterative process, we searched for the n-tuples with higher and lower correlation coefficient to build the attribute vector, and tried with several combinations of amounts of unigrams, bigrams and trigrams.

3. Classifier learning. After several experiments with different Machine Learning algorithms, such as Naïve Bayes, C4.5 decision tree and Adaboost, Adaboost with Naïve Bayes showed to be the most effective. Then, we continued our experiments only with the latter, considering different attributes vectors.

4. Evaluation of text classifiers. The linear increment on the amount of n-tuples used increases precision and F-Measure, but results in a poorer recall. The increment in the amount of bigrams and trigrams produces

a higher recall, but with lower precision and F-measure. The experiments performed with the training set proved that increasing the number of attributes would produce similar results, but not necessarily better.

The participation of our team in the PPI task of the Biocreative competition has primary served us as a proof-of-concept for our systematic approach to feature engineering in text classification tasks. We believe we have obtained reasonable results with respect to the effort we have invested in the competitions, moreover not considering external resources apart from the documents themselves.

[1] Buenaga, M.; Maña, M.; Gachet, D. and Mata, J. The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library. 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006. 548-551.

Enhancing Recall without Degrading too much Precision by Unifying Multiple Backward Parsing Models for Gene Mention Tagging

Han-Shen Huang, Yu-Shi Lin, Kuan-Ting Lin, Cheng-Ju Kuo, **Yu-Ming Chang**, Bo-Hou Yang, I-Fang Chung, Chun-Nan Hsu

Institute of Information Science, Academia Sinica, Taipei, Taiwan
Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan
Department of Electrical Engineering, Chang-Gung University, TaoYuan, Taiwan

Because of a great quantity of biomedical literature has been published, how to extract useful information from these literature becomes very important. Therefore, the named entity task of gene and gene product is a crucial first step for information extraction. We apply two well-known machine learning algorithms, namely, SVM and CRF to evaluate it in Gene Mention Tagging task(GM) of the 2^{nd} BioCreAtIvE. We combine results of two different models to achieve high generalization performance. Results of the 2^{nd} BioCreAtIvE competition in GM show that our approach achieves the 1st quartile among 21 groups.

ProtIR Prototype: Abstract Relevance for Protein-Protein Interaction in BioCreAtIvE2 Challenge, PPI-IAS Subtask

Yan Hua Chen[1], Heri Ramampiaro[1], Astrid Lægreid[2], Rune Sætre[3]

[1.] Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway
[2.] Norwegian University of Science and Technology, Trondheim, Norway
[3.] Department of Computer Science, University of Tokyo, Bunkyo-ku, Tokyo, Japan

ProtIR is a prototype developed for the IAS subtask of the BioCreAtIvE2 task, protein-protein interaction (PPI) extraction. This poster gives a description of the skeleton of the system and demonstrates the motivation for the current solution.

Our idea is to adapt information retrieval (IR) techniques for this task to classify and rank a set of abstracts that may contain protein interaction. By using a list of well-known protein interaction related keywords, and a list of protein and gene symbols and names collected from the GeneTools' annotation database of NTNU, we experimented the bag-of-words approach to explorer its advantages and limitations in such a task of biomedical domain. In the phase of recognizing a protein mention, we introduced a name evidence scoring scheme, that use the inverse document frequency (idf) as a weighted factor. By including this factor, the system can easier discriminate between a term in the protein name that are specific to the protein and a term that are not.

The preliminary result was evaluated by BioCreAtIvE, and attained a f-score of 68,2% on the test corpus. We believe that extending the system to integrate sentence context information will improve the performance. Engaging protein information and protein interaction information from other public biomedical databases will most likely enhance the performance.

We compared and analyzed the ROC and precision-recall graphs for abstract relevance prediction using weighted and un-weighted scoring based on protein and connection keyword occurrence, and scoring solely based on connection keyword occurrence. The un-weighted scoring of PPI and the prediction using connection keywords occurrence alone, had the best results. For this, further suggestion for improvement will be to normalize the name evidence score.

Multi Model Approach for Alternative Taggings

Roman Klinger, **Christoph M. Friedrich**, Juliane Fluck

Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany

One characteristic in BioCreative 2006 compared to common NER tasks is that the training data contains acceptable alternatives for gene and protein names next to the gold standard. We address that problem with a multi model approach using Conditional Random Fields. Because of ambiguities in the allocation of annotations to the gold standard file (GENE.eval) and the acceptable alternatives file (ALTGENE.eval) we build two different training files which have different annotation lengths. Having different annotations from the two models built on these training files, we use different combination strategies resulting in the possibility to influence the ratio between recall and precision.

