Critical Assessment of Information Extraction Systems in Biology

Proceedings of BioCreative III Workshop

September 13 -15, 2010
Bethesda, MD USA

Editors:

Cecilia Arighi
Kevin Cohen
Lynette Hirschman
Martin Krallinger
Zhiyong Lu
Alfonso Valencia
John Wilbur
Cathy Wu

ISBN 978-1-4507-3685-5

2010 BioCreative III Workshop Proceedings
Table of Contents

Overview Papers

Systems Descriptions

Preface

Welcome to the BioCreative III workshop being held in Bethesda, Maryland, USA on September 13-15. On behalf of the Organizing Committee, we would like to thank you for your participation and hope you enjoy the workshop.

The BioCreative (Critical Assessment of Information Extraction systems in Biology) challenge evaluation consists of a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain (http://www.biocreative.org/). Its aim is to promote the development of text mining and text processing tools which are useful to the communities of researchers and database curators in the biological sciences. The main emphasis is on the comparison of methods and the community assessment of scientific progress, rather than on the purely competitive aspects.

The first BioCreative was held in 2004, and since then each challenge has consisted on a series of defined tasks, areas of focus in which particular NLP tasks are defined. BioCreative I focused on the extraction of gene or protein names from text , and their mapping into standardized gene identifiers (GN) for three model organism databases, and functional annotation, requiring systems to identify specific text passages that supported Gene Ontology annotations for specific proteins, given full text articles. BioCreative II (2007) focused on GN task but for human genes or gene products mentioned in PubMed/MEDLINE abstracts, and on protein-protein interaction (PPI) extraction, based on the main steps of a manual protein interaction annotation workflow. BioCreative II.5 (2009) focus on the PPI, the tasks were to rank articles for curation based on curatable PPIs; to identify the interacting proteins in the positive articles, and to identify interacting protein pairs.

The BioCreative III continues the tradition of a challenge evaluation on several tasks judged basic to effective text mining in biology, including a gene normalization (GN) task and two protein-protein interaction (PPI) tasks (interaction article classification, and interaction method detection). It also introduces a new interactive task (IAT), run as a demonstration task. The goal of IAT is to develop an interactive system to facilitate a user’s annotation of the unique database identifiers for all the genes appearing in an article. This task includes ranking genes by importance based preferably on the amount of described experimental information regarding genes.

The Biocreative III workshop includes two special panels, one to discuss the challenges and opportunities for text mining in biology led by several funding agencies and publishers, and a panel to discuss about system interoperability led by systems developers.

We would like to thank all participating teams, panelists and all the chairs and committee members.

The BioCreative III Workshop was supported by NSF grant DBI-0850319

Organizing Chairs
Cecilia Arighi, University of Delaware, USA
Cathy Wu, University of Delaware and Georgetown University, USA

BioCreative III Committees

Steering Committee

Cecilia Arighi, University of Delaware, USA
Kevin Cohen, University of Colorado, USA
Lynette Hirschman, MITRE Corporation, USA
Martin Krallinger, Spanish National Cancer Centre, CNIO, Spain
Zhiyong Lu, National Center for Biotechnology Information(NCBI), NIH, USA
Alfonso Valencia, Spanish National Cancer Centre, CNIO, Spain
John Wilbur, National Center for Biotechnology Information(NCBI), NIH, USA
Cathy Wu, University of Delaware and Georgetown University, USA

Scientific Committee

Sophia Ananiadou, University of Manchester, UK
Nigel Collier, Japanese Science and Technology Agency (JST), Japan
Mark Craven, University of Wisconsin, USA
Anna Divoli, University of Chicago, USA
Henning Hermjakob, EBI, UK
Eivind Hovig, Oslo University Hospital, Norwey
Lars Juhl Jensen, EMBL, Germany
Michael Krauthammer, Yale University, USA
Claire Nedellec, MIG, France
Goran Nenadic, University of Manchester, Manchester, UK
Jong C. Park, KAIST, South Korea
Dietrich Rebholz, EBI, UK
Andrey Rzhetsky, University of Chicago, USA
Hagit Shatkay, Queen's University, Canada
Neil Smalheiser, University of Illinois, USA
Larry Smith, NCBI, USA
Jun'ichi Tsujii, University of Tokyo, Japan and University of Manchester, UK
Karin Verspoor, University of Colorado Denver, USA
Andrew Chatr-aryamontri, University of Rome Tor Vergata, Italy
Phoebe Roberts, Pfizer, USA
Pascale Gaudet, Northwestern University, USA
Eva Huala, Stanford University, USA
Ian Harrow, Pfizer, UK
Michelle Giglio, University of Maryland, USA
Lois Maltais, Jackson Laboratory, USA

BioCreative III Committees (cont.)

Local Organizing Committee
Cecilia Arighi, University of Delaware, USA
Sun Kim, National Center for Biotechnology Information(NCBI), NIH, USA
Won Kim, National Center for Biotechnology Information(NCBI), NIH, USA
Zhiyong Lu, National Center for Biotechnology Information(NCBI), NIH, USA
Susan Phipps, University of Delaware, USA
Oana Catalina Tudor, University of Delaware, USA
John Wilbur, National Center for Biotechnology Information(NCBI), NIH, USA
Cathy Wu, University of Delaware and Georgetown University, USA

Proceedings Committee
Kevin Cohen, University of Colorado, USA
Katie Lakofsky, University of Delaware, USA

BioCreative III Workshop Agenda

September 13 - 15, 2010
Double Tree Hotel
Bethesda, Maryland, USA

Meeting Location: Second level
Registration: Outside Ballroom A/B
All presentations: Grand Ballroom A/B
Demo: Grand Ballroom A/B
Posters and Reception: Ballroom C

