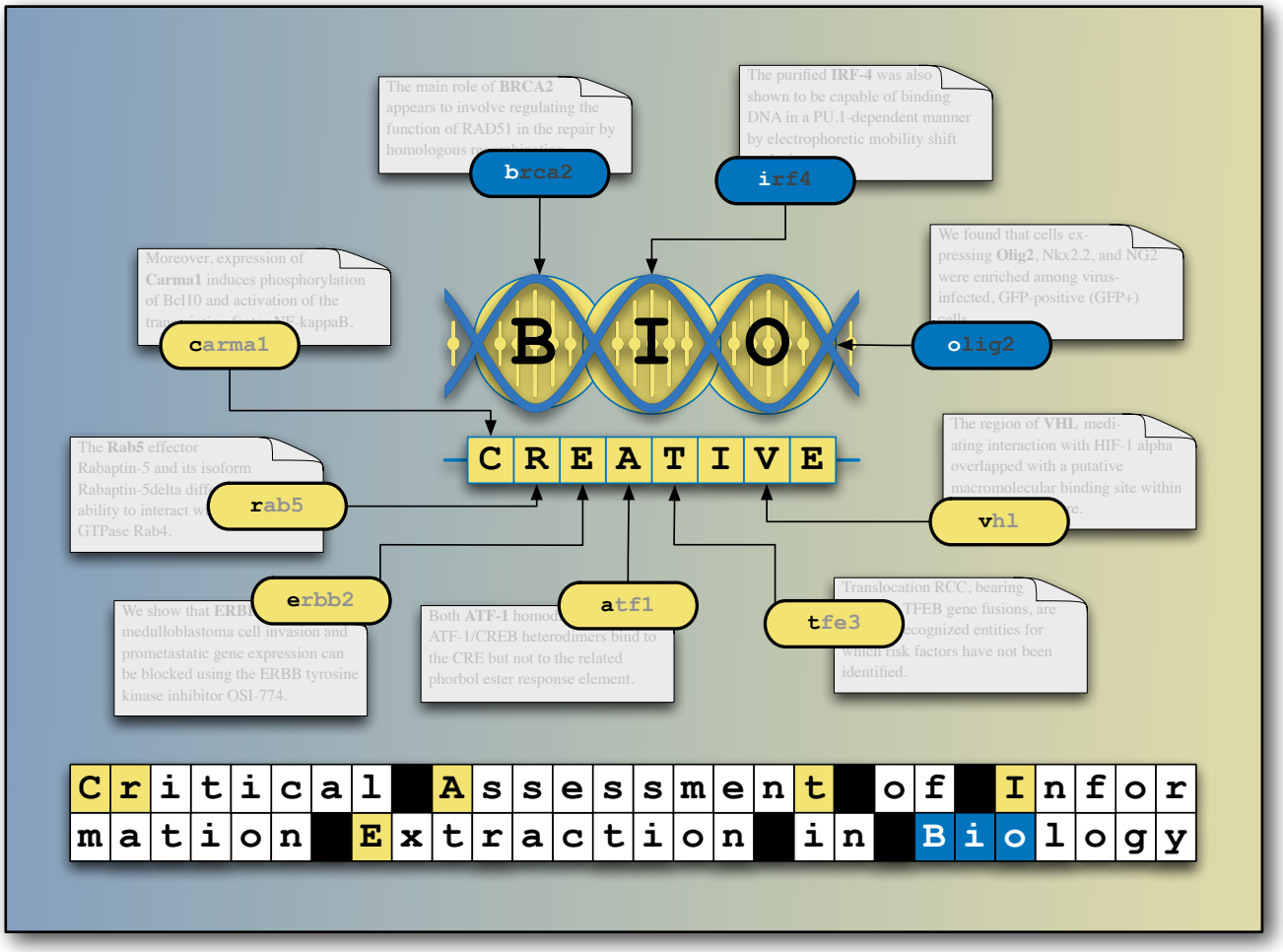
BioCreative II.5 Workshop 2009

Special Session on Digital Annotations

cnio
Centro Nacional de Investigaciones Oncológicas



October 7th - 9th, 2009

www.BioCreative.org

MBRACE Grid Network of excellence

ENFIN Enabling Systems Biology www.enfin.org

INB

Genoma España

BioCreative II.5 Workshop 2009
special session | **Digital Annotations**

Auditorium of the Spanish National Cancer Research Centre (CNIO)
Madrid, Spain, October 7th - 9th 2009

Organizers

- **Belén Bañeres**
  Spanish National Cancer Research Centre (CNIO), Spain
- **Gianni Cesareni**
  Department of Biology, University of Rome Tor Vergata, Italy
- **Lynette Hirschman**
  Biomedical Informatics for the Information Technology Center (MITRE), USA
- **Martin Krallinger**
  Spanish National Cancer Research Centre (CNIO), Spain
- **Florian Leitner**
  Spanish National Cancer Research Centre (CNIO), Spain
- **Alfonso Valencia**
  Spanish National Cancer Research Centre (CNIO), Spain

Sponsors

BioCreative is supported by the Spanish National Bioinformatics Institute (www.inab.org) - a platform of Genoma España, and the Spanish National Cancer Research Centre (CNIO). The BioCreative II.5 workshop is supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and EMBRACE LSG-CT-2004-512092.

table | **Contents**

workshop | **Programme**

wednesday | **Oct 7**

digital annotations *the proponents*	08:30	09:00	**Registration**
	09:00	09:20	Workshop opening and welcome *Alfonso Valencia, CNIO*
	09:20	10:00	Digital Annotations: the publishers *Adriaan Klinkenberg, Elsevier*
	10:00	10:40	Active Learning for More Efficient Corpus Annotation *Udo Hahn, Friedrich-Schiller-Universität Jena*
	10:40	11:20	Digital Annotation for Production Bioinformatics Systems *Judith Blake, The Jackson Laboratory*
			coffee | **Break**
digital annotations *the proponents*	11:40	12:20	UniProt KnowledgeBase (UniProtKB) annotation and text mining *Alan Bridge, Swiss Institute of Bioinformatics*
	12:20	13:00	Text Mining needs of bioinformaticians in the Pharmaceutical Industry *Ian Harrow, Pfitzer Global Research and Development*
			lunch | **Break**
digital annotations *febs & biocreative*	14:30	15:10	The curation process in the context of the FEBS letters SDA experiment *Gianni Cesareni, University of Rome Tor Vergata*
	15:10	15:50	The BioCreative II.5 challenge overview *Martin Krallinger, CNIO*
	15:50	16:30	Analysis of Author, Expert and System Curation Results for the INT (Interacting Proteins) Task *Lynette Hirschman, MITRE*
			coffee | **Break**
digital annotations *febs & biocreative*	17:00	17:40	BioCreative II.5: Evaluation and Ensemble System Performance *Florian Leitner, CNIO & Lynette Hirschman, MITRE*
	17:40	18:30	**Discussion** | Digital annotations, the FEBS experiment, and the BioCreative challenge

thursday | **Oct 8**

participant systems	09:00	09:50	**Keynote** | Rethinking the goals of BioNLP research: What are we trying to accomplish? *Larry Hunter, University of Colorado Denver School of Medicine*
	09:50	10:15	Normalizing Interactor Proteins and Extracting Interaction Protein Pairs using Support Vector Machines *Yifei Chen, Vrije Universiteit Brussel*
	10:15	10:40	Online protein interaction extraction and normalization at Arizona State University *Jörg Hakenberg, Arizona State University*
			coffee | **Break & Poster Session**
participant systems	11:20	11:45	IASL-IISR interactor normalization system using a multi-stage cross-species gene normalization algorithm and SVM-based ranking *Hong-Jie Dai, Institute of Information Science*
	11:45	12:10	OntoGene in BioCreative II.5 *Fabio Rinaldi, University of Zurich*
	12:10	12:35	AkaneRE Relation Extraction: Protein Normalization (INT) and Interaction (IPT) in the BioCreAtIvE II.5 Challenge *Rune Sætre, University of Tokyo*
	12:35	13:00	Classification of protein-protein interaction documents using text and citation network features *Luis M. Rocha, Indiana University and Instituto Gulbenkian de Ciência*
			lunch | **Break**
applied systems	14:30	15:20	**Keynote** | Predicting the future of prediction *Anna Tramontano, Sapienza University of Rome*
	15:20	15:50	WikiGenes - Collaborative scientific publishing on the Web *Robert Hoffmann, Memorial Sloan-Kettering Cancer Center*
	15:50	16:20	Reflect: An Augmented Browsing Tool for Scientist *Seán I. O'Donoghue, EMBL*
			coffee | **Break & Photo Session**
applied systems	16:50	17:20	The OKKAM authoring platform for entity annotation: the SDA case *Stefano Bocconi, Elsevier Labs and University of Trento*
	17:20	17:50	The BioCreative Meta-Server in the challenge and as a collaborative text-mining platform *Florian Leitner, CNIO*
	17:50	18:30	**Discussion** | Applied systems in text mining - now and where we need to go
			dinner | **Workshop**

friday | **Oct 9**

participant systems	09:30	10:20	**Keynote** | CALBC: Producing a large-scale biomedical corpus is a challenge *Dietrich Rebholz-Schuhmann, European Bioinformatics Institute*
	10:20	10:35	Combining regular expressions and lexical frequency for online extraction of protein-protein interactions *Frederic Ehrler, University Hospital of Geneva*
	10:35	10:50	Biocreative II.5: A Probabilistic Dimensional Data Model for Protein Identification, Disambiguation, and Interaction Discovery *Jay Urbain, Milwaukee School of Engineering*
			coffee | **Break**
participant systems	11:20	11:40	Information Extraction of Normalized Protein Interaction Pairs Utilizing Linguistic and Semantic Cues *Karin Verspoor, University of Colorado Denver School of Medicine*
	11:40	12:00	Empirical investigations into full-text protein interaction article categorization task (ACT) in the BioCreative II.5 Challenge *Man Lan, Institute for Infocomm Research*
	12:00	12:15	Applying Lazy Local Learning in BC II.5 Article Categorization Task *Cheng-Ju Kuo, Institute of Information Science*
	12:15	12:45	Outlook: BioCreative III and closing session *Alfonso Valencia, CNIO and Lynette Hirschman, MITRE*

Digital Annotations: the publishers

Adriaan Klinkenberg

Executive Publisher Elsevier Life Sciences, Elsevier BV, Radarweg 29, 1043 NX Amsterdam, Netherlands
F.Klinkenberg@elsevier.com

In recent years, Publishers have successfully migrated print journals to an online environment, and in doing so, increased usage and accessibility. The next step in this migration process is to start making full use of web 2.0, HTML, linking and implementation of semantic annotation and text mining tools so that richer versions of articles become hyperlinked to factual databases and offer annotation and graphical features beyond the article as we currently know it. Recent examples include the FEBS SDA Experiment and the integration of the NextBio text-mining interface into ScienceDirect. In the coming years, Publishers will need to redesign their workflows, from e-submission tools to final appearance online. The role of Publishers may well be expanding beyond traditional boundaries, as their expertise and knowledge in content management and in managing information technology processes can usefully be deployed in the pre-publication process and data gathering steps, as well as post-publication in increasing distribution and usefulness of the data. This presentation will highlight some of the (annotation) challenges and opportunities that we are currently facing in the publishing industry.