For selecting our model 50-fold bootstrapping is used to avoid the selection of overfitted models. Parameters analysed are e.g. tokenisation details and features to extract from text. We use a rich set of morphological features (some automatically generated), dictionaries, offset-conjunction and part-of-speech/shallow parsing information from the GeniaTagger. Additionally we use the tagging information of the ProMiner, a normalising tagger achieving a very high precision. The combination to a resulting tagging is followed by a postprocessing step correcting bracket and quotation errors. Additionally a concept study using latent semantic analysis for acronym disambiguation has been introduced, which affects 4 highly ambiguous acronyms. This disambiguation concept works here only at the sentence level but can be shown to be more powerful, if the full sentence context will be available. We achieve an F-score of 86.33 % with a precision of 87.27 % and a recall of 85.41 % on the test set.

From Interaction Mentions to Curatable Interactions

Barry Haddow, Michael Matthews

Language Technology Group, University of Edinburgh, Edinburgh, Scotland

The IPS submission from team 6 made use of the first prototype release of a biomedical information extraction pipeline to identify mentions of protein-protein interactions, together with additional modules to normalise proteins to Uniprot and to identify the curatable interactions from amongst the interaction mentions. The pipeline is being developed to extract information on protein-protein interactions and tissue expression from research papers.

The information extraction pipeline includes preprocessing, named entity recognition, term normalisation and relation extraction components. In the preprocessing stage, the text is tokenised, part-of-speech tagged and chunked , and then information about abbreviations and species words is added to the document. The named entity recogniser uses an extended version of the Curran and Clark Maximum Entropy Markov Model NER tagger, trained on data annotated with protein names. The term normaliser aims to map the protein names to identifiers drawn from a RefSeq derived lexicon, using a combination of fuzzy matching rules and machine learning based species tagging, but the output of the term normaliser was not used directly for IPS as the subtask employed a different lexicon. The relation extractor uses a maximum entropy model with a variety of shallow linguistic features, trained on annotated data, to extract mentions of protein-protein interactions from the text.

In order to extend the pipeline for the IPS submission, two extra components were implemented: a Uniprot term normaliser and a curation filter. The role of the Uniprot normaliser was to map the proteins in the interaction mentions output by the pipeline to identifiers drawn from Uniprot, and two different methods were used for normalisation: a string similarity based fuzzy matcher, and an exact matcher which just performed a case-free match. Where multiple matches were found for a given protein the species words occurring in the text were used to choose the most probable match. The role of the curation filter was to pick out the curatable interactions from amongst the interactions predicted by the pipeline. It was implemented using a support vector machine (SVM) model trained on the data provided for IPS, and using features based on how the mentions occurred in the text, the proteins involved in the interaction mentions, and their predicted species.

IntAct- The Molecular Interaction Database

Jyoti Khadake, B. Aranda, C. Derow, S. Kerrien, J. Khadake, C. Leroy, L. Montecchi-Palazzi, S. Orchard, J. Risse, D. Thorneycroft, R. Apweiler, H. Hermjakob

EMBL Outstation—European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus Hinxton, Cambridge, UK

IntAct provides an open source database and toolkit for the storage, presentation and analysis of protein interactions. The data is manually curated from both existing literature and direct submissions received prior to publication. High quality manual curation is time consuming and effective text-mining would improve both the efficiencyof the curation process and our coverage of the data available in literature.. IntAct uses the UniProtKB as a common platform for the identification of proteins allowing for the disambiguation of the identifiers used throughout the literature. The IntAct web interface provides both textual and graphical representations of protein interactions, and allows the exploration of interaction networks in the context of the GO annotations and InterPro signatures of the interacting proteins. Data can be retrieved in both PSI-MI XML2.5 or a, tab-delineated text format MITAB2.5 [ref]. IntAct currently contains 130,000 binary and complex interactions, making intensive use of controlled vocabularies to ensure consistency of data annotation. All IntAct software, data and controlled vocabularies are available at http://www.ebi.ac.uk/intact.

Integrating Knowledge Extracted from Biomedical Literature: Gene Normalization

Graciela Gonzalez, **Robert Leaman**, Amanda Zeigler, Chitta Baral

Department of Biomedical Informatics & Department of Computer Science and Eng, School of Computing and Informatics, Fulton School of Engineering, Arizona State University, Tempe, USA

We present a multi-stage metric-based gene normalization system which utilizes an existing entity recognition system (ABNER) to find mentions of genes and their products. The mentions found are filtered to improve specificity and are expanded using the Stanford Biomedical Abbreviation database and their provided Java code. For each mention, a similarity score is computed for each gene name in the list of standard gene names, and the system is architected to allow efficient computation. This similarity score is based on the Dice coefficient, which reflects the number of tokens which are contained in both the gene mention and the standard gene name, scaled to reflect the lengths of both, and is similar to Jaccard except that it gives twice the weight to agreements. Several modifications to the standard calculation are applied, including applying weights to the tokens according to their frequency in the training data so as to limit the effect of high-frequency terms. Also, tokens from the gene mention which are not found in the list of standard gene names are matched approximately. Potential matches with a similarity lower than an empirically-determined threshold are discarded. The gene for the most similar name found for each mention is output unless determined to be a duplicate using a heuristic based on the idea that a single abstract sometimes refers to the same gene by different names.