September 13, 2010 (Monday)	
7:30 AM – 5:00 PM	**Registration**
7:30 AM – 8:30 AM	Breakfast (provided, outside of Ballroom A/B in the EMC Foyer)
8:30 AM – 8:40 AM	**Workshop Opening** – Cathy Wu
8:40 AM – 9:10 AM	**Discussion of systems results (GN)** – Zhiyong Lu, NCBI
9:10 AM – 10:40 AM	**Challenge Participant systems 1 (GN)** – chaired by John Wilbur Team 63 University of Colorado by Karin Verspoor Team 65 University of Zurich by Fabio Rinaldi Team 68 Arizona State University by Martin Gerner Team 70 Universidade de Aveiro, IEETA by Sergio Matos Team 74 Institute of Information Science, Academia Sinica, Taiwan by Chun-Nan Hsu
10:40 AM – 11:00 AM	Break
11:00 AM – 1:00 PM	**Challenge Participant systems 2 (GN) –** chaired by Zhiyong Lu Team 78 University of Iowa by Sanmitra Bhattacharya Team 83 NCKU IKMlab by Chih-Hsuan Wei Team 89 University of Wisconsin-Milwaukee by Shashank Agarwal Team 93 University of Tokyo by Naoaki Okazaki Team 97 Georgetown University by Hongfang Liu Team 98 Tsinghua University by Minlie Huang Team 101 Yuan Ze University by Richard Tzong-Han Tsai
1:00 PM – 2:30 PM	Lunch (provided, hotel Restaurant)
2:30 PM – 3:30 PM	**Keynote: “**Communicating Scientific Knowledge: Is progress lagging?” **David Lipman**, Director of NCBI, NIH, USA
3:30 AM – 3:45 PM	Break
3:45 PM – 5:45 PM	Introduction by workshop sponsor Peter McCartney: Program Director, Division of Biological Infrastructure, NSF **Panel Discussion: Challenges and Opportunities for Text Mining in Biology: Government/Publisher Perspectives** • Government panel (Moderator: Cathy Wu) Valerie Florance: Director, Extramural Programs, NLM, NIH Karin Remington: Director, Center for Bioinformatics & Computational Biology, NIGMS, NIH Susan Gregurick: Program Manager, Computational Biology, DOE Sylvia Spengler: Program Director, Division of Information and Intelligent Systems, NSF Dietrich Rebholz-Schuhmann, Research Group Leader, Literature Analysis, European Bioinformatics Institute (EBI), UK [ELIXIR] • Publisher panel (Moderator: Lynette Hirschman) Anita de Waard: Disruptive Technologies Director, Elsevier Labs Virginia Benson Chanda, Editorial Manager, John Wiley & Sons Kevin Cohen, Biomedical Text Mining Group Lead, Center for Computational Pharmacology, U. Colorado School of Medicine
5:45 PM – 6:00 PM	Break
6:00 PM – 7:30 PM	**Reception and Poster Session**

September 14, 2010 (Tuesday)	
7:30 AM – 12:00 PM	**Registration**
7:30 AM – 8:30 AM	Breakfast (provided, outside of Ballroom A/B in the EMC Foyer)
8:30 AM – 9:00 AM	**Discussion of systems results (PPI)** – Martin Krallinger, CNIO
9:00 AM – 10:30 AM	**Challenge Participant systems 3 (PPI) –** chaired by Florian Leitner Team 65 University of Zurich by Fabio Rinaldi Team 69 Arizona State University by Robert Leaman Team 70 Universidade de Aveiro, IEETA by Sérgio Matos Team 73 NCBI/NIH by Sun Kim
10:30 AM – 11:00 AM	Break
11:00 AM – 12:30 PM	**Challenge Participant systems 4 (PPI) –** chaired by Martin Krallinger Team 81 University of Delaware and Indiana University by Hagit Shatkay and Luis Rocha Team 89 University of Wisconsin-Milwaukee by Shashank Agarwal Team 90 National Centre for Text Mining by Rafal Rak Team 100 NCBI\NLM\NIH by Rezarta Islamaj Dogan
12:30 PM – 1:00 PM	**Challenge Discussion Panel**
1:00 PM – 2:30 PM	Lunch (provided, hotel Restaurant)
2:30 PM – 3:30 PM	**Keynote: “**Text Bound Annotation and Evaluation -- The Perspectives of BioNLP Shared Tasks and GENIA” **Jun’ichi Tsujii**, Professor, Department of Computer Science, University of Tokyo, Japan. School of Computer Science, University of Manchester, UK
3:30 AM – 3:45 PM	Break
3:45 PM – 5:45 PM	**Developing Text Mining Systems for Biocuration and Biological Discovery: Developer Perspectives** •System Interoperability Gully Burns, Information Science Institute, University of Southern California Florian Leitner, Structural and Computational Biology Group, Spanish National Cancer Research Centre, Spain Andreas Prlic, Protein Data Bank, University of California, San Diego Dietrich Rebholz, European Bioinformatics Institute Karin Verspoor, Center for Computational Pharmacology and the Computational Bioscience Program University of Colorado
5:45 PM – 6:00 PM	Break
6:00 PM – 7:30 PM	**Systems Demo Session**

September 15, 2010 (Wednesday)	
7:30 AM – 8:30 AM	Breakfast (provided, outside of Ballroom A/B in the EMC Foyer)
8:30 AM – 8:45 AM	**Interactive task overview (IAT)** – Cecilia Arighi, CBCB, UD
8:45 AM – 10:20 AM	**Interactive task session 1** (**IAT) –** Chaired by Cecilia Arighi Team 61 "MyMiner", David Salgado, Australian Regenerative Medicine Institute - Monash University Team 65 "ODIN: The OntoGene Document Inspector", Fabio Rinaldi, University of Zurich Team 68 "Gene View", Philippe Thomas, Humboldt Universität zu Berlin Team 78 "Online Gene Indexing and Retrieval for BioCreative III at the University of Iowa", Sanmitra Bhattacharya, University of Iowa Team 89 "Interactive NLP system for normalized gene annotations", Shakshank Agarwal, University of Wisconsin-Milwaukee Team 93 "GNSuite", Rune Sætre, University of Tokyo
10:20 AM – 11:00 AM	**Interactive task session 2** (**IAT)** •Report from User Advisory Group – Andrew Chatr-aryamontri, BioGrid, Wellcome Trust Centre for Cell Biology, University of Edinburgh, UK
11:00 AM – 11:30 AM	Break
11:30 AM – 12:30 PM	**Developing Text Mining Systems for Biocuration and Biological Discovery: User Perspectives** •Discussion session **–** Led by Phoebe Roberts, Pfizer, USA -Metrics for evaluation User Studies and System Evaluation in Information Retrieval, Ben Carterette, Department of Information and Computer Sciences, University of Delaware
12:30 PM – 1:30 PM	Lunch (provided, hotel Restaurant)
1:30 PM – 2:30 PM	**Biocreative III and BioCreative IV** (Lynette Hirschman and Alfonso Valencia) •Special issue •Future task challenges
2:30 PM	Workshop closing

Overview Papers

Biocreative III Interactive Task: an Overview

Cecilia N. Arighi[1§], Lynette Hirschman[2] and Cathy H. Wu[1]

[1]Center for Bioinformatics and Computational Biology, University of Delaware
[2]The MITRE Corporation
[§]Corresponding author

Email addresses:
CNA: arighi@dbi.udel.edu
LH: lynette@mitre.org
CHW: wuc@udel.edu

Abstract

The interactive task (IAT) in Biocreative III is a demonstration task, and is focused on indexing (identifying which genes are being studied in an article and linking these genes to standard database identifiers) and gene-oriented document retrieval (identifying papers relevant to a selected gene) applied to full-length articles. A User Advisory Group (UAG), made up of curators and industry representatives, was set up to provide system requirements as well as testing of the systems. For the evaluation, six developer teams each provided an interface for testing by UAG members. The comparison of manual vs. system-assisted annotations will facilitate the definition of metrics and acquisition of data that are necessary for designing the evaluation of the interactive systems in the future BioCreative IV challenge. This will be discussed in the IAT session during the workshop.