Active Learning for More Efficient Corpus Annotation

Udo Hahn

Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Germany
hahn@coling-uni-jena.de

Building gold standards for the evaluation of the performance of systems for text analytics (information extraction, text mining, text summarization, etc.) or their methodological foundations (tokenization, POS tagging, chunking, parsing, named entity recognition, relation extraction, anaphora resolution, etc.) constitutes a serious bottleneck for research and applications. This process is known to be slow (human annotators have to be trained, often in lots of iterative cycles) and costly (both in terms of time and money to be spent to deliver good quality of meta data).

Active Learning (AL) has recently been proposed by an increasing number of researchers as a methodology to combat some of the problems implied by standard annotation procedures. The main idea behind AL is based on the intuition that some annotations are "easy" to encode, while others may be much harder to deal with. In AL jargon, the first are less informative for training a classifier, while the latter are much more informative for the classification decision problem and thus have to be selectively picked up for consideration.

In my talk, I will introduce the AL paradigm and discuss some of the achievement we have made in the past years of experience creating corpora within the AL paradigm involving a variety of different annotation tasks (with a focus on biomedical named entities though). Altogether, the data seems to indicate that we can get annotation data with AL in a cheaper but still reliable way (compared with standard random selection of entities to be annotated). Rather than just advertising AL, I will also discuss some of the caveats related to AL and put it in a larger perspective for advanced gold data annotation.

Digital Annotation for Production Bioinformatics Systems

Judith A. Blake

The Jackson Laboratory Bar Harbor, ME, USA
judith.blake@jax.org

The Mouse Genome Informatics resource is a model organism database resource for the laboratory mouse, an important experimental organism for the study of human biology and disease. MGI has rigorous document triage and annotation procedures designed to identify articles about mouse genome biology and determine whether those articles should be curated. Each month, over 1000 journal articles are processed for Gene Ontology terms, gene mapping, gene expression, phenotype data and other key biological information. Curators utilize a variety of ontologies and terminologies to facilitate integration and retrieval of information. MGI regularly provides gold-standard curated biomedical literature for NLP research and evaluations. Recently, MGI evaluated dictionary-based text mining tools for integration into our production system. Although we don't foresee that human curation tasks can be fully automated in the near future, we are eager to implement entity name recognition and gene tagging tools that can help streamline our curation workflow. I will discuss efforts to identify and utilize digital annotation tools into our production bioinformatics system.

UniProt KnowledgeBase (UniProtKB) annotation and text mining

Alan Bridge

Alan Bridge[1a], Lydie Bougueleret[1], Amos Bairoch[2], and Ioannis Xenarios[1,3]

[1] Swiss-Prot group, Swiss Institute of Bioinformatics, CMU, 1 Michel Servet, 1211 Geneva 4, Switzerland
[2] Department of Structural Biology and Bioinformatics, Faculty of Medicine, University of Geneva, Switzerland
[3] Vital-IT Group, Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland
a) alan.bridge@isb-sib.ch

The UniProt Knowledgebase (UniProtKB) is an expertly curated database of protein information with cross-references to multiple external sources. The UniProtKB/Swiss-Prot section contains manually annotated records covering many species with a particular focus on human and other mammals, non-mammalian vertebrates such as Xenopus and zebrafish, Drosophila melanogaster, Caenorhabditis elegans, plants, fungi, bacteria and archaea, viruses and toxins. Complete manual annotation of a UniProtKB/Swiss-Prot record requires a precise description of protein function(s), enzyme-specific information (including kinetic parameters where available), biologically relevant domains and sites, post-translational modifications, subcellular location(s), tissue specificity, developmental specific expression, structure, interactions, splice isoform(s), associated diseases or deficiencies or abnormalities, etc. UniProtKB makes use of a number of controlled vocabularies and ontologies including the Gene Ontology (GO), enzyme (EC) nomenclature and reactions, hierarchical descriptions of biochemical pathways, tissues, strains, subcellular locations and topologies (http://www.uniprot.org/docs/#vocabulary). UniProtKB keywords precisely summarize the content of each entry, and, like other controlled vocabularies, are fully mapped to their corresponding GO terms where appropriate.

The information used to create these annotations is extracted by UniProtKB annotators from the full text of articles, including figures and tables, methods, and supplementary materials, and is combined with annotator-evaluated computational analysis and information from other external databases. Although PubMed remains our most widely used search engine for document retrieval and annotation prioritization, we also employ text mining tools for this and other annotation tasks. One such tool is STRING (Search Tool for Retrieval of INteracting Genes/proteins), a meta-database of known and predicted protein interactions (http://string.embl.de/). STRING integrates text mining results with information on genomic context, expression patterns, and experimentally determined protein interactions and pathways, and provides links to other information sources such as PDB and Online Mendelian Inheritance in Man (OMIM). STRING incorporates an intuitive user interface for easy navigation of protein interaction networks, which facilitates the coherent annotation of groups of interacting proteins. In addition to external resources like STRING we have also developed in-house tools for specific annotation tasks, including a method for retrieving texts describing single amino acid polymorphisms in human proteins. This tool uses UniProtKB/Swiss-Prot sequence annotations (such as signal peptides and alternatively spliced isoform sequences) to correct potential discrepancies in mapping numbered sequence positions from article abstracts to the appropriate UniProtKB entry. Such tools demonstrate the utility of an integrated approach to text mining, allowing curators to directly consider several sources of biological information when prioritizing entries and selecting documents for annotation.

Text Mining needs of bioinformaticians in the Pharmaceutical Industry

Ian Harrow

Ian Harrow[1a] and Phoebe Roberts[2]

[1] eBiology Group, Pfizer Global Research and Development, Sandwich Laboratories, Kent, CT13 9NJ, UK
[2] Computational Science Centre of Emphasis, Pfizer Global Research and Development, RTC, Cambridge, USA
a) ian.harrow@pfizer.com

This presentation describes global and local applications of text mining at Pfizer. For broad delivery of text mining results, we employ a range of solutions, starting with dictionary-based retrieval of single entities and co-occurring entities, building to more powerful information extraction using Natural Language Processing (NLP), disambiguation, and relevance ranking. Common queries central to the drug development process provide comprehensive assessment of biological and chemical literature, and results are displayed in our main knowledge base, Pharmamatrix, which is continuously updated by searches across Medline and Embase. Scaling this volume of text mining within Pfizer's enterprise environment, which is largely based on Oracle, is achieved through a grid computing approach. This allows us to integrate text mining results with many different types of structured data from public and proprietary databases to deliver relevant information to our scientists.

Sophisticated text mining solutions, including NLP, fact extraction and entity normalization, is made accessible to end users via user-friendly commercial tools. A work flow for mining user-defined full text corpora involves using QUOSA to retrieve full text documents, which are exported to Linguamatics I2E for mining. The same dictionary resources developed for Pharmamatrix can be harnessed by the NLP engine, I2E. This has allowed us to respond rapidly and iteratively to specific biological questions, frequently resulting in supplementation or creation of database content. A use case to obtain kinetic parameters used by systems biologists will be described. Patterns of usage from such custom full text mining will be given along with learning and outstanding challenges.

The curation process in the context of the FEBS letters SDA experiment

Gianni Cesareni

Department of Biology, University of Rome Tor Vergata, Italy
cesareni@uniroma2.it

I will describe the curation strategies and rules adopted by the IMEx databases and I will report how the curators' practice and performance has changed and matured during the months of this experiment. I will use the quantitative results of the SDA experiment to illustrate the pros and cons of an editorial procedure based on a collaboration between authors, curators, editors and publishers and, finally, I will discuss how this procedure can be improved by automatic tools.

Analysis of Author, Expert and System Curation Results for the INT (Interacting Proteins) Task

Lynette Hirschman

Lynette Hirschman[a] and Scott Mardis

Biomedical Informatics for the Information Technology Center, MITRE Corporation, USA
a) lynette@mitre.org

This talk presents an analysis of the author-curated submissions which were created as part of the initial FEBS Letters Structured Digital Abstract experiment. For a set of 48 articles, we have data from the author curations, the MINT expert curation (providing a Gold Standard), and over 40 training runs from 10 groups who submitted automated systems for evaluation.

Taken as a group, the authors achieved micro and macro averaged f-measure a little over 0.6 (with recall somewhat lower than precision). However, if we use the pooled system results and a simple voting algorithm to both filter and supplement the author annotations, it is possible to improve the joint author+automated system performance to around 0.7 f-measure. These results are significant because they show how automated tools can provide authors with tools to improve the quality of their annotations.

In addition, we will characterize the differences between author and system submissions, particularly proteins that were missed by the authors compared to those that were missed by the automated systems.

The BioCreative II.5 challenge overview

Martin Krallinger

Martin Krallinger, Florian Leitner, and Alfonso Valencia[a]

Spanish National Cancer Research Centre (CNIO), Madrid, Spain
a) avalencia@cnio.es

This talk will provide a general introduction to the motivation behind the BioCreative initiative and related efforts in the context of community evaluations for text mining and information extraction applied to the biological literature. Some of the specific differences between previous BioCreative challenges (I and II) and the current BioCreative II.5 will be pointed out. BioCreative II.5 is unique not only in the sense of a collaborative setting that involved publishers, database curators and the text mining community, but also could connect authors themselves into the data preparation cycle. BioCreative efforts were using traditional offline evaluation models, while BioCreative II.5 focused on integrating the challenge into an online setting, where systems can be made available to the public immediately.

The main concern of BioCreative II.5 centered on the creation of an environment where the objectives were selected by author- and curator-annotation requirements. BioCreative II.5 assesses the feasibility of assisting these annotators with automated, online IE systems. The challenge is firmly based on the context of the FEBS Letters experiment, a collaborative effort between the interaction database MINT and publisher house Elsevier/FEBS Letters to promote full-text article annotation by the authors of the papers published in FEBS Letters. In this experiment, authors were asked to annotate experimentally shown protein-protein interactions (PPIs) and this data was then used to construct Structured Digital Abstracts (SDAs), consisting of the relevant proteins in terms of the corresponding database identifiers, the interaction type, and the interaction detection method - the latter two as Molecular Interaction ontology terms.