[1] B. Settles, “ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text,” Bioinformatics, vol. 21, pp. 3191-3192, 2005.

[2] J. D. Wren, J. T. Chang, J. Pustejovsky, E. Adar, H. R. Garner, and R. B. Altman, “Biomedical term mapping databases,” Nucl. Acids Res., vol. 33, pp. D289-293, 2005.

[3] L. Egghe and C. Michel, “Strong similarity measures for ordered sets of documents in information retrieval,” Information Processing and Management, vol. 38, pp. 823-848, 2002.

[4] M. Bilenko and R. Mooney, “Adaptive Duplicate Detection Using Learnable String Similarity Measures”, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pp.39-48, 2003

XXXXXXXXXXX XXXXXXXXXXXXXX XXXXXXXXXX

Hamis Mwessa

Oncological laboratory, Masunzu Medical Store, Dar es salaam, Tanzania

The project is two-fold: a regional part and a African part. The regional part consists of a demonstrator in Paediatrics. The European part consists of a series of studies on existing Telematics services and infrastructures, on ongoing projects of potential interest, organisational and security aspects of a African Cancer telematics network, and a large study to determine the requirements of the Oncological community, leading to a demonstrator of Telematics services build on the existing AFRICODE network

Many valuable computer tools for cancer research, treatment, education, screening and prevention have been developed in the past decade. Parallel to this, the Internet community has produced generic software tools for the exchange and dissemination of information in a worldwide accepted and technically outstanding fashion. TeleSCAN, the African demonstrator within the Masunzu project, aims at the synthesis of these achievements, and by providing these tools to the oncology community, mobilising.

A Study for Application of Discriminative Models in Biomedical Literature Mining

Chengjie Sun

ITNLP Lab.,
Harbin Institute of Technology, Harbin, China

By automatically identifying gene and protein names and mapping these to unique database identifiers, it becomes possible to extract and integrate information from a large amount of biomedical literature. This paper presents the attempts of use discriminative models to automatically detect gene name mention and normalize gene mentions. Conditional Random Fields model is adopted to solve gene mention task and Maximum Entropy model is used to do gene mention normalization task. The evaluation results of biocreative2006 are also reported.

Keywords: discriminative model, conditional random field, maximum entropy, text mining.

Gene/Protein Name Detection Using Online Resources

Manabu Torii, Hongfang Liu, Zhangzhi Hu, Cathy Wu,

Georgetown University Medical Center, Washington DC, USA

We developed a gene/protein name recognition system exploiting online terminology sources as well as existing natural language processing tools for BioCreAtIvE 2 Gene Mention task.

Our recognition method consists of three steps. The first step is dictionary-lookup where terms in the text are looked up in terminology sources, BioThesaurus [3] and UMLS [1]. The second step is machine learning that integrates the dictionary knowledge, part of speech (POS) information, and contextual clues for gene/protein name recognition. The POS information was obtained using GENIA tagger [6]. As the machine learning component, Conditional Random Field (CRF) implementation of Mallet [4] was used. The last step in the system is post-processing that corrects apparent errors and incorporates acronym/abbreviation information [5].

We submitted three runs. For the first submission, we used the base system above. With the aim to improve the recall measure, we considered two ad-hoc approaches. For the second run, we retrieved one or more MEDLINE abstract (if found) for each excerpt, from which the excerpt may be originated. The base system was applied to the abstracts, rather than to excerpts, in order to exploit different occurrences of gene/protein phrases in the abstracts. For the third submission, noticing that some of the (true) gene/protein phrases initially mapped to BioThesaurus entries were (falsely) untagged by the machine learning component in the base system, we considered regaining phrases initially mapped to BioThesaurus. To that end, besides the base system, we applied another recognition system, a long-distance character language model-based chunker provided in the LingPipe suite [4], and phrases recognized by both the dictionary look-up component and this recognition system were supplemented in the submission. Our system achieved the recall as high as 89.3% (precision 82.7%) for the third submission, and the precision as high as 85.7% (recall 84.8%) for the first submission.

[1] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res, 32, D267-270, 2004.

[2] LingPipe. http://www.alias-i.com/lingpipe/.

[3] Liu H, Hu ZZ, and Wu C. BioThesaurus: a thesaurus of gene and protein names. BioInformatics, Jan 1;22(1):103-5, 2006.