Background

The biological literature represents the repository of biological knowledge. The ever increasing scientific literature now available electronically and the exponential growth of large-scale molecular data have prompted active research in biological text mining and information extraction to facilitate literature-based curation of molecular databases and bio-ontologies. Many text mining tools and resources have been developed and there are community efforts, including BioCreative, for evaluating text mining systems applied to the biological domain [1,2,3]. However, these tools are still not being fully utilized by the broad biological user communities. Such a gap is partly due to the intrinsic complexity of the biological literature for text mining, and partly to the lack of standards and limited interactions between the text mining and the user communities of biological researchers and database curators. Previous BioCreative challenges have involved experienced curators from specialized databases (e.g., the MINT, BioGrid and IntAct protein-protein interaction databases in Biocreative II, and II.5), to generate gold standard data for training and testing of the systems. However, after that step, there was no curator intervention. Although this approach is valuable to address the system performance, it did not address system usage and adoption by curators or biologists. Therefore, in Biocreative III we introduced a demonstration interactive task (IAT) that aims at providing key component modules for text mining services for biocuration. Our goal in BioCreative III has been to provide an interface to support curators and other specialized end users, with results that can be integrated in the curation workflow.

Results

Establishment of the User Advisory Group

A critical aspect of the BioCreative III is the active involvement of the end users to guide development and evaluation of useful tools and standards. This prompted the creation of the UAG by recruiting members from major public databases as well as a small number of interested advisors from industry. This group is currently composed of 13 members (http://www.biocreative.org/about/biocreative-iii/UAG/)

The roles of the User Advisory Group included:

- **Developing end user requirements for interactive text mining tools:** the UAG provided the guidelines for the requirements delivered to the participants in the BioCreative III interactive task(http://www.biocreative.org/tasks/biocreative-iii/iat/).
- **Providing users for the interactive task:** UAG members tested the systems, and provided feedback.
- **Providing gene normalization annotation of a corpus of full text articles** for use in developing baseline statistics (inter-annotator agreement, and time for task completion) as well as a gold standard of articles correctly annotated for gene/protein normalization.

Interactive Task

Monthly discussions with the UAG provided important insight into what would be of general interest for literature curation. The group was able to identify a common need underlying the different specific curation tasks (e.g. model organism database, protein-protein interaction database, and protein sequence database); this common need was the identification, within an article of the genes/proteins that have some experimental data, and their linkage to appropriate database identifiers (e.g., EntrezGene or UniProt identifiers). For this activity it is critical to consider the full-length article in order to rank the corresponding genes by relevance in context of the overall article. A natural extension to this task is the retrieval of additional articles for which the gene in question has experimental information. So in addition to a gene normalization and ranking task, a document retrieval task was included.

IAT System Requirements

A web based interface should be user friendly and assist curators to easily find the desired information.

Indexing task: For this subtask, the input requirement was a PubMed Central ID, and the output would return a list of gene/protein identifiers linked to the appropriate database identifiers from the selected full-text article. The list of genes/proteins should be ranked for their importance or "centrality" to the article. Such a ranking could, for example, take into consideration the frequency of gene/protein mentions, but might also weight the sections of the article where the gene is mentioned. For example, a gene with associated experimental results that is mentioned with low frequency should rank higher than a high frequency mentioned gene with no experimental results.

Retrieval task: For this subtask, the input was a user-selected gene; the system output was a ranked list of documents from PubMedCentral (with links to the full text) which would be relevant to provide information on the selected gene.

IAT System Testing

Six systems were made available for testing. All systems were run against the same set of articles. Members from the UAG curated the papers using the systems. To familiarize her/himself with the system, each evaluator first went over an article previously curated by the group (they were familiar with results). Each evaluator was assigned a primary system, but could access others as well. Curators recorded the time spent with the system while curating and also answered a questionnaire related to interface usability and task performance. The results were collected and compared to the manually annotated set.

Discussion

IAT in Biocreative III is a demonstration task and was designed to facilitate the definition of metrics and acquisition of data that are necessary for designing the evaluation of the interactive systems in the BioCreative IV challenge. Both the participating teams and the UAG are instrumental in accomplishing this goal. The analysis of the overall results will be presented and discussed during the IAT workshop session.

Methods

All information about the IAT task is available at http://www.biocreative.org/tasks/biocreative-iii/iat/). The full text articles in XML format from the PubMed Central Open Access collection was made available at http://www.biocreative.org/resources/corpora/biocreative-iii-corpus/, IAT PubMedCentral XML Data.

Acknowledgements

We would like to acknowledge the UAG and participating systems for their participation and key contributions to this effort, as well as the six participating groups who fielded systems for evaluation. This work was supported under NSF grant DBI-0850319.

References

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

2. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, Hirschman L, Valencia A: **Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.** *Genome Biol.* 2008;9 Suppl 2:S1. Epub 2008 Sep 1.

3. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An Overview of BioCreative II.5.,** *IEEE/ACM Trans Comput Biol Bioinform***.** 2010 Jul-Sep;7(3):385-99.

Results of the BioCreative III Interaction Method Task

Martin Krallinger[1], Miguel Vazquez[1], Florian Leitner[1], and Alfonso Valencia[1*]

[1]Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

*Corresponding author

Email addresses:

MK: mkrallinger@cnio.es

MV: mvazquezg@cnio.es

FL: fleitner@cnio.es

AV: avalencia@cnio.es

Abstract

Background

A considerable effort has been made to standardize the annotation process of protein interaction data through the development of the Molecular Interaction (MI) ontology. Among other terms, it contains a hierarchy for the experimental context used to determine protein-protein interactions (PPIs). These experimental approaches provide qualitative information on the type and reliability of an interaction. The Interaction Method Task (IMT) aims to promote the implementation of automated systems for detecting associations between articles and these interaction detection method concepts with the goal of facilitating the manual curation strategies.

Results

A total of eight teams submitted predictions for the IMT. Each team was allowed to send up to 5 runs, plus an additional 5 using the BioCreative Meta-Server; in total, we received 42 runs. These were compared to a Gold Standard, manually generated annotations done by trained domain experts from the BioGRID and MINT databases. The annotations consisted in associations of full text articles to the interaction detection method concepts as defined in the MI ontology that supported protein interactions described in the articles. The highest AUC iP/R achieved by any run was 53%, the best MCC score 0.55. In case of competitive systems with an acceptable recall (above 35%) the macro-averaged precision ranged between 50% and 80%, with a maximum F-Score of 55%.