The BioCreative II.5 challenge focused on the reproduction of parts of the SDAs in an online setting. The specific focus on experimentally verified protein-protein interactions (PPI) reported in the literature, with the final aim of assisting authors and curators in the annotation process. BC II.5 included three tasks relevant in the process of construction of SDAs: 1) the selection of articles that contain relevant, experimentally proven PPIs (Article Classification Task, ACT), 2) the identification of the relevant interacting proteins and their mapping to the corresponding database entries (Interactor Normalization Task, INT), and 3) the extraction of interactions by detecting pairs of interacting proteins (Interaction Pair Task, IPT). The corpora for the evaluation consisted of 1190 full-text articles from FEBS Letters, split evenly into training and test set. The training set were articles from 2008 containing the publicly available SDA-containing articles from FEBS, while the test set was composed of MINT curator annotations from 2007 and earlier. The complete annotation process was modeled online, with a client (the BioCreative Meta-Server, BCMS) sending annotation requests containing the full-text to the participant's servers ("Annotation Servers") that in turn had to respond with the annotation data within a ten-minute time frame. We will present the background of the challenge in this talk, an overview of the primary evaluation results in the follow-up talk this day, and some insight on the technical implications of this novel online evaluation strategy and the BCMS in a specialized talk on Thursday.

BioCreative II.5: Evaluation and ensemble system performance

Florian Leitner and Lynette Hirschman

Scott Mardis[1a], Florian Leitner[2b], and Lynette Hirschman[1c]

[1] Biomedical Informatics for the Information Technology Center, MITRE Corporation, USA
[2] Spanish National Cancer Research Centre (CNIO), Madrid, Spain
a) mardis@mitre.org
b) fleitner@cnio.es
c) lynette@mitre.org

For each task in the challenge (ACT, INT, IPT - see challenge overview), participants were asked to report ranked result lists; an ordered list of articles by their likelihood to contain PPIs (ACT), the list of UniProt (release 14.8) accessions for each article (INT), and the list of interaction pairs (as UniProt accessions) for the IPT task. In correspondence to the organization of the SDAs participants were ask to provide the list of entities but not to map the entities in the corresponding text.

The main evaluation measures the performance of the systems with respect to producing ranked lists of results that would match the gold standard as close as possible, ideally having the true hits in the top ranks. Therefore, the evaluation on the raw data calculates the macro-averaged performance (average values per paper) of the systems defined as the area under the interpolated precision/recall-curve (AUC iP/R), a measure of precision and recall with respect to the ranked list of results generated by the systems. The highest AUC iP/R scores achieved by the best systems were: 0.678 for ACT, 0.435 for INT, and 0.222 for IPT. The participants were asked to report identifiers for the complete set of UniProt accessions, implying the correct identification of the species. This is a non-trivial task even for the human curators, in part because authors often fail to mention the species and because experiments carried out with proteins from different species are common in the PPI literature. However, it would be rather simple for authors and curators to improve the results from the automatic systems by identifying the relevant organisms. Orthologs (proteins evolutionary related and performing equivalent functions in different organisms) could then be mapped to the correct organisms. Therefore, as an alternative evaluation strategy, we consider it more realistic to count orthologs as correctly identified if they were also homonymous to the gold standard annotation and removed (filtered) proteins from the results that did not belong to the same species as the proteins annotated in the gold standard. This practice would favor high-recall, low-precision systems - to counter this tendency, we used the balanced F-score for this second evaluation. The best systems with their associated precision (P) and recall (R) values achieved: for the interactor normalization (INT), a P: 67%, R: 52%, and F-score: 0.551; for the interaction pairs (IPT) a P: 38%, R: 30%, and F-score: 0.301.

Furthermore, we explore the collected results of the protein normalization and protein pair identification tasks to determine if pooled results will yield better performance than any particular participating system for a range of BioCreative II.5 metrics. To do this, we select the best system run from each team for each metric of interest, in order to select maximally independent approaches for the pooled results. We explore how the composite results change as different metrics and different selection criteria are used. We discuss the potential user impact of improvements achieved through creating such composite systems. Preliminary results indicate that both simple voting, as well as more sophisticated learning methods, produce improvements that will positively impact users. Because users rarely have time and attention to explore extensive lists of candidate annotations, we will describe the impact of limiting the number of ranked responses in each run. We show how such limits impact the composite classifiers, clarifying the extent to which automated methods make use of the lower ranked results.

Rethinking the goals of BioNLP research: What are we trying to accomplish?

Larry Hunter

Center for Computational Pharmacology, University of Colorado Denver School of Medicine,
PO Box 6511, MS 8303, Aurora, CO 80045, USA
Larry.Hunter@ucdenver.edu

For the past decade, BioNLP research has primarily focused on entity identification and extraction of relationship triples from texts. BioCreative and other shared tasks demonstrate that substantial progress has been made in these efforts, although there remains room for improvement. However, now that relatively high performance on these tasks is obtainable, the question remains how such capabilities facilitate biomedical research. Providing automated assistance to gene ontology curators and other manual biomedical knowledge representation efforts is one commonly cited application. However, the relatively poor performance of BioNLP tools in this specific application (see, e.g. Alex, et al., and Caporaso, et al., from PSB 2008) calls into question whether application of high F-score entity recognition algorithms are really the best way to improve curators' efficiency. Furthermore, assisting curators is a rather cramped view of the utility of BioNLP; isn't there something of more direct utility to bench scientists that BioNLP methods make possible? The TREC Genomics evaluations further call into question the utility of NLP methods compared to other approaches to information retrieval. So what should be trying to accomplish?

Leach, et al., PLoS Comp Bio 2009 lays out a new, more expansive role for BioNLP in building and maintaining knowledge-based systems for the analysis of high-throughput biomedical data. Gene- and ontology-centric knowledge representation schemes facilitate the alignment of existing knowledge with data from expression arrays, genome-wide association studies, shotgun proteomics, nucleosome footprinting and other genome-scale technologies. Part of the difficulty bench scientists face in interpreting the results of such experiments is placing them in the context of what is already known, particularly since taking full advantage of such global analyses requires integrating information large amounts of information that cross many disciplinary boundaries. However, BioNLP techniques alone are not sufficient for this goal; they must be integrated with systems that also capture knowledge from non-textual sources, that make inferences about implicit information, that assess the certainty and/or value of information, and that present the results in a way that supports the needs of users. While extrinsic evaluation of the contribution of BioNLP to such systems is complex, the insight-based evaluation methods developed by the scientific visualization community offer an important model. Only by considering BioNLP in the context of larger goals for knowledge-based systems in biomedicine is it possible to lay out a future research (evaluation!) agenda that will make our field genuinely relevant to biomedical science.

Predicting the future of prediction

Anna Tramontano

Department of BioChemical Sciences, Sapienza University of Rome, Italy
anna.tramontano@uniroma1.it

The Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment has been around for 15 years [1], it involves hundreds of research groups around the world [2] and has been used as a model for several similar experiments. A number of aspects of protein structure prediction have been addressed by CASP in its lifetime, but perhaps more importantly, trends fostered, or even driven, by the experiment are becoming apparent.

There have been landmarks in CASP history that are important to mention. Some might look trivial, such as standardizing formats and methods for gathering the predictions, but have had a serious impact nevertheless. Others are more scientifically interesting, for instance the assessment of the maturity of some areas such as secondary structure prediction, the validity of novel approaches, for example fold recognition or fragment based methods, the effectiveness of advanced technical solutions as exemplified by metaservers, the need of introducing methods for tackling emerging areas, e.g. the detection of intrinsically disordered regions in proteins or the identification of functional sites in proteins of unknown function.

Every round of the experiment seizes the present state of methods before they make their way in everybody's toolbox, but their value is not only the assessment of the present [2]; they permit comparison with the past to identify progress and what led to them [3], and provide data to attempt the prediction of the future in protein structure prediction.

Perhaps the most relevant questions that we can ask by analysing CASP history are: How much is technology driving discoveries? Where are we all going in terms of methods? Which specific aspects of the problem are we likely to focus on in the future and how much is the experiment itself affecting future directions?

1. Moult J, Pedersen J, Judson R & Fidelis K (1995) **A large-scale experiment to assess protein structure prediction methods.** Proteins 23, ii-v.
2. Moult J, Fidelis K, Kryshtafovych A, Rost B & Tramontano A (2009) **Critical assessment of methods of protein structure prediction (CASP) - round VIII.** Proteins, *in press.*
3. Kryshtafovych A, Fidelis K & Moult J (2009) **CASP8 results in context of previous experiments.** Proteins, *in press.*

CALBC: Producing a large-scale biomedical corpus is a challenge

Dietrich Rebholz-Schuhmann

European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
rebholz@ebi.ac.uk

The CALBC initiative is a challenge in several senses: (1) generating a large-scale biomedical corpus, (2) harmonising the annotations from different sources, and (3) enabling a TM challenge on the harmonised corpus. Altogether, the CALBC project aims to provide a large-scale biomedical text corpus that contains semantic annotations for tagged named entities of different kinds.

In the first phase, the annotation systems from 5 participants have been collected in a common annotation format. The format provided concept ids in the boundary annotations and was used to automatically compare and align the results.

In the first phase of the project, the produced results from the participants have been integrated into a single harmonised corpus ("silver standard" corpus). In the next step, the submissions of the participants have been evaluated against the silver standard corpus.

The annotated corpus will be used at a later stage for a public challenge measuring the performance of annotation systems. We expect that our approach is suitable to generate a large-scale annotated corpus.

WikiGenes - Collaborative scientific publishing on the Web

Robert Hoffmann

Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, USA
roberth@MIT.edu

Over the past decades, a hypercycle of technological advancement, scientific hypothesis and increasing amount of data has led to a shift in perspective toward system thinking. The way scientific insights are published and communicated, however, has remained essentially the same since the days of Charles Darwin.

As a result, facts pertaining to specific genes, chemicals and pathologies are scattered over hundreds and even thousands of different articles and journals. This is a concern to individual scientists whose contributions lay idle in unstructured archives. Most importantly, this situation is detrimental to scientific progress as a whole, which is an incremental process and thus dependent on the consolidation and accessibility of information in every step. Whereas this risk of losing information is recognized for other kinds of data, text mining is the only attempt to recover the valuable information from thousands of biomedical articles published every day.

WikiGenes, a collaborative knowledgebase for biology and medicine, is the first system to unite the values of traditional publishing with the unique opportunities of wikis and the Internet. In WikiGenes scientists can publish their insights while contributing at the same time to a resource of structured, semantically enriched and integrated knowledge. Authors who invest valuable time and knowledge are provided due credit in WikiGenes, since a powerful authorship tracking technology enables the unambiguous attribution of every sentence and every word to its author. WikiGenes is open access and available online at http://www.wikigenes.org.