[4] McCallum AK. MALLET: A Machine Learning for Language Toolkit, 2002. http://mallet.cs.umass.edu.

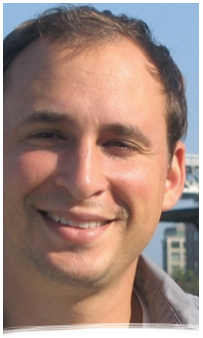
[5] Schwartz A and Hearst M. A simple algorithm for identifying abbreviation definitions in biomedical texts. Pacific Symposium on Biocomputing, 2003.

[6] Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, and Tsujii J. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005.

Proceedings of the Second BioCreative Challenge Evaluation Workshop

INVITED SPEAKERS *PORTFOLIO*

Casey Bergman



Casey Bergman is a Lecturer in Bioinformatics at the University of Manchester, and previously did post-doctoral research at the

University of Cambridge and the Berkeley Drosophila Genome Project.

His primary interests are in understanding the function and evolution of noncoding DNA sequences with an emphasis on cis-regulatory sequences that control transcription. Dr. Bergman has been awarded fellowships in Bioinformatics from the NSF and Royal Society, and is a contributing member of the Faculty of 1000 subsection in Bioinformatics. He has been involved in several large-scale curatorial efforts to annotate noncoding sequences in the Drosophila genome, including the development of the flyreg database of transcription factor binding sites, which was a founding dataset in the Open Regulatory Annotation (www.oreganno.org) database. Together with an international consortium of researchers in North America and Europe, Dr. Bergman co-organized the RegCreative Jamboree in November 2006, which brought together researchers in regulatory bioinformatics and text mining to explore the interface between these two emerging disciplines.

Suzanna Lewis

Suzanna Lewis is a Staff Scientist at Lawrence Berkeley National Laboratory (LBNL). She is the PI for the Bioinformatics Core of the National Center for Biomedical Ontologies and a PI of the Gene Ontology (GO) Consortium, of which she is one of the founders. Ms.

Lewis has 20 years of extensive experience in the development and implementation of large genome informatics programs, particularly in coordinating multiple groups. She directed the Berkeley Drosophila Genome Project (BDGP) Informatics group for over ten years, managing and coordinating software developers, biological curators, students, and post-doctoral researchers. Her group was responsible for designing and developing much of the software used by FlyBase, the Drosophila model organism database. She is one of the founders of the Generic Model Organism Database project, a critical component of genomics infrastructure. She is an organizer of the cold Spring Harbor Genome Informatics conference and an instructor for the CSH course “Programming for Biologists”. Ms. Lewis has dedicated her career to providing the scientific community with complete access to the information that is essential for pursuing their research. In 2000, Ms. Lewis was a co-winner of the American Association for the Advancement of Science Newcomb Cleveland Prize for the most cited paper of the year, “The genome sequence of D. melanogaster”. Her contributions to the field were acknowledged in 2005 when she was elected a Fellow of the American Association for the Advancement of Science.

W. John Wilbur

John Wilbur is a Senior Investigator in the Computational Biology Branch of the National Center for Biotechnology Information. He is a principal investigator leading a research group in the study and development of statistical text processing algorithms. While at NCBI he developed the algorithm that produces PubMed related documents and the algorithm that in PubMed allows fuzzy phrase matching. His group has developed algorithms for phrase identification in natural language text that are used in NCBI's electronic textbook project and allow for easy reference from MEDLINE documents to related textbook material. Most recently he has developed a spell checking algorithm that is used to offer suggestions for correcting PubMed queries. He has a strong interest in machine learning and natural language processing techniques and one focus of current research is improvements in named entity recognition in the field of molecular biology and medicine.

Lynette Hirschman

Lynette Hirschman is Chief Scientist and Director, Biomedical Informatics for the Information Technology Center at the MITRE Corporation in Bedford, MA. Dr. Hirschman is responsible for overseeing MITRE's research activities in Biotechnology, and is leading research projects on "Genomics for Bioforensics" and the NSF-funded BioCreative: Critical Assessment for Information Extraction in Biology. She is a founding organizer of the ISMB BioLINK SIG for text mining in biology and has organized workshops or sessions on biomedical natural language processing for the Association of Computational Linguistics, the Pacific Symposium on Biocomputing, as well as the KDD Challenge Cup in 2002. She received a B.A. in Chemistry from Oberlin College in 1966, a M.A. in German Literature from University of California, Santa Barbara, in 1968, and a Ph.D. in formal linguistics from University of Pennsylvania in 1972, under Aravind Joshi. She is the author of over 100 publications in the areas of natural language processing, speech understanding, logic programming, human-computer interaction and bioinformatics.