Conclusions

Participating systems where able to achieve competitive results despite the difficulties of this task: the variability of method term mentions, challenges due to pre-processing of full text articles provided as PDF files, and the heterogeneity and different granularity of method term concepts encountered in the ontology. As the document associations had to be accompanied with supporting evidence text passages for human interpretation, such systems may serve to generate text-mining assisted coarse level annotations efficiently. In the case of online systems (team 89), full-text articles can be annotated with method concepts on average in 3.7 seconds (sd: +/-0.35 sec), while still achieving competitive results; thus, such tools could be integrated into biological annotation workflows.

Background

Biomedical sciences require a strong support of generated discoveries by their experimental approach. The experimental context is crucial for the interpretation of biological assertions as well as to determine the reliability of given biological finding [1]. An important aspect for the annotation of protein interactions is to identify the experimental techniques ("interaction detection methods") described in an article to support the interactions [2]. Annotation of experimental techniques or "evidence" is also common with other annotation efforts, such as the Gene Ontology Annotations (GOA; in the form of evidence codes) [3]. Knowing the experimental method that provided the evidence for an interaction serves as "credibility" or likelihood indicator that the reported interaction actually occurs in a living organism ("*in vivo*") or cell culture ("*in vitro*").

These types of text classification tasks are based on associating standardized terms from a controlled vocabulary to the text in question. In the case of protein-protein interaction annotations, efforts have been made to develop a controlled vocabulary ("ontology") about interaction detection methods in order to standardize the terminology serving as experimental evidence support. A considerable amount of database curation work is devoted to the manual extraction of the experimental evidence support for the interaction pairs described in articles [4]. A relevant work with this respect was the implementation of a system for detecting experimental techniques in biomedical articles by Oberoi and colleagues [5]. Also the construction of a text mining system with a particular focus on interaction detection methods using statistical inference techniques has been explored recently [6], motivated by the Interaction Method Task of the BioCreative II challenge [7], where two different teams provided results [8, 9].

For BioCreative III, participants were asked to provide a list of interaction detection methods identifiers for a set of full-text articles (publications), ordered by their likelihood of having been used to detect the PPIs described in each article. These identifiers correspond to the standardized experimental detection method terms from the PSI-MI ontology for an experimental detection method. The evaluation of the results is oriented to lessen to database curation efforts by providing with a list of the most likely PSI-MI identifiers so as to facilitate their identification and subsequent assignment.

Results

Overview

In total, eight teams participated in this task. The official evaluation results of each run are shown in Table 2, measuring the performance on the documents for which the system provided results. The evaluation of the overall performance of the systems on the whole test set is shown in Table 3. The team information is shown in Table 1. Teams could participate offline, sending the results via e-mail, as well as online via the BioCreative Meta-Server (BCMS) [10]. The highest AUC iP/R achieved by any run was 53%, the best MCC score measured was 0.55 (see Methods for a quick explanation of these two scoring schemas).

Team	Contact/Leader	Organization
65	Fabio Rinaldi	University of Zurich
69	Robert Leaman	Arizona State University
70	Sérgio Matos	Universidade de Aveiro, IEETA
81	Luis Rocha	Indiana University
88	Ashish Tendulkar	IIT Madras
89	Shashank Agarwal	University of Wisconsin-Milwaukee
90	Xinglong Wang	National Centre for Text Mining
100	Zhiyong Lu	NCBI\NLM\NIH

Table 1: IMT Participants

List of IMT participants by team ID, team leader/main contact and institution.

Team	Run/Srvr	Docs	Precision	Recall	F1 Score	AUC iP/R
T65	RUN_1	**222**	9.35%	**83.21%**	0.16322	0.47884
T65	RUN_2	**222**	2.45%	**100.00%**	0.04750	0.44034
T65	RUN_3	**222**	9.99%	79.38%	0.17163	0.47650
T65	RUN_4	**222**	33.48%	42.88%	0.35403	0.30927
T65	RUN_5	**222**	*2.44%*	**100.00%**	*0.04735*	0.50111
T69	RUN_1	214	54.87%	57.91%	0.52392	0.52112
T69	RUN_2	211	57.01%	57.35%	0.53415	0.51844
T69	RUN_3	203	60.24%	56.41%	0.54454	0.51470
T69	RUN_4	199	62.46%	55.17%	**0.55060**	0.51013
T69	RUN_5	190	64.24%	52.44%	0.54354	0.49390
T70	RUN_1	143	51.78%	35.01%	0.37838	0.31402
T70	RUN_2	72	71.76%	36.81%	0.45608	0.36215
T70	RUN_3	*30*	**80.00%**	41.50%	0.51508	0.41500
T70	RUN_4	205	31.65%	38.72%	0.31747	0.32295
T70	RUN_5	159	36.36%	*21.26%*	0.24754	0.18976
T81	RUN_1	**222**	4.44%	63.91%	0.08191	0.22022
T81	RUN_2	221	9.39%	41.92%	0.14117	0.19766
T81	RUN_3	**222**	13.51%	28.35%	0.17414	*0.17010*
T81	RUN_4	**222**	13.21%	29.57%	0.17341	0.20388
T81	RUN_5	209	21.93%	24.64%	0.21339	0.18733
T88	RUN_1	219	29.10%	45.04%	0.33601	0.38590
T88	RUN_2	220	28.67%	45.53%	0.33353	0.38373
T89	RUN_1	200	54.78%	53.37%	0.50905	0.46061
T89	RUN_2	200	54.95%	53.23%	0.50760	0.46423
T89	RUN_3	201	54.05%	53.25%	0.50234	0.45330
T89	RUN_4	199	54.48%	54.18%	0.51254	0.47211
T89	RUN_5	201	55.30%	56.12%	0.52377	0.47807
T89	SRVR_4	200	55.33%	55.61%	0.52112	0.47636
T89	SRVR_5	199	54.09%	54.00%	0.50962	0.47650
T89	SRVR_6	201	55.14%	56.12%	0.52350	0.48047
T89	SRVR_7	203	50.46%	55.66%	0.50064	0.47392
T89	SRVR_8	199	54.04%	54.05%	0.50840	0.47534
T90	RUN_1	200	56.11%	51.59%	0.50720	0.44687
T90	RUN_2	203	56.37%	53.19%	0.51203	0.47159
T90	RUN_3	217	55.29%	59.90%	0.54616	**0.52974**
T90	RUN_4	177	63.98%	46.89%	0.51355	0.44118
T90	RUN_5	164	66.26%	46.78%	0.52021	0.44458
T100	RUN_1	213	47.26%	54.97%	0.47062	0.43312
T100	RUN_2	**222**	41.19%	54.61%	0.44178	0.43238
T100	RUN_3	**222**	35.29%	45.53%	0.37496	0.32459
T100	RUN_4	**222**	35.29%	45.53%	0.37496	0.32459
T100	RUN_5	125	56.40%	30.65%	0.37011	0.29387
Team	**Run/Srvr**	**Docs**	**Precision**	**Recall**	**F1 Score**	**AUC**

Table 2: Primary evaluation results on the annotated documents

Macro-averaged results when evaluating only documents for which the system reported results (i.e., measuring the average per-document performance only on the documents each run produced annotations for). The highest score for each evaluation column is show in bold typeface, the lowest in italics. **Run/Srvr**: RUN=offline run, SRVR=online server run via BCMS; **Docs**: number of documents annotated; **AUC iP/R**: Area under the interpolated precision/recall curve.