Reflect: An Augmented Browsing Tool for Scientists

Seán I. O'Donoghue

Seán I. O'Donoghue[1a], Heiko Horn[1,2], Evangelos Pafilis[1], Michael Kuhn[1], Reinhard Schneider[1], and Lars J. Jensen[1,2]

[1] European Molecular Biology Laboratory, Heidelberg, Germany
[2] NNF Center for Protein Research, University of Copenhagen, Denmark
a) sean@mandala.cc

Introduction

Anyone who regularly reads life science literature often comes across names of genes, proteins, or small molecules that they would like to know more about. In such cases it would be helpful if these entities were tagged so as to allow the reader easy access to further information about each specific entity. However, to systematically add such semantic capabilities usually requires publishers to invest in considerable re-engineering of their backend content services. For these reasons, only a tiny faction of all web content currently has systematic semantic annotations.

Results

To make this process easier, we have developed a new service called Reflect [1] that can be installed as a plug-in to Firefox or Internet Explorer. Reflect lets end-users systematically tag gene, protein, and small molecule names in any web page, typically within a few seconds, and without affecting document layout. Clicking on a tagged gene or protein name opens a popup showing a concise summary that includes synonyms, database identifiers, sequence, domains, 3D structure, interaction partners, subcellular location, and related literature. Clicking on a tagged small molecule name opens a popup showing 2D structure and interaction partners. The popups also allow navigation to commonly used databases. Reflect also has SOAP, REST (HTTP post), and JavaScript interfaces, allowing publishers and content providers to access Reflect programmatically and provide tagged content to directly to end-users.

Discussion

Usage of Reflect has grown rapidly within the life sciences, and while currently only genes, protein and small molecule names are tagged, we plan to soon expand the scope to include general knowledge. The popularity of Reflect demonstrates the use and feasibility of letting end-users decide how and when to add semantic annotations. Ultimately, we believe that semantics is in the eye of the end-user, and hence we predict that end-user driven, augmented browsing approaches such as Reflect will become increasingly important in the near future, and will change dramatically how scientists use the web.

Availability

Reflect is freely available at http://reflect.ws

1. Pafilis, E., O'Donoghue, S. I., Jensen, L. J. et al. **Reflect: Augmented Browsing for the Life Scientist.** Nature Biotechnology 27 (6), 308 (2009)

The OKKAM authoring platform for entity annotation: the SDA case

Stefano Bocconi

Elsevier Labs and University of Trento, Italy
stefano.bocconi@gmail.com

In this talk I describe an authoring environment for scientific article annotation with specific reference to FEBS Letters SDA experiment. The environment consists of an authoring client (Word) that interfaces via a plugin to online annotation services. Some of these services are provided by the OKKAM platform, some can be existing services such as Whatizit. At the moment we are using Whatizit for proteins, with the goal of integrate other services as well. The goal is to provide the authors with different NER services geared to different type of entities, and output an annotated article. The presentation will include a short demo of the Word plugin functionality.

The BioCreative Meta-Server in the challenge and as a collaborative text-mining platform

Florian Leitner

Florian Leitner, Martin Krallinger, and Alfonso Valencia[a]

Spanish National Cancer Research Centre (CNIO), Madrid, Spain
a) avalencia@cnio.es

In cooperation with the participants of BioCreative II, the first meta-server prototype, collating web-services from various text mining providers, was created. The aim of this platform, called the BioCreative Meta-Server (BCMS), is to provide common annotation types available in the field of biological information extraction: Gene/Protein Mention, Gene/Protein Normalization assignment, Taxa labeling, and a binary classification if a text describes protein-protein interactions. For the BioCreative II.5 challenge, Protein-Protein Interactions were added. The prototype version of the platform consists of three units: (1) a static text collection (the approximately 22,900 PubMed abstracts used during BioCreative II), (2) the annotation servers providing access to the information extraction systems for the meta-server, and (3) the meta-server itself, storing the annotations in a central repository and distributing them to the users in raw and collated forms. Access to the data is provided by three different options: (1) via web-interface; (2) via web-services; (3) as annotation repository download.

The platform was extended for BioCreative II.5 from working with annotations for MEDLINE abstracts to using full-text articles. This specialized BCMS version was used to carry out the whole event in an online scenario: Participants were asked to implement up to five Annotation Servers (representing up to five possible runs for each task in the challenge). The functionality of these servers was asserted by running the training set data through the annotation servers before actually moving all teams to the online phase. To this end, a completely new interface was created that allows the providers of Annotation Servers to monitor and administer their services on the platform itself, including status and error reporting, while the BCMS takes care of distributing, validating, and collating the data. A complete redesign of the internals of the platform had been made in order to handle the load created by the necessity to process full-text with dozens of threads and database requests in parallel.

One main argument for this collaborative effort is that an aggregated repository provides advantages over isolated web-service solutions: results from different systems are directly comparable, all results are contained in a unified data structure and encoding, users are confronted with a single portal to the existing heterogeneous text mining services, and the system is easily extensible to other annotation types as has been shown now for the BioCreative II.5 challenge. At this stage, the public interface elements provided by the BCMS prototype from BC II and the specialized version created for BC II.5 need to be united. This final version, with the necessary community support, would result in a consolidated platform that provides a system for the continuous annotation of biomedical reports and evaluation of text mining systems in the area of protein interactions. It will allow the text mining community to offer and make use of an collaborative access to a continuous stream of annotations provided by independent information extraction tools relevant for Molecular Biology.

Normalizing Interactor Proteins and Extracting Interaction Protein Pairs using Support Vector Machines

Yifei Chen

Yifei Chen[1a], Feng Liu[2b], Bernard Manderick[1]

[1] Computational Modeling Lab, Vrije Universiteit Brussel, Belgium
[2] Vrije Universiteit Brussel, Belgium
a) yifechen@vub.ac.be
b) fengliu@vub.ac.be

For Interactor normalization task (INT), we build a system consisting of five essential components, a gene mention recognizer (GMRer), a species filter, a dictionary generator, a dictionary matcher and a SVMs-based disambiguation filter. A GMRer is trained on the data set of BioCreative II GM task; a species filter tries to detect the species names from the documents; Dictionary here is built to provide protein identifiers and their synonyms. Our dictionary generator here is designed to build a comprehensive dictionary. At the same it can reduce the variety and ambiguity of synonyms in the dictionary as much as possible. The dictionary matcher provides mapping criterion to associate the gene mentions with their Uniprot AC Numbers. Due to the ambiguity of the biological terms, the AC Numbers for gene mentions are not unique. Hence, the SVM-based disambiguation filter is operating on the candidate sets and assign a unique AC Number for each gene mentions. The filter is build on a set of extracted features from the gene mentions themselves, their context and background resources. After learning and prediction, the outputs of our system are the final normalized gene mentions with unique AC Numbers. An interesting remark is that if we heavily make use of the interaction databases, e.g., MINT and IntAct to filter the false positives, the F score of cross validation on the training data can be around 90, which can be done by judging if this protein has a partner protein to consist of a interaction pair in a given document and measuring the cosine similarity between the given document and the evidences (actually the abstracts in PubMed talking about this interaction pair) provided by MINT and IntAct.

For Interaction pair task (IPT), first we need to preprocess the original documents, which is to split some complex sentences into one main sentence and several clauses. Here it should be noticed that we only consider the intra-sentence interactions. Hence we only choose those sentences/clauses within which there exists at least one interaction protein pair to build our training data. Our generic system making use of SVMs is based on the following features: interaction proteins and their context words, the distance in tokens between the two proteins, the number of other identified proteins between the two proteins, interaction words, the position of interaction words and the distance in words between the interaction word and the protein nearest to it. In order to incorporate the domain knowledge for our system, we also design 16 syntactic pattern features and 2 Boolean features to indicate if this interaction pair exists in MINT and IntAct databases, which can greatly improve the performance further. We manually annotate the interaction proteins for our training data according to the SDA information in order to eliminate the negative impacts (i.e., the introduced errors) of the Interactor Normalizer. On this training data, our system can achieve around 94 F score using 10-fold cross validation.

Online protein interaction extraction and normalization at Arizona State University

Jörg Hakenberg

Jörg Hakenberg[1a], Robert J. Leaman[1,2], Nguyen Ha Vo[1], Siddhartha Jonnalagadda[2], Ryan Sullivan[2], Christopher Miller[2], Luis Tari[1], Chitta Baral[1], and Graciela Gonzalez[2b]

[1] Dept Computer Science, Arizona State University, USA
[2] Dept Biomedical Informatics, Arizona State University, USA
a) joerg.hakenberg@asu.edu
b) Graciela.Gonzalez@asu.edu

Biomedical text mining seeks to extract information on entities such as proteins and drugs, and their relationships, from text. The BioCreative 2.5 challenge was a community effort to assess current systems for the tasks of protein-protein interaction extraction (IPT task), normalization of protein names to UniProt IDs (INT), as well as classification of texts regarding relevance to protein interactions (ACT). Notably, training and evaluation data were full text publications.

The system we present here handles the INT and IPT tasks. For INT, we used BANNER to recognize protein names, which was trained on BioCreative 1 and 2 GM data. We used dictionary matching to find likely candidate entries from UniProt, and then applied PNAT, a derivative of the previously published GNAT, to find the correct UniProt ID among the candidates. PNAT uses context information such as species, GeneOntology terms, associations with diseases, cellular locations, etc., to disambiguate proteins given the paragraph they are mentioned in.

For IPT, we extracted all sentences from the training data that contained an annotated protein pair. Note that the annotation provided was on a document level, so we used BANNER/PNAT to find actual occurrences in the full text documents. We further restricted all these sentences to the smallest snippet to contain both proteins and an interaction-indicating keyword ("binds", "co-localized"). All snippets were clustered using alignment as a similarity function. By multiple sentence alignment, we identified a consensus among similar sentences (allowing, for instance, for groups of words at certain positions). These consensus patterns were transformed into OpenDMAP patterns, which we use to spot similar sentences in new text.

The maximum scores achieved among the five configurations we set up were a macro-averaged F-score of 0.55 for INT (including orthologs) and 0.30 for IPT. Maximum AUC iP/R values were 0.53 and 0.27, respectively. Our system is available via the BioCreative MetaService framework for online extraction of protein interactions and normalization of protein names.