Junichi Tsujii

I am Professor of Computational Linguistics and Natural Language Processing of the University of Tokyo and Professor of Text Mining of the University of Manchester, UK. My first degree, from Kyoto University, was in Electronics and was awarded MSc and PhD from Department of Electrical Engineering, Kyoto University. I was Associate Professor of Kyoto University from 1979 to 1988 with a break at CNRS Grenoble, France as invited senior researcher in 1981-1982.

Before taking up the position at the University of Tokyo, I was Professor of Computational Linguistics of University of Manchester Institute of Science and Technology (UMIST) from 1988 to 1995. I was recently appointed Director of National Center for Text Mining

(NacTeM)Å@of the UK in Manchester (2005) and Professor of Text Mining of the University of Manchester. I have been permanent member of International Committee of Computational Linguistics (ICCL) from 1994 as well as Vice-President (2005) and President (2006) of the Association for Computational Linguistics (ACL).

My research interests include grammar formalisms, efficient and effective deep parsing, machine learning for natural language processing, and their application to Text Mining in biology. The research groups in the University of Tokyo and NaCTeM (the University of Manchester) have been successful in developing new ideas such as

(1) Feature forest model for estimating ME (Maximum Entropy) parameters for feature structured objects (FSOs), an extension of the Inside-Outside algorithm-type of dynamic programming to deal with FSOs,
(2) Search algorithms for efficient statistical parsing,
(3) Improved maximum entropy estimator (ME with inequality constraints) which is suited for text mining applications such POS tagger, text classification, NER (Named Entity Recognizer), etc., (4)CFG filtering algorithms for efficient parsing for LTAG, HPSG, etc.

The two teams of Tokyo and Manchester have been applying these new theories and techniques to Text Mining in Biology. The achievements so far includes

(1) TerMine: Term management system based on C-value
(2) AcroMine: Statistics-based acronym discovery system which associates acronyms with their definitions
(3) MEDIE: Intelligent Text Management system using a deep parser (Enju), which stores and indexes the whole Medline
(4) Info-Pubmed: Information Extraction and Retrieval system focusing on Protein-Protein Interaction
(5) GENIA Tool kit: NLP tool kit which are highly adaptable, includes a tagger, a chunker, several NERs, a shallow parser and a deep parser
(Enju)
(6) GENIA Corpus: Richly annotated corpus of Medline abstract (2,000 abstracts), multi-layer annotations include POSs, NEs, Syntactic Trees, Semantic structures (PASs), Co-references and Events.

These NLP tools and systems are available at http://www.nactem.ac.uk/index.php and http://www-tsujii.is.s.u-tokyo.ac.jp/index.html.

Martin Krallinger

Martin Krallinger is currently working at the Spanish National Cancer Center it the group of Prof. Alfonso Valencia. Previous research stays included the National Biotechnology Center (CNB, Madrid, Spain), the Center of Applied Molecular Engineering (CAME, University of Salzburg, Austria), the Center of Surface Biotechnology (Biomedical Center, Uppsala, Sweden) and the former Institute of Molecular Biology of the Austrian Academy of Sciences. His main research interests are related to text mining, information extraction and information retrieval applied to the biomedical and molecular biology literature. He is currently reviewing articles for journals such as Bioinformatics or IEEE Transactions on Information Technology in Biomedicine and served in the Programme Committee of numerous conferences and workshops. He has published articles relevant for text mining in the biology domain in journals such as Genome Biology and Science STKE as well chapters in books such as “Bioinformatics: From Genomes to Therapies (VCH-Wiley). In order to introduce and promote the use of bioinformatics and text mining resources by biologists, he was also in charge of lectures related to information extraction applications for biology at several Spanish universities.

Alfonso Valencia

Alfonso Valencia obtained his Ph.D. in Biochemistry and Molecular Biology in 1981 (Instituto de de Investigaciones Biomédicas - CSIC / Department of Biochemistry, Fac. Medicine U. Autónoma Madrid) for work in the area of Biophysics and protein modeling. From 1988-1994 Alfonso Valencia was a postdoctoral fellow and Visitor Scientist in the European Molecular Biology Laboratory (Heidelberg) working in Dr. Chris Sander's "Protein Design Group". His postdoctoral work was dedicated to the development of methods for protein structure prediction based on the exploration of correlations in multiple sequence alignments (correlated mutation analysis), and for function prediction based on the differential comparison of sequence conservation in protein families (sequencespace method). His work also included the exploration of the sequence / structure space in the ras-p21 and actin-hsc70-FtsA protein families. During this time they also developed was can be considered the first automatic genome annotation method (Genequiz). In 1994 he returned to Spain to form his own group at the Spanish National Biotechnology Centre (CNB-CSIC) to work on the computational analysis of protein families. At that time, the group developed in depth collaborations with experimental groups applying protein modeling and genome analysis techniques to families such as FtsA. In 1998 the group developed the first application of text mining techniques in the area of molecular biology, a work that continued with the publication in 2002 of the first application of text mining to the extraction of protein interactions, that constitutes the basis for

the much of the current development in the area of biological text mining.