Team	Run/Srvr	Precision	Recall	F1 Score	MCC	AUC iP/R
T65	RUN_1	8.77%	84.82%	0.15893	0.23552	0.27588
T65	RUN_2	2.45%	**100.00%**	0.04779	0.06259	0.24484
T65	RUN_3	9.42%	81.78%	0.16892	0.24172	0.27727
T65	RUN_4	33.48%	42.32%	0.37385	0.36166	0.14169
T65	RUN_5	*2.44%*	**100.00%**	*0.04763*	*0.06193*	0.29016
T69	RUN_1	52.07%	55.03%	0.53506	0.52519	0.34302
T69	RUN_2	54.34%	53.51%	0.53920	0.52958	0.33824
T69	RUN_3	57.36%	50.29%	0.53589	0.52796	0.32539
T69	RUN_4	59.25%	48.01%	0.53040	0.52456	0.31711
T69	RUN_5	61.33%	43.64%	0.50998	0.50896	0.29373
T70	RUN_1	48.61%	23.15%	0.31362	0.32617	0.12949
T70	RUN_2	70.00%	11.95%	0.20421	0.28419	0.08731
T70	RUN_3	**80.65%**	*4.74%*	0.08961	0.19270	*0.03826*
T70	RUN_4	31.22%	36.43%	0.33625	0.32216	0.15688
T70	RUN_5	32.69%	15.94%	0.21429	0.21717	0.05734
T81	RUN_1	4.54%	66.03%	0.08496	0.11406	0.07716
T81	RUN_2	8.71%	42.13%	0.14430	0.15560	0.06239
T81	RUN_3	13.51%	28.46%	0.18326	0.17168	0.04657
T81	RUN_4	13.20%	27.70%	0.17881	0.16667	0.05601
T81	RUN_5	21.35%	22.20%	0.21767	0.20090	0.05283
T88	RUN_1	28.44%	45.16%	0.34897	0.34146	0.20244
T88	RUN_2	28.17%	45.92%	0.34921	0.34263	0.20069
T89	RUN_1	52.52%	49.53%	0.50977	0.49997	0.28202
T89	RUN_2	52.02%	48.96%	0.50440	0.49451	0.28589
T89	RUN_3	50.78%	49.34%	0.50048	0.49016	0.27238
T89	RUN_4	52.50%	49.91%	0.51167	0.50181	0.29220
T89	RUN_5	52.58%	52.18%	0.52381	0.51382	0.29980
T89	SRVR_4	52.71%	51.61%	0.52157	0.51163	0.29926
T89	SRVR_5	52.28%	50.10%	0.51163	0.50168	0.30046
T89	SRVR_6	52.28%	52.18%	0.52232	0.51226	0.30049
T89	SRVR_7	49.55%	52.56%	0.51013	0.49972	0.29303
T89	SRVR_8	51.76%	50.29%	0.51011	0.49999	0.29766
T90	RUN_1	53.33%	47.06%	0.50000	0.49113	0.26805
T90	RUN_2	52.56%	48.77%	0.50591	0.49625	0.28386
T90	RUN_3	52.30%	58.25%	**0.55117**	**0.54201**	**0.35423**
T90	RUN_4	61.09%	38.14%	0.46963	0.47436	0.25209
T90	RUN_5	64.24%	35.10%	0.45399	0.46707	0.24270
T100	RUN_1	44.59%	51.61%	0.47845	0.46794	0.26055
T100	RUN_2	39.86%	54.84%	0.46166	0.45448	0.26982
T100	RUN_3	35.29%	44.59%	0.39396	0.38240	0.15734
T100	RUN_4	35.34%	44.59%	0.39430	0.38271	0.15758
T100	RUN_5	54.86%	18.22%	0.27350	0.30847	0.11109
Team	**Run/Srvr**	**Precision**	**Recall**	**F1 Score**	**MCC**	**AUC iP/R**

Table 3: Evaluation of whole document set performance

Micro-averaged results when evaluating all documents (i.e., measuring the overall performance of each run on the whole document set). The highest score for each evaluation column is show in bold typeface, the lowest in italics. **Run/Srvr**: RUN=offline run, SRVR=online server run via BCMS; **MCC**: Matthew's Correlation Coefficient; **AUC iP/R**: Area under the interpolated precision/recall curve (micro-averaged by iterating over the precision/recall values of the highest ranked annotation of all articles, then all second ranked annotations, etc.).

Timing Measures

By using the BCMS framework for participating online, we were able to measure the time it took the systems to report interaction method identifiers for full-text articles. However, there was only one team (89) participating online in this task, albeit with 5 servers and quite competitive results. This team annotated a full-text article on average in 3.7 seconds (sd: +/-0.35 sec), achieved a maximum F-Score score of 52% with an AUC iP/R of 48% (see Methods for an explanation of these measures).

Comparing Participating Systems

The participants were asked to fill in a short questionnaire, and all participants responded. Only one team (81) used other sources of training data than what was provided through the challenge itself, one team made use of the UMLS (69) and two of MeSH terms (90, 100). Most teams relied on the provided text we extracted using the UNIX tool "*pdftotext*", while two teams (65, 100) made use of the PDFs directly. Most teams incorporated lexical analysis of the text (sentence splitting, tokenization and/or lemmatization/stemming), quite a few looked at n-gram tokens (teams 81, 89, 90, 100), but only one also included Part-of-Speech-tagging (team 90), and, interestingly, some teams omitted a specialized Named Entity Recognition approach (NER; teams 81, 89, 100; instead using regex matching). Team 90 even made use of shallow parsing techniques. All teams except 81 and 90 relied on Bag-of-Word vectors, and teams 70 and 88 did not use any supervised classifiers. Teams 90 and 69 were the only teams to use a Logistic Regression classifier trained on each term, team 90 also applied a Support Vector Machine, and team 69 used MALLET for NER. Other than that, no team reported to have made use of existing BioNLP libraries and instead they relied on in-house tools. Only teams 90 and 65 applied gene/protein mention detection. We also asked teams to evaluate the difficulty of the task (easy, medium, hard); No team thought the task was easy, half (70, 89, 90, 100) said is was hard, while the other four classified it as "medium".