IASL-IISR interactor normalization system using a multi-stage cross-species gene normalization algorithm and SVM-based ranking

Hong-Jie Dai

Hong-Jie Dai[1a], Po-Ting Lai[2], Chi-Hsin Huang[1], Yen-Ching Chang[1], Yue-Yang Bow[1], Hsin-Ta Wu[1], Richard Tzong-Han Tsai[2b], and Wen-Lian Hsu[1c]

[1] Institute of Information Science, Academia Sinica, Nankang, Taipei 115, R.O.C, Taiwan
[2] Dept. of Computer Science & Engineering, Yuan Ze Univ., Chung-Li, Taiwan, R.O.C.
a) hongjie@iis.sinica.edu.tw
b) thtsai@saturn.yzu.edu.tw
c) hsu@iis.sinica.edu.tw

The interactor normalization task (INT) is to identify genes which play the interactor role in protein-protein interactions (PPIs), to map these genes to unique IDs, and to rank them according to their normalized confidence. INT has two subtasks: gene normalization (GN) and interactor ranking. The main difficulties of INT GN are identifying genes across species and using full papers, instead of abstracts, as target data. Other INT tasks include identifying if a normalized gene is an interactor and ranking all interactors according to their normalized confidence. To tackle these problems, we developed a multi-stage GN algorithm and a ranking method which exploit information in different parts of a paper.

In our GN algorithm, machine-learning-based gene mention recognition (GMR) and GN is carried out starting from the sections with the richest context information (introduction) to those with the poorest (captions). Then gene names in the keyword/abbreviations fields are used to adjust gene mention boundaries. We have also compiled a full name/abbreviation mapping table and a blacklist extracted from part of the UMLS and MeSH databases. For cross-species identification, the system checks prefixes for species markers, such hOBP (human), and surrounding words for species-related terms. If the species can be determined, only identifiers belonging to that species are reserved for matching. Two dictionary-based matching strategies are employed. The first uses a dictionary compiled by collecting gene names in EntreGene and SWISS-PROT and generating their orthographical variants. Each recognized gene mention is looked up in the dictionary. If an exact match is found, the gene is assigned that entry's ID. To improve normalization recall, we also compiled a larger dictionary by adding the gene entries in TrEMBL to the first dictionary. Because all TrEMBL terms are indexed by the Lucene search engine, we can use the engine to find partial matches for gene mentions. For genes assigned more than one identifier, the ID is determined by taking a weighted vote of several rule-based classifiers.

We rank all normalized IDs using a support vector machine with sixty features. The most important feature checks whether the ID's context matches linguistic patterns describing PPI to determine if the ID is an interactor. Another feature indicates in which section(s) the ID appears, while most remaining features are related to GN.

The training set is compiled from the annotated interactors in the official training set. Three-fold cross-validation was carried out on the training set, showing a promising AUC of 58.35%. Our experimental results also show that with full text, versus abstract only, INT AUC performance was 22.22% higher.

OntoGene in BioCreative II.5

Fabio Rinaldi

Fabio Rinaldi[1a], Gerold Schneider[2], Kaarel Kaljurand[2], Simon Clematide[2]

[1] IFI, University of Zurich, Switzerland
[2] University of Zurich, Switzerland
a) rinaldi.fabio@gmail.com

OntoGene is a text mining research activity partially supported by the Swiss National Science Foundation, with additional supported provided by NITAS, Text Mining Services, Novartis.

Our system is based on a common core consisting of a pipeline of standard NLP components and an internally developed dependency parser. The core architecture is expanded by input filters, which allow easy adaptation to novel input formats, output filters, which allow a presentation targeted to the specific needs of the application, and pluggable modules which allow the customization to specific tasks.

The pipeline has been applied with minor adaptations to the BioNLP shared task (event extraction), BioCreative II.5 (detection of protein-protein interactions), and an internal project aiming at simulating the process of curation of IntAct annotators.

In all the tasks, our results have been consistently highly ranked.

In the full paper, we will describe in detail the customizations undertaken for the IPT task of BioCreative II.5, in particular concerning (1) detection and grounding of domain entities (2) disambiguation based on our "focus organism" approach, (3) optimal ranking of candidate interactions. The latter point takes into account various factors including the 'novelty' of the information presented by the candidate interaction, the zone in which it is found, the syntactic path between the two interacting proteins.

AkaneRE Relation Extraction: Protein Normalization (INT) and Interaction (IPT) in the BioCreAtIvE II.5 Challenge

Rune Sætre

Rune Sætre[1a], Kazuhiro Yoshida[1], Makoto Miwa[1], Takuya Matsuzaki[1], Yoshinobu Kano[1], and Jun'ichi Tsujii[1,2]

[1] Department of Computer Science, The University of Tokyo, Japan
[2] NaCTeM (National Center for Text Mining), University of Manchester, UK
a) rune.saetre@is.s.u-tokyo.ac.jp

Objective
To evaluate the new Akane Relation Extraction (RE) system on a general Protein-Protein Interaction (PPI) Information Extraction (IE) task. AkaneRE was used to produce results for the INT and IPT tasks in the BioCreative II.5 (BC2.5) text mining challenge (http://www.biocreative.org/). The AkaneRE system used for BC2.5 can be divided into two parts: The general RE engine and the BC2.5-specific module. By changing the BioCreative-specific part, the same system can also be applied to other types of IE tasks, like the BioNLP shared task on event extraction (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/). In this report, we will provide a comparison of the BC2.5 and BioNLP tasks and work-flows.

Method
The old AkanePPI has been generalized into the new AkaneRE, which can extract general relation between any textual entities. To control the flow of information through the system, a UIMA based platform called U-Compare was used (http://u-compare.org/). U-Compare provides a drag-and-drop interface to create a work-flow, so the modules can easily be exchanged with other U-Compare compatible components. Our BC2.5 work-flow consists of the following components: Receiving UTF-8 encoded journal papers from the web, removing all the text markup and tokenizing the text into an internal plain text ASCII representation. The Genia-tagger is used for Part-of-speech tagging and to lemmatize the plain text. A fast version of Enju (Mogura) and the Genia Dependency parser (GDep) are used to parse the text. MedTNer recognizes protein names and normalizes them to the set of most probable UniProt identifiers. The main AkaneRE modules predict pairs and rank them according to the interaction probability. The article-level probability is calculated by using features such as the number of the sentence-level mentions of the single proteins, their co-occurrences, and the sections in which they are mentioned. Species disambiguation is resolved by the re-ranking module, using features both from the text, and from the citations and the reference list.

Results
The Area Under the interpolated Precision/Recall Curve (iPR-AUC) is maximized by the re-ranker, by ranking all possible predictions, but it is possible to set a threshold to optimize the Precision, Recall or F-score values instead. Our best filtered result in the online challenge was an iPR-AUC of 50% for the Interacting protein Normalization Task (INT), and 30% for the Interacting Pairs Task (IPT). In the offline challenge the performance was an iPR-AUC of 58% for the INT task, and 40% for the IPT task.
For reference, the old AkanePPI system was number six of sixteen systems in the BioCreative2 Interacting Proteins Sub-task (IPS), with scores P=18, R=27 and F=19.
The same RE engine was used in the BioNLP shared task, and was ranked as number six among 24 systems with official scores P=54, R=28, F=37.

Classification of protein-protein interaction documents using text and citation network features

Luis M. Rocha

Artemy Kolchinsky[1,2], Alaa Abi-Haidar[1,2], Jasleen Kaur[1], Ahmed Hamed[1], and Luis M. Rocha[1,2a]

[1] School of Informatics, Indiana University, USA
[2] FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Portugal
a) rocha@indiana.edu

We participated in the *Article Classification Task* (ACT): binary classification of full-text documents relevant for protein-protein interaction (PPI). We used two distinct classifiers for the online and offline challenges: (1) the lightweight *Variable Trigonometric Threshold* (VTT) linear classifier we previously introduced successfully in BioCreative 2 (BC2), and (2) a more computationally expensive Naive Bayes classifier using features from the citation network of the relevant literature, respectively. We supplemented the supplied training data with full-text documents previously classified in BC2, as well as from the MIPS database. For VTT on the **online part** of this challenge, we used word-pair features computed from a relatively small number of words: the top 1000 words obtained from the rank product of the ranks computed via the TF.IDF measure on all documents and a score that maximizes the difference between the probabilities of occurrence in relevant and irrelevant documents (after removal of stop words and stemming). VTT's linear decision surface is further controlled by the number of proteins mentions in abstracts and figures as identified by ABNER. The parameters for the submitted decision surface were obtained via an exhaustive search for maximum performance of accuracy and F-Score measures on K-Fold sets of the training data described above. For the citation network classifier on the **offline part** of this challenge, we first implemented a Naive Bayes classifier using citation features such as: (1) cited PMIDs (2) citation authors and (3) citation author,year pairs (all weighted equally). Additionally, to exploit the known network structure, we further used co-citation data. Our approach involved harvesting approximately 18500 PDF files, from which about 16000 PMIDs, 316000 referenced PMIDs, and 637500 citations were extracted. Looking at the F-Score and AUC, we can see that our offline submissions substantially outperformed our online submissions. It is worth noticing that our best online submissions are those that did not train on the additional MIPS documents we used, and also did not use the entity counts via ABNER (online runs 4 and 5). As for the offline submissions, the integration of the citation classifier with cocitation information and the VTT classifier was our top overall classifier for both F-Score and AUC performance (offline run 5). Therefore, the inclusion of citation network data was shown to be very promising. Clearly, there is much more to do in this domain, especially by harvesting more reliable network data then entire biomedical bibliome. However, the very good performance of the citation network classifier in tandem with VTT (offline run 5) shows how including even incomplete citation network information is very beneficial.

Combining regular expressions and lexical frequency for online extraction of protein-protein interactions

Frederic Ehrler

Frederic Ehrler[1a] and Patrick Ruch[2]

[1] University hospital of Geneva, Switzerland
[2] HES Geneva, Switzerland
a) Frederic.Ehrler@unige.ch

Background
Following the interest taken into the BioCreAtIvE MetaServer integration platform developed at the end of the previous campaign, BioCreAtIvE II.5 aimed at evaluating more formally online systems for protein interactions extraction. BCII.5 goes one-step further to reflect at best real annotation task by recognizing protein-protein interactions from full-text coming from the FEBS Letters and by providing the result through a standardized annotation server.
Purpose
Compared to the previous editions of BioCreAtIvE, two significant modifications have been introduced. The former concerns the use of full-text from the FEBS Letter journal instead of using only abstracts. The latter pertains to the test of interaction extraction methods in real time through an annotation server that receives documents and returns a list of their corresponding interactions.
Methods
We tested several methods and combination of methods to identify protein names and to extract interactions. On the side of the protein recognition, we compared two methods. The first method consisted to recognize the gene names based on the low probability of their specific terminology to occur in usual English language. The second method attempted to recognize the protein names based on predefined patterns, with verbal heads as main interaction trigger (68 items such as binds, interacts…). Additionally to these two methods, we capitalized on protein name frequency. Indeed, it was expected that the frequency was a good criteria to identify proteins occurring in a given document and to discard erroneous names found in our gene dictionary (GPSDB). On the side of the interactions themselves, we also tested two methods: a simple one that attempted to link the two closest proteins to a verbal head; and a second that employed manually designed regular expressions.
Results
As expected with our simple dictionary-based methods, performances on named entity normalization remain relatively low. F-scores range from 15% to 28% with a recall that goes up to 35%. As performances on interaction detection are bounded by those of named entity normalization, they are naturally relatively low, from 3% to 12% regarding F-score. The best run is obtained using regular expressions to identify both protein names and interacting protein pairs. Further, we observe that filtering strongly protein names based on frequency results in missing interaction pairs and significantly affect recall.
Conclusion
Performances obtained by protein interaction annotators remain low. In our system, the weakness of the named entities recognition model can clearly be improved. Since we opted for a time efficient approach with high recall, our system tends to generate false positive candidates that influence negatively the identification of the interactions. From a qualitative perspective, the system has been significantly improved and it is now part of the EAGLi engine, see http://eagl.unige.ch/EAGLi/ (or http://eagl.unige.ch/EAGLi/jsp/more.jsp?pmid=19098309&query=what%20proteins%20can%20interact%20with%20cftr%20?#PP for an example of the protein interaction service.)