The group is well known in the field by its activity as assessor in the protein structure competition (CASP) and by being the organizer of the main text mining challenge (BioCreative).

In 2004 Alfonso Valencia was elected as director of the Spanish Bioinformatics Institute (INB) organized by the Genome Spain foundation.

The INB is a large collaborative programme in which 10 groups provide the bioinformatics infrastructure to support the National genomics projects. The INB will continue its activity until 2009 after the successful evaluation of its activity during the period (2004-06).

Valencia moved to the Centro Nacional de Investigaciones Oncológicas (CNIO) in 2006 as director of the Structural Biology and Biocomputing Programme. His mission as director of the CNIO's Programme is to organize a combined computational and structural approach to study the molecular basis of cancer processes, building on the possibilities offered by current high-throughput genomics approaches, and collaborating with the other CNIO's groups.

Alfonso Valencia is a CSIC Research Professor and EMBO member since 2005. He is a founder, former Vice-President and member of the board of International Society for Computational Biology (ISCB). He has been chair of the Systems Biology and/or Text Mining tracks of the main Computational Biology Annual Conference (ISMB) since 2003. He is also founder of the organization behind the European Conference of Computational Biology for which he co-organized the European annual conference in 2005. He was member of the steering committee of the ESF programme on ·"Functional Genomics" (2000-2005) and since 2006 he is co-director of the new "Frontiers of Functional Genomics" five-years ESF program. He serves in the EMBL and BioZemtrum Scientific Advisory Committees. Alfonso Valencia is co-Executive Editor of "Bioinformatics" published by Oxford University Press, that is the main journal in the field. Among many other grants it is worth mentioning the participation of the group in the three main VI Framework Programme European Networks in Bioinformatics / Computational Biology (Biosapiens, EMBRACE and ENFIN). Alfonso Valencia published his first paper in Computational Biology in 1986, since them he has published more than 160 papers (H factor 40) and more than 20 invited reviews and book chapters.

Proceedings of the Second BioCreative Challenge Evaluation Workshop

LIST OF INVITED SPEAKERS *AND PARTICIPANTS*

List of Invited Speakers and Organizers

Casey Bergman
casey.bergman@manchester.ac.uk
Manchester, UK, University of Manchester

Gianni Cesareni
cesareni@uniroma2.it
University Rome Tor Vergata, Rome, Italy

Aaron M. Cohen
cohenaa@ohsu.edu
Oregon Health & Science University, Oregon, USA

Matthew Day
M.Day@nature.com
Database Publisher Nature Publishing Group, London, UK

Lynette Hirschman
lynette@mitre.org
Biomedical Informatics for the Information Technology Center (MITRE Corporation), Bedford, USA

Samuel Kerrien
skerrien@ebi.ac.uk
EMBL Outstation, European Bioinformatics Institute, Wellcome Trust, Cambridge, UK

Martin Krallinger
mkrallinger@cnio.es
Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain

Suzanna Lewis
suzi@berkeleybop.org
Lawrence Berkeley National Laboratory (LBNL), Berkeley, USA

Junichi Tsujii
tsujii@is.s.u-tokyo.ac.jp; j.tsujii@manchester.ac.uk
University of Tokyo and University of Manchester, Tokyo/Manchester, Japan/UK

List of Invited Speakers and Organizers

Alfonso Valencia
avalencia@cnio.es
Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain

W. John Wilbur
wilbur@ncbi.nlm.nih.gov
Computational Biology Branch of the National Center for Biotechnology Information (NCBI, NIH), Bethesda, USA

List of Participants

B

Beatrice Alex
balex@inf.ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom

Rie Ando
rie1@us.ibm.com
IBM, Hawthorne, USA

Christian Blaschke
blaschke@bioalma.com
Alma Bioinformatics, SL, Tres Cantos, Madrid, Spain

Paul Boddie
paul.boddie@biotek.uio.no
University of Oslo, Oslo, Norway

Pedro Carmona-Saez
pcarmona@cnb.uam.es
CNB, Madrid, Spain

Francisco Manuel Carrero
francisco.carrero@uem.es
Universidad Europea de Madrid, Villaviciosa de Odón, Spain

Monica Chagoyen
monica.chagoyen@cnb.uam.es
Centro Nacional de Biotecnologia - CSIC, Madrid, Spain

Yu-Ming Chang
porter@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica, Taiwan, ROC. Taipei, Republic of China (Taiwan)

Yifei Chen
yifechen@vub.ac.be

Vrije Universiteit Brussel, Brussels, Belgium

Yan Hua Chen
yanhua@idi.ntnu.no
Norwegian University of Science and Technology, Trondheim, Norway