Discussion

Given the performance, it is possible that humans could make use of the provided results in a limited number of ways. Mostly, the short time needed by the servers makes it seem reasonable that online, automated systems could be used for this task.

However, when looking at the overall performance of the best result on the whole set (F-Score 55%, AUC iP/R 35%, MCC 0.54), it becomes apparent that automated annotation of interaction methods is not yet "solved" and more work in this area is required. It seems that there is still much room for improvement in the future. However, for the purpose of helping curators or biologist to identify interaction methods in an article, the results might be sufficient given that systems are doing an acceptable job at articles they identify as relevant (highest AUC iP/R: 53%). This could be further the case if the evidence passages the teams were asked to provide for their annotations are meaningful – a future evaluation target.

Conclusions

In summary, it is likely the future will see significant improvements for this kind of nearly novel entity normalization task, and the current results are possibly promising enough to aid humans in the annotation process. Performance will hopefully increase with the amount of readily available training data and as more interest in this particular area of entity types is raised. On the positive side, the relatively good performance (w.r.t. the global results) of the online team (89) combined with their very competitive server annotation times (3.7 sec/article) clearly demonstrates that online, high-quality BioNLP can be implemented in ways where processing times are acceptable to serve end-users on demand.

Methods

Result Structure

For each article, participants had to return zero or more PSI-MI detection method term identifiers, and for each term annotation they had to provide a confidence score (in the range (0,1]), and an overall (unique) rank for each term annotated on an article, from the most to the least relevant. In addition, participants were asked to return the most decisive text passage that gave rise to their annotation - data that we will need to evaluate manually at a later stage. For each participation methods – offline via e-mail, online via the BCMS - teams could submit five runs for a total of ten if they participated in both settings.

Task Data

Participants were supplied with 2,035 articles as training set, received an additional 587 articles as development set shortly before the test phase, and were then tested on 305 unseen publications, 222 of which were annotation-relevant articles. Both the training and test set have a highly distorted representation of the 115 possible method detection terms found in PSI-MI, with only 4 methods representing roughly half of all annotations made on the articles, both in the training and the test set. These 4 high-frequency terms are (from most to least frequent): “anti bait coimmunoprecipitation”, “anti tag coimmunoprecipitation” (these two represent 1/3 of all annotations), “pull down”, and “two hybrid”.

Evaluation

The evaluation is based on comparing automatically generated PSI-MI IDs (terms) with a manual annotated set of 222 full-text publications (“test set”, see Task Data). The evaluation functions are the same as already discussed for BioCreative II.5 [11]: The overall set performances of the systems are evaluated using the “traditional” F-Score (using micro-averaging) and the MCC score (Matthew’s Correlation Coefficient). Similar to the Article Classification Task evaluation, the primary evaluation score is based on the average per-article annotation performance (macro-averaging) given its ranking; To this end, the area under the (interpolated) Precision/Recall curve is measured (AUC iP/R), averaging the AUC from the individual scores on each article. Each team was allowed to submit five runs, and teams participating online (via the BCMS), could submit another five runs via their servers (i.e., independent of the number of “offline” runs), using the same setting as for BioCreative II.5. This online setting allowed us to measure how much time it took the servers to generate the annotations.

Authors' contributions

MK was the main organizer of the task, managed the teams, assembled the corpus and arranged for the Gold Standard annotations. MV contributed to corpus assembly, evaluation and the software development. FL wrote the software used for this task and ran the online challenge via the BCMS as well as the initial evaluation. AV, as a lead organizer of BioCreative, was the responsible for this task**.**

Acknowledgements

We especially want to thank the BioGRID (Andrew Chatr-aryamontri) and MINT (Gianni Cesareni) databases and their curators for generating the data collections for this task; without their enormous efforts, the IMT would not have been possible.
The work on BioCreative was supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and by grants BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovación and by the Spanish National Bioinformatics Institute (http://www.inab.org/).

References

1. Krallinger M: **Importance of negations and experimental qualifiers in biomedical literature**. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* 2010:46-49.
2. Orchard S, Montecchi-Palazzi L, Hermjakob H, Apweiler R: **The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments**. *Pac Symp Biocomput* 2005:186-196.
3. Consortium GO: **The Gene Ontology project in 2008**. *Nucleic Acids Research* 2008, **36**(Database issue):D440-444.
4. Chatr-aryamontri A, Kerrien S, Khadake J, Orchard S, Ceol A, Licata L, Castagnoli L, Costa S, Derow C, Huntley R *et al*: **MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data**. *Genome Biol* 2008, **9 Suppl 2**:S5.
5. Oberoi M, Struble C, Sugg S: **Identifying experimental techniques in biomedical literature**. *Proc Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis* 2006:122-123.
6. Wang H, Huang M, Zhu X: **Extract interaction detection methods from the biological literature**. *BMC Bioinformatics* 2009, **10 Suppl 1**:S55.
7. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II**. *Genome Biol* 2008, **9 Suppl 2**:S4.
8. Ehrer F, Gobeill J, Tbahriti I, Ruch P: **GeneTeam site report for BioCreative II: customizing a simple toolkit for text mining in molecular biology.** *Proceedings of the BioCreative II Workshop* 2007:199-207.
9. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Clematide S, Hess M, von Allmen J-M, Parisot P, Romacker M *et al*: **OntoGene in BioCreative II**. *Genome Biol* 2008, **9 Suppl 2**:S13.
10. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo C-J, Hsu C-N, Tsai RT-H, Hung H-C, Lau WW *et al*: **Introducing meta-services for biomedical information extraction**. *Genome Biol* 2008, **9 Suppl 2**:S6.
11. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An Overview of BioCreative II.5**. *IEEE/ACM Trans Comput Biol Bioinform* 2010, **7**(3):385-399.

Results of the BioCreative III (Interaction) Article Classification Task

Martin Krallinger[1], Miguel Vazquez[1], Florian Leitner[1], David Salgado[2], and Alfonso Valencia[1*]

[1]Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

[2]ARMI - Australian Regenerative Medicine Institute,
EMBL Australia

*Corresponding author

Email addresses:

MK: mkrallinger@cnio.es

MV: mvazquezg@cnio.es

FL: fleitner@cnio.es

AV: avalencia@cnio.es

Abstract

Background

The detection of annotation-relevant articles is a common step required by annotation databases. In case of complex biological events such as physical protein-protein interactions (PPIs) for which a considerable amount of articles are published each month, the use of keyword-based search strategies may not provide satisfactory results. Additionally, there exists a general interest in the biological community in determining if a given article describes the characterization of interactions. This motivates the construction of automated systems able to classify and rank large sets of potentially relevant abstracts. To build such systems for detecting PPI-describing articles, participating teams were provided with a balanced training set of 2,280 abstracts in total, a development set of 4,000 abstracts reflecting the same class imbalance as the test set (15% positive examples), and a set of 6,000 abstracts were used as test set. Domain experts had manually labelled all data collections.