A Probabilistic Dimensional Data Model for Protein Identification, Disambiguation, and Interaction Discovery

Jay Urbain

Electrical Engineering and Computer Science Deptartment, Milwaukee School of Engineering, USA
urbain@msoe.edu

For Biocreative II.5, we explore development of a text mining system based on a probabilistic dimensional data model for efficient identification and disambiguation of protein entities and their interactions in context.

Traditional text mining systems follow a sequential process to identify, disambiguate, and relate named entities where each stage in this process is dependent on the accuracy of prior stages to be effective.

We propose an alternative model, a model that integrates multiple sources of evidence for entity identification, disambiguation, and relation discovery simultaneously within the framework of a probabilistic graphical model. In such a model, the strength of each individual component is strengthened by evidence of other components in the model. For example, the disambiguation of one protein helps with the identification of another. In turn, the discovery of a relation between two proteins helps with the disambiguation of other proteins.

Such a multievidentiary model requires efficient search and aggregation of term and entity statistics at multiple levels of document and protein database structure, including indexing of individual words, entities, sentences, paragraphs, and document. To meet the needs of such multievidentiary models, we present a new text mining system based on a dimensional data model, and our first experiences using the system to develop probabilistic graphical models for protein identification, disambiguation, and interaction discovery for Biocreative II.5.

Information Extraction of Normalized Protein Interaction Pairs Utilizing Linguistic and Semantic Cues

Karin Verspoor

Karin Verspoor[a], Christophe Roeder, Helen L. Johnson, K. Bretonnel Cohen, William A. Baumgartner Jr., and Lawrence Hunter

Center for Computational Pharmacology, University of Colorado Denver School of Medicine, USA
a) karin.verspoor@ucdenver.edu

The Center for Computational Pharmacology approach to information extraction for the BioCreative II.5 challenge takes advantage of our open-source tools for Biomedical Natural Language processing (BioNLP, http://bionlp.sourceforge.net), specifically the OpenDMAP system [1] and the BioNLP-UIMA framework, built on the Apache Unstructured Information Management Architecture [2] (http://incubator.apache.org/uima). We submitted responses to the gene normalization task (INT) and the interaction pair task (IPT).

For the gene normalization task, we utilized a two-step process of (1) dictionary lookup of gene names using the SwissProt subset of the UniProt database and (2) ambiguity resolution of competing candidates utilizing various document-internal clues. For the dictionary lookup, we normalize both the dictionary terms and input text by making the string lowercase, eliminating punctuation such as apostrophes, hyphens, and parentheses, converting Greek letters and Roman numerals to a standard form, and finally removing spaces. We search for any matches within sentences of the input text, starting on a token boundary and ending on a token boundary, unless there is a plural in which case the right token boundary constraint is relaxed. If there is more than one database match over a given span, an ambiguity resolution algorithm is applied, primarily utilizing species information in the document but also considering any detected abbreviations. Both global and local strategies for species detection are utilized, and a confidence score is attributed to specific gene normalizations based on what combination of evidence supports the normalization.

For the interaction pair task, we employed the concept recognition mechanisms of OpenDMAP to search for phrases in the input documents that match pre-defined patterns of expression for various interaction types. We used the patterns from our BioCreative II submission [3], with minor modifications. We additionally took advantage of a coordination module that we utilized in our BioNLP'09 system [4] to support handling of coordinated lists of interactors. The semantic grammar generally defines an interaction as the co-occurrence of two proteins in conjunction with an interaction event concept, in various linguistic constructs. This approach has been shown to result in high-precision extraction of interaction events [4].

1. Hunter L, Lu Z, Firby J, Baumgartner Jr. WA, Johnson HL, Ogren PV, and Cohen KV. 2008. **OpenDMAP: An open-source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-specific gene expression.** BMC Bioinformatics, 9(78).
2. Ferrucci D and Lally A. 2004. **Building an example application with the unstructured information management architecture.** IBM Systems Journal, 43(3):455– 475, July.
3. Baumgartner WA, Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A, White EK, Medvedeva O, Cohen KB and Hunter L: **Concept recognition for extracting protein interaction relations from biomedical text.** Genome Biology 2008, 9(Suppl 2):S9.
4. Cohen KB, Verspoor K, Johnson H, Roeder C, Ogren P, Baumgartner W, White E, Tipney H, and Hunter L. (2009) **High-precision biological event extraction with a concept recognizer.** In Proceedings of the Workshop on BioNLP:Shared Task, pages 50-58, Boulder, CO, June 2009.

Empirical investigations into full-text protein interaction article categorization task (ACT) in the BioCreative II.5 Challenge

Man Lan

Man Lan[a] and Jian Su[b]

Institute for Infocomm Research, Singapore
a) mlan@i2r.a-star.edu.sg
b) sujian@i2r.a-star.edu.sg

The selection of protein interaction document is one important application of information management for scientific community and has a direct impact on the quality of downstream biomedical text mining applications, such as information extraction and retrieval, summarization, QA etc. The BioCreative II.5 Challenge protein interaction Article Categorization task (ACT) is actually a binary text classification task which concerns whether a given structured full-text biomedical article contains protein interaction information. This may be the first attempt of full-text protein interaction classification in wide research community. In this paper, we compare and evaluate the effectiveness of different section types of full-text articles in the protein interaction classification task. Moreover, in practice, the rare number of true positive samples results in unstable performance and unreliable classifier model trained on it. Previous research on learning with skewed class distributions has altered the class distribution using up-sampling and down-sampling. In this paper, we also discuss various investigations into skewed protein interaction document classification and analyze the effect of various issues related to the choice of external sources, up-sampling training sets, classifiers. We report on the various factors above to show that (1) full-text biomedical article contains a wealth of scientific information important to users that may not be completely represented by abstracts and/or indexing terms, which improves the accuracy performance of classification; (2) reinforcing true positive samples significantly increases the accuracy and stability performance of classification.

Applying Lazy Local Learning in BC II.5 Article Categorization Task

Cheng-Ju Kuo

Cheng-Ju Kuo[1a], Maurice HT Ling[3,4c], and Chun-Nan Hsu[1,2b]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] USC/Information Science Institute, Marina del Rey, CA, USA
[3] School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore
[4] Department of Zoology, The University of Melbourne, Parkville, Victoria, Australia
a) cju.kuo@gmail.com,
b) chunnan@iis.sinica.edu.tw,
c) mauriceling@acm.org

Article categorization task (ACT) in BC II.5 is to classify whether the input article contains protein interaction descriptors.

In our approach, given an article for classification, the system will invoke the selection of top 100 abstracts from the 5,495 abstracts in the BC II - IAS data corpus based on the document cosine similarity between the query article and each abstract in the IAS dataset. These 100 abstracts were used to train a classifier by AdaBoost (with Decision Tree as the weak learner), implemented in MALLET. The trained classifier will then be used to class the query article into either containing protein interaction or not. Therefore, for each query, there will be a unique classifier trained by a different set of selected abstracts. This is a lazy local approach in the sense that a classifier will not be trained until a query article is given and the training examples are selected locally in the neighborhood of the query article in the feature space.

In order to reduce the feature size generated from full-text article, only title, abstract and captions of figures of FEBS full-text article were used to generate feature vectors for model training and testing after the removal of common words and stemming of tokens.

Independent evaluation demonstrated 17.4% precision for identifying articles containing protein interaction descriptors (58 true positives out of 333 identified positives) but 98.9% precision in identifying articles not containing protein interaction descriptors (259 true negatives out of 262 identified negatives). This result suggests that our system can be applied as an effective filter to eliminate articles without protein interaction descriptors and thus greatly reduce the number of articles for subsequent processing of extracting protein interactions.

BioAlvis II, NLP-based semantic mining of literature on molecular biology of bacteria

Sophie Aubin, Philippe Bessières, Robert Bossy, Laurent Gillard, Julien Jourde, Frédéric Papazian, Philippe Veber, and Claire Nedellec

MIG Lab, INRA, France

A large part of the biomedical knowledge is still only available in documents in natural language, despite of the growing number of available structured databases. Research efforts towards the Semantic Web aim "at replacing the current web of links with a web of meaning" producing large-scale methods for automating deep semantic analysis documents and markup suitable for information extraction or information retrieval applications in the biomedical domain.

In bacterial molecular biology and genetics, the lack of available semantic resources (lexicon and ontology) is a major obstacle to document annotation and to the development of knowledge management systems that integrate textual, experimental and in silico information. Corpus-based Machine Learning provides an attractive alternative to the manual acquisition of resources provided that the linguistic and biological specificities are taken into account. We describe the principles of the development of the BioAlvis framework and an experimental evaluation of BioAlvis in an IR task on PubMed references about prokaryotes.

Using Full Text from Scientific Articles in Portable Document Format

Roman Klinger, Robert Pesch, Heinz Theodor Mevissen, and Juliane Fluck

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

Many full texts of journals are only available electronically in the Portable Document Format (PDF). Contrary to publications available in the Extensible Markup Language (XML) or similar, these documents are lacking information about the structure of the text, which is even not trivial to extract. Therefore, several challenges occur when well-established methods for typical data sources like Medline abstracts should be applied on that data source.

These include extraction of plain text from the document with newly discovered cases of character encoding and reading order reconstruction as well as special tokenization issues. Latter are e.g. the lack of encoded white space (typically only the position of glyphs is stored) or word-wraps at lines (because of narrow columns), column breaks or page breaks.

Unlike in file formats including logical markup the document structure needs to be reconstructed. This includes detection of tables, figures, captions, headers, footers, footnotes as well as the typical "Problem-Solution" structure of scientific articles.