Andreas Doms
adoms@biotec.tu-dresden.de
Biotec/Dept. of Computing, TU Dresden, Dresden, Germany

Jae-Hong Eom
jheom@bi.snu.ac.kr
Seoul National University, Seoul, Republic of Korea

Alejandro Figueroa
alejandro.figueroa.a@gmail.com
DFKI, Saarbruecken, Germany

Juliane Fluck
juliane.fluck@scai.fraunhofer.de
Fraunhofer Institute SCAI, Sankt Augustin, Germany

Christoph M. Friedrich
friedrich@scai.fhg.de
Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany

Katrin Fundel
katrin.fundel@bio.ifi.lmu.de
LMU, München, Germany

Kuzman Ganchev
kuzman@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA United States

List of Participants

B

Julien Gobeill
julien.gobeill@sim.hcuge.ch
Hôpitaux Universitaires de Genève, Geneve, Suisse

Jose Maria Gomez Hidalgo
jmgomez@uem.es
Universidad Europea de Madrid, Villaviciosa de Odon, (Madrid), Spain

Graciela Gonzalez
graciela.gonzalez@asu.edu
Arizona State University, Tempe, United States

Barry Haddow
bhaddow@inf.ed.ac.uk
University of Edinburgh, Edinburgh, Scotland

Jörg Hakenberg
hakenbergj@biotec.tu-dresden.de
Technical University Dresden, Dresden, Germany

Chun-Nan Hsu
chunnan@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica, Taiwan Nankang, Taipei, Taiwan

Minlie Huang
aihuang@tsinghua.edu.cn; minlie.huang@hotmail.com
Tsinghua University, Beijing, China

Lawrence Hunter
HunterOnSabbatical@gmail.com
University of Colorado School of Medicine, Aurora, USA

Sophia Katrenko
katrenko@science.uva.nl

University of Amsterdam, Amsterdam, the Netherlands

Jyoti Khadake
jyoti@ebi.ac.uk
EMBL-EBI Hinxton, Cambridgeshire, United Kingdom

Sun Kim
skim@bi.snu.ac.kr
Seoul National University, Seoul, South Korea

Roman Klinger
roman.klinger@scai.fhg.de
Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany

Cheng-Ju Kuo
cju.kuo@gmail.com
Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan

Man Lan
lanman.sg@gmail.com
National University of Singapore, Singapore, Singapore

William Lau
lauwill@mail.nih.gov
National Institutes of Health, Center for Information Technology, Bethesda, USA

James Leaman
bob.leaman@asu.edu
Arizona State University Tempe, Arizona, United States

Florian Leitner
fleitner@cnio.es
Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain

List of Participants

B

Ulf Leser
leser@informatik.hu-berlin.de
Humboldt-Universität zu Berlin, Berlin, Germany

Heng-Hui Liu
liuhh@cad.csie.ncku.edu.tw
National Cheng-Kung University, Tainan, Taiwan

Hongfang Liu
hl224@georgetown.edu
Georgetown University Washington, DC, USA

ThaiBinh Luong
thaibinh.luong@yale.edu
Yale University New Haven, CT, United States

Manuel J. Maña López
manuel.mana@diesia.uhu.es
Universidad de Huelva, Huelva, Spain

Julio Martinez
martinez@bioalma.com
Alma Bioinformatics, SL, Tres Cantos, Madrid, Spain

Jacinto Mata Vázquez
mata@uhu.es
Universidad de Huelva, Huelva, Spain

Michael Matthews
m.matthews@ed.ac.uk
University of Edinburgh, Edinburgh, UK

Antonio Molina Marco
amolina@dsic.upv.es

Universidad Politécnica de Valencia, Valencia, Spain

Hamis Mwessa
mwessa2003@yahoo.com
Masunzu Medical Store Dar es salaam, Tanzania

Goran Nenadic
g.nenadic@manchester.ac.uk
University of Manchester, Manchester, UK

Guenter Neumann
neumann@dfki.de
DFKI, Saarbruecken, Germany

Mariana Neves
marianaln@hotmail.com
Universidad Complutense de Madrid, Madrid, Spain

Arzucan Ozgur
ozgur@umich.edu
University of Michigan Ann Arbor, Michigan, United States

Ruch Patrick
patrick.ruch@sim.hcuge.ch
University and Hospitals of Geneva, Geneva, Switzerland

Ferran Pla Santamaría
fpla@dsic.upv.es
Universidad Politécnica de Valencia, Valencia, Spain

Conrad Plake
conrad.plake@biotec.tu-dresden.de
Biotechnological Centre of TU Dresden, Dresden, Germany

List of Participants

B

Richard Povinelli
richard.povinelli@marquette.edu
Marquette University Milwaukee, WI, United States