Results

We measured the performance of ten participating teams in this task, each of whom could submit five runs (offline, all teams) and another five online via the BioCreative Meta-Server (only teams 81 and 89 took advantage of this option), for a total of 52 runs. The highest MCC score measured was 0.55 at an accuracy of 89%, the best AUC iP/R was 68%. Most of the participating teams relied on machine learning methods, and some of them explored the use of lexical resources such as MeSH terms, PSI-MI concepts or particular lists of verbs and nouns. Some integrated NER approaches.

Conclusions

With the current state-of-the art performance, text-mining tools for article classification can be used to report ranked lists of relevant articles for manual selection (68% AUC iP/R). To rely purely on automated results would require either further improvement of the systems or determining more stringent selection cut-offs, as they still return a high number of false positives (1/3 for the best result) and miss many relevant articles (43%). This issue may be partially explained by a general problem for text classification of biomedical literature, consisting in the considerable class imbalance between relevant and non-relevant articles.

Background

The selection of relevant articles for further manual inspection to derive biological annotations is a common step across almost all biological annotation databases [1]. Commonly, relevant articles are defined as a list of PubMed entries, derived from a keyword search or a journal of interest. Often, such search strategies are carried out periodically to articles that will be examined more carefully during the database curation process [2]. However, in case of complex biological events like protein-protein interactions (PPIs), simple keyword queries are often very inefficient in detecting relevant articles [3]. Therefore, promoting the development of article classification systems have a long tradition, e.g. the TREC Genomics tracks [4].

The aim of the (Interaction) Article Classification Task is to promote the development of automated systems that are able to classify articles as relevant for protein-protein interaction (PPI) database curation efforts. The resulting text mining tools should be able to simplify the identification of relevant articles [5] for a range of journals known to publish protein interaction reports. This task was inspired by former challenges, namely BioCreative II [6, 7] and II.5 [8], with specific modifications that address practical aspects of the resulting systems. These changes include:

a) The use of PubMed abstracts as opposed to full text articles, as they are not subjected to existing hurdles in case of availability and formatting surrounding full text articles.
b) The use of a large range of journals considered as relevant by biological databases to avoid inclusion of journals that have no usefulness in terms of the curation process or are not published any longer.
c) The construction of a large manually classified training, development, and test set to enable the implementation of supervised learning methods and a statistical sound evaluation.
d) The use of a publication time range selection criteria to focus on recent articles and provide a more coherent data collection.
e) A sampling of articles in the case of the development and test set that reflects the real class imbalance encountered for these journals.

Participating systems were allowed to use any additional information provided through the PubMed abstract, such as the linked publication or the MeSH (Medical Subject Heading) terms, but from an evaluation perspective, only content from the abstracts was considered for the creation of the (human) Gold Standard annotations. The Gold Standard annotations were generated by domain experts through inspection of a randomly sampled set of abstracts following classification guidelines which were refined during several rounds of classification based on the feedback of the BioGRID and MINT database curators.

Team	Contact/Leader	Organization
65	Fabio Rinaldi	University of Zurich
70	Sérgio Matos	Universidade de Aveiro, IEETA
73	W John Wilbur	NCBI
81	Luis Rocha	Indiana University
88	Ashish Tendulkar	IIT Madras
89	Shashank Agarwal	University of Wisconsin-Milwaukee
90	Xinglong Wang	National Centre for Text Mining
92	Keith Noto	Tufts University
100	Zhiyong Lu	NCBI\NLM\NIH
104	Jean-Fred Fontaine	Max Delbrück Center

Table 1: ACT Participants

List of ACT participants by team ID, team leader/main contact and institution.

Team	Run/Srvr	Accuracy	Specificity	Sensitivity	F-Score	MCC	AUC iP/R
T65	RUN_1	88.68%	97.64%	38.57%	50.83%	0.48297	63.85%
T65	RUN_2	87.93%	93.07%	59.23%	59.82%	0.52727	63.89%
T65	RUN_3	67.05%	64.19%	83.08%	43.34%	0.34244	41.74%
T65	RUN_4	73.68%	74.13%	71.21%	45.08%	0.34650	41.74%
T65	RUN_5	88.00%	94.40%	52.20%	56.89%	0.50255	62.39%
T70	RUN_1	*56.45%*	*49.70%*	94.18%	39.62%	0.31789	56.76%
T70	RUN_2	87.41%	96.11%	38.79%	48.32%	0.43346	56.76%
T70	RUN_3	81.92%	83.61%	72.53%	54.91%	0.46563	56.76%
T70	RUN_4	47.77%	39.04%	**96.59%**	35.95%	0.27060	56.76%
T70	RUN_5	86.84%	98.62%	20.99%	32.62%	0.34488	56.76%
T73	RUN_1	87.55%	91.81%	63.74%	60.83%	0.53524	65.91%
T73	RUN_2	**89.15%**	94.95%	56.70%	61.32%	**0.55306**	67.96%
T73	RUN_3	87.78%	92.61%	60.77%	60.14%	0.52932	65.89%
T73	**RUN_4**	88.88%	94.34%	58.35%	**61.42%**	0.55054	**67.98%**
T73	RUN_5	87.62%	92.18%	62.09%	60.33%	0.53031	65.37%
T81	RUN_1	59.03%	58.76%	60.55%	30.96%	0.13949	19.93%
T81	RUN_2	58.47%	57.86%	61.87%	31.12%	0.14219	19.69%
T81	RUN_3	25.37%	14.72%	84.95%	25.66%	-0.00344	15.66%
T81	RUN_4	63.45%	69.16%	31.54%	20.74%	0.00538	16.20%
T81	RUN_5	69.17%	77.35%	23.41%	18.72%	0.00645	*15.63%*
T81	SRVR_10	85.38%	99.61%	5.82%	10.78%	0.17771	50.25%
T81	SRVR_11	84.73%	99.86%	0.11%	*0.22%*	*-0.00272*	46.02%
T81	SRVR_12	84.30%	98.86%	*2.86%*	5.23%	0.05244	32.11%
T81	SRVR_13	84.88%	99.92%	0.77%	1.52%	0.05791	18.59%
T81	SRVR_9	84.88%	**99.98%**	0.44%	0.88%	0.05220	44.19%
T88	RUN_1	42.63%	35.11%	84.73%	30.94%	0.15238	21.97%
T88	RUN_2	56.92%	53.73%	74.73%	34.47%	0.20417	26.04%
T89	RUN_1	80.02%	80.90%	75.06%	53.26%	0.44911	61.29%
T89	RUN_2	81.00%	81.75%	76.81%	55.08%	0.47242	62.13%
T89	RUN_3	82.40%	83.85%	74.29%	56.15%	0.48180	60.48%
T89	RUN_4	87.73%	94.79%	48.24%	54.40%	0.47967	43.76%
T89	RUN_5	87.27%	91.81%	61.87%	59.58%	0.52082	48.47%
T89	SRVR_4	77.80%	77.84%	77.58%	51.46%	0.43152	57.44%
T89	**SRVR_5**	78.05%	78.15%	77.47%	51.71%	0.43424	57.56%
T89	SRVR_6	79.90%	81.00%	73.74%	52.67%	0.44073	54.97%
T89	SRVR_7	86.25%	92.06%	53.74%	54.24%	0.46156	41.58%
T89	SRVR_8	86.87%	90.39%	67.14%	60.80%	0.53336	47.40%
T90	RUN_1	88.73%	95.15%	52.86%	58.73%	0.52736	51.14%
T90	RUN_2	88.70%	94.97%	53.63%	59.01%	0.52890	51.65%
T90	RUN_3	88.32%	93.93%	56.92%	59.64%	0.52914	65.24%
T90	RUN_4	88.93%	96.03%	49.23%	57.44%	0.52237	49.26%
T90	RUN_5	88.60%	95.05%	52.53%	58.29%	0.52204	50.83%
T92	RUN_1	86.22%	90.77%	60.77%	57.22%	0.49155	50.99%
T100	RUN_1	88.77%	96.82%	43.74%	54.15%	0.50005	61.62%
T100	RUN_2	88.27%	93.89%	56.81%	59.49%	0.52732	61.86%
T100	RUN_3	81.13%	82.69%	72.42%	53.80%	0.45256	60.25%
T100	RUN_4	81.85%	82.85%	76.26%	56.04%	0.48270	63.75%
T104	RUN_1	80.12%	80.69%	76.92%	53.99%	0.45999	53.67%
T104	RUN_2	80.07%	80.47%	77.80%	54.21%	0.46370	53.67%
T104	RUN_3	64.93%	59.86%	93.30%	44.66%	0.38161	53.67%
T104	RUN_4	69.78%	66.25%	89.56%	47.34%	0.40530	53.67%
T104	RUN_5	86.27%	98.47%	18.02%	28.47%	0.30064	53.67%
Team	**Run/Srvr**	**Accuracy**	**Specificity**	**Sensitivity**	**F-Score**	**MCC**	**AUC iP/R**