Solving all these problems leeds to the ability of application of elaborated text mining techniques. To present their results to an end-user, valuable methods of visualization need to be implemented allowing for the accommodating presentation of text mining results in the original layout of an article.

We present our approaches together with preliminary results.

Judith **Blake**

judith.blake@jax.org

My research focuses on functional and comparative genome informatics. I work on the development of systems to integrate and interrogate genetic, genomic and phenotypic information. I am one of the leaders of the Gene Ontology (GO) project and I have been deeply involved with the work of the GO Consortium since its inception. The Gene Ontology project is an international effort to provide controlled structured vocabularies for molecular biology that serve as terminologies, classifications and ontologies to further data integration, analysis and reasoning. My interest in bio-ontologies stems as well from the work I do as a principal investigator with the Mouse Genome Informatics (MGI) project at The Jackson Laboratory. The MGI system is a model organism community database resource that provides integrated information about the genetics, genomics and phenotypes of the laboratory mouse. MGI identifies and curates over 14,000 publications My current research projects combine bio-ontologies and database knowledge systems to represent disease processes with the objective of discovering molecular elements that contribute to particular pathologies such as lung cancer.

Stefano **Bocconi**

stefano.bocconi@gmail.com

Stefano Bocconi is currently working for the OKKAM European project, which aims at implementing an infrastructure for uniquely assigning identifiers to entities. This implies services for the definition of new identifiers for entities which have not one yet and the mapping of existing identifiers that refer to the same entities. This infrastructure is used to support authoring of news items and scientific papers, specifically by annotating the entities contained in the text with their identifiers, and to provide additional knowledge-based services that take advantage of the lack of ambiguity due to the use of unique identifiers.

Stefano Bocconi is mainly interested in entity-based semantic integration of different knowledge sources, which encompasses identity problems (level of granularity, identity across space and time) and the definition of similarity/equivalence relations between entities, as well as the practical side of integrating different sources such as the Open Linked Data cloud.

Other interests are the use of semantics in multimedia for video generation (the subject of his PhD at the CWI in Amsterdam) and model-based diagnosis (two years post-doc at the University of Turin).

Alan **Bridge**

alan.bridge@isb-sib.ch

Alan Bridge is a senior annotator at the UniProtKB/Swiss-Prot database which is based at the Swiss Institute of Bioinformatics in Geneva. A trained biologist, he holds a PhD in cell biology and an MSc in bioinformatics. His current work comprises all aspects of quality assurance and the development of automatic annotation processes for UniProtKB/Swiss-Prot. In the context of BioCreative his major interests include the integration of text mining tools into the annotation workflow at UniProtKB/Swiss-Prot, particularly in the area of annotation prioritization and updates, and developing links to papers with structured digital abstracts.

Gianni **Cesareni**

cesareni@uniroma2.it

Gianni Cesareni is a Full Professor of Genetics at the University of Rome Tor Vergata (Italy). After obtaining a degree in physics at the University of Rome La Sapienza he spent three years in Cambridge in the laboratory of Sidney Brenner. He then moved to the EMBL in Heidelberg where he led a group working on the mechanisms controlling plasmid DNA replication. Since 1989 he teaches and works in Rome. He is interested in the interplay between specificity and promiscuity in the protein interaction network mediated by protein recognition modules. He is the founder of the MINT protein interaction database.

Udo **Hahn**

hahn@coling-uni-jena.de

Udo Hahn is a full professor of Computational Linguistics and Language Technology at Friedrich-Schiller-Universität Jena (Germany) and head of the Jena University Language and Information Engineering (JULIE) Lab. He has been working in the field of biomedical NLP for more than a decade. Basically, he has been involved in two streams of work. One is dedicated to the design, the development and the engineering of biomedical ontologies from a principled methodological basis which is

rooted in description logics (e.g., the formalization of part-whole relations). His second stream of work covers the whole variety of natural language processing tasks that are relevant to support medical and biological researchers and developers, including but not limited to applications such as (multi-lingual) document retrieval, semantic search engines, information extraction and text mining systems.

System building for BioNLP applications requires a solid engineering approach to properly combine a large variety of single NLP components. Besides living up to common software engineering standards, his Lab's work is embedded in the framework of the Unstructured Information Management Architecture (UIMA), now an Apache incubator project. In recognition of his activities and achievements, Udo Hahn received two UIMA Innovation Awards (2007, 2008) from IBM. Contributions to a commonly shared UIMA type system and a strict dedication to Open Source activities have guided JULIE Lab's team work. Quite recently, JULIE Lab's contributions were ranked on second position (among 24 participants) for Task 1 of the "BioNLP'09 Shared Task on Event Extraction" competition.

The annotation of corpora has soon turned out as a major stepping stone for building large-scale systems for text analytics (not only) in the biomedical domain. The focus of JULIE Lab's activities in this area of work is on speeding up annotation processes (while keeping the level of commonly accepted standards and quality of annotations) for richly and diversely annotated corpora (usually involving a large number of entity classes). Out of these requirements we started our efforts in annotation methodology and, more concrete, in exploring the potential of Active Learning to build up large corpora in a more efficient, less costly but still reliable and valid manner.

Ian **Harrow**

ian.harrow@pfizer.com

Ian Harrow is a Senior Principal Scientist in the eBiology group which is part of Computational Biology in Sandwich, UK. He has been involved with large scale application of text mining technologies to support drug discovery over the last five years. He joined the Pfizer in 1984 working on the discovery of anti-parasitic agents for the Animal Health division. He moved into bioinformatics and human drug discovery in 1997 at the inception of the industrial scale genome era. Ian holds a BSc in Zoology from the University of Nottingham, a PhD in neurobiology and electrophysiology at University of Cambridge and gained 3 years post-doctoral experience at Columbia University, New York.

Lynette **Hirschman**

lynette@mitre.org

Lynette Hirschman is Director, Biomedical Informatics in the Information Technology Center at the MITRE Corporation in Bedford, MA. She received a B.A. in Chemistry from Oberlin College, a M.A. in German literature (University of California, Santa Barbara), and a Ph.D. in mathematical linguistics (University of Pennsylvania, 1972). At MITRE, Dr. Hirschman has worked in the areas of human computer interaction, human language understanding, and since 2000, bioinformatics and text mining for biomedical applications. She was PI for the DARPA-funded MiTAP project capturing disease outbreak from open source feeds. She is a founding organizer of BioCreative (Critical Assessment of Information Extraction for Biology), the first international challenge evaluation of text mining for the biomedical domain. She is currently working with the Genome Standards Consortium on metadata capture for metagenomics (under NSF funding), as well as on anonymization of records for protection of privacy, genotyping of influenza, and bioinformatics tools for glycobiology.

Robert **Hoffmann**

roberth@mit.edu

Robert Hoffmann is a computational biologist, affiliated with the Computational Biology Center at the Memorial Sloan-Kettering Cancer Center in New York. He studied Genetics at the University of Vienna and received his Master's degree in Bioinformatics under the tutelage of Dr. Peter Schuster at the Institute of Molecular Pathology (IMP). In 2001, Robert moved to Madrid to achieve his doctoral thesis in the Protein Design Group of Dr. Alfonso Valencia. During his PhD he contributed original work to the areas of text mining in the biomedical literature (www.ihop-net.org), the comparative analysis of protein networks and the evolution and dissemination of scientific knowledge. His works have been published in leading scientific journals, including Nature Genetics, PNAS and Trends in Genetics.

In 2006, Robert was awarded with the Society in Science Branco Weiss Fellowship, to explore novel ways of data and information management with the aim to improve the efficiency of global scientific research and also bring science closer to society. The initial part of this project was carried out at the Massachusetts Institute of Technology (MIT) at the group of Sir Tim Berners-Lee and has led to the development of the WikiGenes project (www.wikigenes.org). Robert is happy father of a child.

Selected Publications

- Hoffmann, R. **A wiki for the life sciences where authorship matters.** Nature Genetics 40, 1047 - 1051 (2008)
- Hoffmann, R., Valencia, A. **Implementing the iHOP concept for navigation of biomedical literature.** Bioinformatics 21(suppl. 2), ii252-ii258 (2005)
- Hoffmann, R., Valencia, A. **A gene network for navigating the literature.** Nature Genetics 36, 664 (2004)
- Pfeiffer, T. & Hoffmann, R. **Temporal patterns of genes in scientific publications.** PNAS 104 (29), 12052–12056 (2007)

Larry **Hunter**

larry.hunter@ucdenver.edu

Dr. Lawrence Hunter is the Director of the Computational Bioscience Program and of the Center for Computational Pharmacology at the University of Colorado School of Medicine, and a Professor in the departments of Pharmacology, Computer Science (Boulder), and Preventive Medicine and Biometrics. He received his Ph.D. in computer science from Yale University in 1989, and then spent more than 10 years at the National Institutes of Health, ending as the Chief of the Molecular Statistics and Bioinformatics Section at the National Cancer Institute. He inaugurated two of the most important academic bioinformatics conferences, ISMB and PSB, and was the founding President of the International Society for Computational Biology. Dr. Hunter's research interests span a wide range of areas, from cognitive science to rational drug design. His primary focus recently has been the integration of natural language processing, knowledge representation and machine learning techniques and their application to interpreting data generated by high throughput molecular biology.

Adriaan **Klinkenberg**

f.klinkenberg@elsevier.com

After graduating at the Agricultural University in Wageningen in virology, Adriaan joined Gist Brocades (now DSM) to work in their Scientific Information department, before moving to Elsevier Science in Amsterdam where he has held various positions in diverse disciplines in life sciences for the past twenty years. For the past two years, Adriaan has been instrumental in getting the FEBS SDA Experiment launched at the publisher's side, and getting a number of innovative "content-enhancing" projects initiated within Elsevier. His interests in improving research infrastructure in Europe were first raised when collaborating with MINT in Rome on SDAs and launching a new journal *Molecular Oncology* with FEBS, which introduced him to the issues of bioinformatics, biobanking and biocuration. He currently works on launching a new bioinformatics and biocuration title for Elsevier, which will hopefully also be a playground for testing new ideas and concepts.

Martin **Krallinger**

mkrallinger@cnio.es

Martin Krallinger is currently working at the Structural Biology and Biocomputing group of the Spanish National Cancer research Center (CNIO). He has a strong research record in biomedical text mining, including numerous highly cited publications in the field. He has been part of several international scientific conference committees (e.g. ISMB, BioLINK, ECCB, NETTAB or LBM2007) and carried out referee activities of over 12 prestigious journals in the field (including Bioinformatics, BMC Bioinformatics, Genome Biology or PLoS Computational Biology). In addition to the co-organization of scientific events such as the Second BioCreAtIvE Challenge Workshop or the workshop on Text Mining for the BioCuration Workflow at the 3rd International Biocuration conference he was responsible of several tutorials and lectures on biomedical text mining.