Fabio Rinaldi
rinaldi@ifi.unizh.ch
University of Zurich, Zurich, Switzerland

Luis Rocha
rocha@indiana.edu
Indiana University, School of Informatics, Bloomington, USA

Carlos Rodriguez Penagos
crodriguezp@cnio.es
Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain

Loic Royer
royer@biotec.tu-dresden.de
Biotec, TU-Dresden, Dresden, Germany

Rune Saetre
satre@idi.ntnu.no
University of Tokyo, Tokyo, Japan

Pablo Sánchez
sanchez@bioalma
Alma Bioinformatics, SL, Tres Cantos, Madrid, Spain

Michael Schroeder
ms@biotec.tu-dresden.de
TU Dresden, Dresden, Germany

Martijn Schuemie
m.schuemie@erasmusmc.nl

Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands

Hagit Shatkay
shatkay@cs.queensu.ca
School of Computing, Queen's University Kingston, Ontario, Canada

Mark Stevenson
marks@dcs.shef.ac.uk
University of Sheffield, Sheffield, United Kingdom

Craig Struble
craig.struble@marquette.edu
Marquette University Milwaukee, WI, United States

Jian Su
sujian@i2r.a-star.edu.sg
Institute for Infocomm Research, Singapore, Singapore

Chengjie Sun
cjsun@insun.hit.edu.cn
Harbin Institute of Technology, Harbin, China

Manabu Torii
mt352@georgetown.edu
Georgetown University Medical Center Washington, DC, USA

Rafael Torres
torres@bioalma.com
Alma Bioinformatics, SL, Tres Cantos, Madrid, Spain

Richard Tzong-Han Tsai
thtsai@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica, Taipei, Taiwan

List of Participants

B

Miguel Vazquez
miguel.vazquez@fdi.ucm.es
Universidad Complutense de Madrid, Madrid, Spain

Andreas Vlachos
av308@cl.cam.ac.uk
University of Cambridge, Cambridge, UK

Xinglong Wang
xwang@inf.ed.ac.uk
University of Edinburgh, Edinburgh, Scotland

John Wilbur
wilbur@ncbi.nlm.nih.gov
National Institutes of Health Bethesda, Maryland, USA

Bo-Hou Yang
ericyang@iis.sinica.edu.tw
Institute if Information Science, Academia Sinica, Taipei, Taiwan

Isik Yulug
yulug@fen.bilkent.edu.tr
Bilkent University, Ankara, Turkey

List of Participants

Calendar of CNIO**ACTIVITIES 2007**

JANUARY

Translating Oncoproteomics into Clinical Applications

Organizers: Ignacio Casal (CNIO, Madrid, Spain); Marta Sánchez-Carbayo (CNIO, Madrid, Spain)

Date: January 15, 2007

FEBRUARY

Molecular mechanisms in Lymphoid Neoplasm - CNIO Cancer Conference

Organizers: E. Campo (Hospital Clinic, Barcelona), R. Dalla Favera (Columbia University, New York), E. Jaffe (NCI, Bethesda), M. A. Piris (CNIO)

Dates: February 19 - 21, 2007

MARCH

Conference, Workshop & Exhibition "High Content Analysis Spain 2007"

Organizers: Alberto Álvarez (CNIC), Wolfgang Link (CNIO), Enrique O´Connor (CIPF-UVEG)

Dates: March 26-27, 2007

APRIL

Second BioCreAtive Challenge Evaluation: Assessment of Text Mining Methods in Molecular Biology- Frontiers of Functional Genomics (ESF Activities)

Organizers: Alfonso Valencia (CNIO), Martin Krallinger (CNIO) Lynette Hirschman (MITRE)

Dates: April, 23-25, 2007

JUNE

Myc and the Transcriptional Control of Proliferation and Oncogenesis - CNIO Cancer Conference

Organizers: R. N. Eisenman (Fred Hutchinson Cancer Research Center, Seattle), M. Eilers (Univ. Marburg), J. León (Univ. Cantabria, Santander)

Dates: June 11-13, 2007

OCTOBER

Nature-CNIO Conference - Oncogenes and Human Cancer: The Next 25 years

Organizers: M. Barbacid, E.Hutchinson, D. Lane, C. Marshall, B. Marte, F. McCormick, B. Pulverer, C. Sawyers, K. Vousden, R. Weinberg

Dates: October 3-6 2007

NOVEMBER

Links Between Cancer, Replication Stress and Genomic Integrity- CNIO Cancer Conference

Organizers: O. Fernández-Capetillo (CNIO, Madrid), J. Lukas (DCS, Copenhagen), J. Méndez (CNIO, Madrid), A. Nussenzweig (NCI/NIH, Bethesda)

Dates: November 5-7, 2007