Florian **Leitner**

fleitner@cnio.es

Florian completed his Masters in Molecular Biology, specializing on Computational Biology, with Frank Eisenhaber at the IMP in Vienna, Austria. He previously worked for Rebecca Wade at the EML in Heidelberg, Germany, and for Markus Jaritz at the Novartis Research Campus in Vienna. Currently, he is doing his PhD at the CNIO in Spain with Alfonso Valencia, and has begun to specialize on text mining over the past few years.

His current interests are in BioNLP (Natural Language Processing for Computational Biology) and distributed (web-) servcies. He is developing a meta-service for BioNLP, the BioCreative Meta-Server for gene and protein annotation extraction from text. This tool is designed for exchanging, unifying, and comparing BioNLP data and act as a hub for both the BioNLP as well as the general molecular and computational biology community. In addition, he is interested in ontology learning and expansion and establishing an (digital) annotation standard for BioNLP that could become part of the BCMS. For the time being, Florian was mainly tasked with the arrangement of the BioCreative II.5 challenge and this workshop.

Seán I. **O'Donoghue**

sean@mandala.cc

Seán O'Donoghue is a research scientist at the Structural and Computational Biology programme at the European Molecular Biology Laboratory, Heidelberg, Germany. He received his B.Sc. (Hons) and PhD in biophysics from the University of Sydney, Australia. He has been actively engaged in bioinformatics research since 1988, mostly at the EMBL. His most significant contribution to date has been contributing to the development of ARIA, a method widely used for calculating 3D structures from NMR data. His currently coordinating a variety of research projects, including Reflect (http://reflect.ws), Martini (http://martini.embl.de), SRS 3D (http://srs3d.org), and an EMBO Workshop on Visualizing Biological Data (http://vizbi.org).

Dietrich **Rebholz-Schuhmann**

rebholz@ebi.ac.uk

Dietrich Rebholz-Schuhmann, MD, Ph.D., studied Medicine (University of Düsseldorf) and Computer Science (University of Passau). He worked in medical informatics research and at LION bioscience AG, Heidelberg, Germany, where he headed a research group in text mining and led the EUREKA research project "Bio-Path". In 2003 he joined the EBI as research group leader in literature analysis.

He was member of the Network of Excellence "SemanticMining" (NoE 507505) and was project partner in the FP6-IST project "BOOTStrep" (www.bootstrep.eu). His team is project partner in the UKPMC project (British Library, University of Manchester). Currently the Rebholz group is coordinating the support action CALBC, which is a challenge to the biomedical text mining community (www.calbc.eu).

His research interest lies in information extraction in the biomedical domain, development of novel text mining solutions and the standarisation of the scientific literature through integration into the bioinformatics database infrastructure. He is editor-in-chief of the journal of biomedical semantics (www.jbiomedsem.com) and has served a number of organizing and programme committees of international conferences including ISMB, ECCB, SMBM and LBM.

Anna **Tramontano**

anna.tramontano@uniroma1.it

Anna Tramontano was trained as a physicist but she soon became fascinated by the complexity of biology and by the promises of computational biology. After a post-doctoral period at UCSF, she joined the Biocomputing Programme of the EMBL in Heidelberg. In 1990 she moved back to Italy to work in the Merck Research Laboratories near Rome. In 2001, she returned to the academic world as a Chair Professor of Biochemistry in "La Sapienza" University in Rome where she continues to pursue her scientific interests on protein structure prediction and analysis in the Department of Biochemical Sciences.

She now leads a truly interdisciplinary group of about twenty scientists with a background in physics, biology, chemistry, engineering and computer science and their interests can be divided, with some overlap, into two main directions: development and/or improvements of computational biology methods and their application to the study of problems of biomedical interest. To the first area, belongs the development of methods for the prediction of the structure and function of proteins and nucleic acid. The other aspect of the research is the investigation of important biomedical problems related to pathologies induced by foreign agent such as Hepatitis C and malaria and genetic diseases as well as the study of biotechnologically important molecules.

Anna is a member of the European Molecular Biology Organization, the Scientific Council of Institute Pasteur - Fondazione Cenci Bolognetti, the organizing Committee of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) initiative, the EMBL Scientific Advisory Committee, the EBI Advisory Committee, the Scientific Advisory Board of the University of Zurich Research Priority Program "Systems biology/Functional Genomics". She is Associate Editor of Bioinformatics and Proteins and a member of the Editorial Board of The FEBS Journal.

She was awarded the KAUST Investigator Award, the prize for Natural Sciences of the Italian Government, the "Marotta Prize" of the Italian National Academy of Science and the Minerva Prize for Scientific Research and has published four books (Bioinformatica - Zanichelli; The ten most wanted solutions in Protein Bioinformatics - CRC Press; Protein Structure Prediction - Wiley; Introduction to Bioinformatics - CRC Press).

Alfonso **Valencia**

avalencia@cnio.es

Alfonso Valencia is a biologist with formal training in population genetics and biophysics which he received from the Universidad Complutense de Madrid.He was awarded his PhD in 1988 at the Universidad Autónoma de Madrid.

He was a Visiting Scientist at the American Red Cross Laboratory in 1987 and from 1989 - 1994 was a Postdoctoral Fellow at the laboratory of C. Sander at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

In 1994 Alfonso Valencia set up the Protein Design Group at the Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC) in Madrid where he was appointed as Research Professor in 2005.

He is a member of European Molecular Biology Organization, founder and former Vice President of the International Society for Computational Biology where he has been Chair of the Systems Biology and/or Text Mining tracks of its annual conference since 2003. He is a founding organiser of the annual European Computational Biology Conferences and co-organised the meeting in 2005.

He also serves on the Scientific Advisory Board of the Swiss Institute for Bioinformatics, Biozentrum, Basel, as well as the Steering Committee of the European Science Foundation Programme on Functional Genomics (2006-2011). His group participates in the three main bioinformatics Networks of Excellence organised under the 6th European Framework Programme (BioSapiens, EMBRACE and ENFIN).

Alfonso Valencia is co-organizer of the BioCreative challenges, co-executive editor of Bionformatics, serves on the Editorial Board of EMBO Journal and EMBO Reports, and is Director of the Spanish National Bioinformatics Institute (INB), a platform of Genoma España.

overview | **Participants**

Judith **Blake**	*The Jackson Laboratory*	judith.blake@jax.org USA
Stefano **Bocconi**	*Elsevier Labs and University of Trento*	stefano.bocconi@gmail.com Italy
Robert **Bossy**	*INRA*	Robert.Bossy@jouy.inra.fr France
Alan **Bridge**	*Swiss Institute of Bioinformatics*	alan.bridge@isb-sib.ch Switzerland
Gianni **Cesareni**	*University of Rome Tor Vergata*	cesareni@uniroma2.it Italy
Yifei **Chen**	*Vrije Universiteit Brussel*	yifechen@vub.ac.be Belgium
Hong-Jie **Dai**	*Institute of Information Science*	hongjie@iis.sinica.edu.tw Taiwan
Frederic **Ehrler**	*University Hospital of Geneva*	Frederic.Ehrler@unige.ch Switzerland
Laurent **Gillard**	*INRA*	laurent.gillard@jouy.inra.fr France
Udo **Hahn**	*Friedrich-Schiller-Universität Jena*	hahn@coling-uni-jena.de Germany
Jörg **Hakenberg**	*Arizona State University*	joerg.hakenberg@asu.edu USA
Ian **Harrow**	*Pfizer Global Research and Development*	ian.harrow@pfizer.com UK
Lynette **Hirschman**	*MITRE*	lynette@mitre.org USA
Robert **Hoffmann**	*Memorial Sloan-Kettering Cancer Center*	roberth@MIT.edu USA
Larry **Hunter**	*University of Colorado Denver School of Medicine*	Larry.Hunter@ucdenver.edu USA
Julien **Jourde**	*INRA*	julien.jourde@jouy.inra.fr France
Jin-Dong **Kim**	*Tokyo University*	jdkim@is.s.u-tokyo.ac.jp Japan
Roman **Klinger**	*Fraunhofer SCAI*	roman.klinger@scai.fhg.de Germany
Adriaan **Klinkenberg**	*Elsevier BV*	F.Klinkenberg@elsevier.com Netherlands

- Martin **Krallinger**
  CNIO
  mkrallinger@cnio.es
  Spain
- Cheng-Ju **Kuo**
  Institute of Information Science
  cju.kuo@gmail.com
  Taiwan
- Man **Lan**
  Institute for Infocomm Research
  mlan@i2r.a-star.edu.sg
  Singapore
- Florian **Leitner**
  CNIO
  fleitner@cnio.es
  Spain
- Sérgio **Matos**
  Universidade de Aveiro
  aleixomatos@ua.pt
  Portugal
- Manickam **Muthuraman**
  Wageningen University
  manickam.muthuraman@wur.nl
  Netherlands
- Claire **Nedellec**
  INRA
  claire.nedellec@jouy.inra.fr
  France
- Mariana **Neves**
  Centro Nacional de Biotecnologia, CSIC
  mlara@cnb.csic.es
  Spain
- Seán **O'Donoghue**
  European Molecular Biology Laboratory
  sean@mandala.cc
  Germany
- Alberto **Pascual Montano**
  National Center for Biotechnology
  pascual@cnb.csic.es
  Spain
- Dietrich **Rebholz-Schuhmann**
  European Bioinformatics Institute
  rebholz@ebi.ac.uk
  UK
- Fabio **Rinaldi**
  University of Zurich
  rinaldi@cl.uzh.ch
  Switzerland
- Luis **Rocha**
  Indiana University
  rocha@indiana.edu
  USA
- Rune **Sætre**
  University of Tokyo
  rune.saetre@is.s.u-tokyo.ac.jp
  Japan
- Isabel **Segura Bedmar**
  Universidad Carlos II de Madrid
  isegura@inf.uc3m.es
  Spain
- Anna **Tramontano**
  Sapienza University of Rome
  anna.tramontano@uniroma1.it
  Italy
- Jay **Urbain**
  Milwaukee School of Engineering
  urbain@msoe.edu
  USA
- Alfonso **Valencia**
  CNIO
  avalencia@cnio.es
  Spain
- Philippe **Veber**
  INRA
  pveber@free.fr
  France
- Karin **Verspoor**
  University of Colorado Denver
  Karin.Verspoor@ucdenver.edu
  USA
- Pierantonio **Zocchi**
  Sapienza University of Rome
  pierantonio.zocchi@gmail.com
  Italy

If you would like to learn more about BioCreative II - ask the organizers for your personal copy of the Genome Biology special issue or visit the Genome Biology website at http://genomebiology.com/supplements/9/S2.