Table 2: (previous page) Article Classification Task results

Preliminary evaluation results based on the unrefined Gold Standard, in terms of Accuracy, MCC Score and AUC iP/R. The highest score for each evaluation column is show in bold typeface, the lowest in italics. Also, the best offline and online run is marked in bold. **Legend:** Run/Srvr (RUN=offline run/SRVR=online run via the BCMS), MCC (Matthew's Correlation Coefficient), P@full_R (Precision at full Recall), AUC iP/R (Area under the interpolated Precion/Recall curve).

Results

Overview

In total, ten teams participated in this task. The individual results of each run are shown in Table 2, the team ID associations are show in Table 1. Teams could participate offline, sending their results via e-mail, as well as online via the BioCreative Meta-Server [9]. For each of these participation methods, teams could submit five runs for a total of ten if they participated both offline and online. The highest AUC iP/R achieved by any run was 68%, the best MCC score measured was 0.55 (see Methods for an explanation of these scoring schemas).

Timing Measures

By using the BioCreative Meta-Server (BCMS) framework for participating online, we were able to measure the time it took the systems to report a classification. There were two teams (81 and 89) participating online, each with 5 servers. Team 81 annotated an article in 20 seconds (average, sd: +/-12 sec) and although the maximum MCC score was only 0.11 (see Table 2, Server 10), their best AUC iP/R was a respectable 50% (Server 10) (see Methods for an explanation of the evaluation measures). The second team, 89, did even better with an average of 1.9 seconds (sd: +/- 0.57 sec) per abstract achieving a maximum MCC score of 0.61 (Server 8); their best AUC iP/R score was 58% (Server 5).

Comparing Participant Systems

The participants were asked to fill in a short questionnaire after the test phase. Interestingly, four teams (73, 81, 92, 100) used other sources of training data than what was provided through the challenge itself (e.g., data from former BioCreative challenges). We also asked teams to evaluate the difficulty of the task (easy, medium, hard); No team thought the ACT task was easy, four (73, 81, 100, 104) said is was hard, while the others classified it as "medium". All teams did some amount of lexical analysis of the text (sentence splitting, tokenization and/or lemmatization/stemming was done by all teams), and many included Part-of-Speech-tagging (teams 65, 73, 89, 90) or even Named Entity Recognition (teams 65, 70, 73, 81, and 90). Team 73 even used dependency parsing on the abstracts.

For generating their predictions all teams relied on the title and abstract, half used the MeSH terms, too, and one teams even was also able to explore full text information for some of the articles. For feature selection or weighting purposes, approaches used by participating teams include statistical methods like Chi-Square, mutual information, frequency cut-off and Bayesian weights as well as other selection criteria such as the restriction to particular Part-of-Speech types. Teams 81, 89, 100 and 104 also used dimensionality reduction techniques on their features.

A common characteristic of most of the participating teams was the use of machine learning techniques in general. Half of them used Support Vector Machines (SVM) for the classification (teams 81, 89, 90, 92, 100), and most of those combined the SVM with other supervised methods (81: (their own) Linear VTT classifier, 89: Naïve Bayes, 90: Logistic Regression, 100: Nearest Neighbour). Team 70 used Nearest Neighbour, 104 Naïve Bayes, 73 Large Margin class./Huber loss function, and team 65 used a Max. Entropy classifier.

Discussion

Given the performance of systems, for example the high-AUC-iP/R servers, it is likely that humans could make use of the provided results to quickly identify the most relevant articles in a set. Therefore, the time spent by the text mining pipelines should be put in contrast to the time a human would need to select relevant articles, a number we will establish in future work with the annotators and curators who provided the Gold Standard. It gives rise to reasonable belief that online, automated systems could have a strong impact on reducing the time required to locate relevant articles.

Conclusions

As the best run in terms of MCC score is 0.55, with a Precision of 67% (Precision is not shown in Table 2) and at a Sensitivity (recall) of 57%, using automated classification results (only relying on the class) without a manual revision would incur significant amount of missed (false negative; 43%) and wrong (false positives; approx. 1/3, calculated from Precision) articles. However, they do likely perform well enough for the envisioned use-cases where a lower performance is sufficient by focusing on the ranking (i.e., the rank/confidence that had to be reported for each result) of the relevant articles for a human user, such as a biologist or curators (highest AUC iP/R: 68%). In summary, current state-of-the-art systems are likely to have a significant impact on simplifying (but not completely automating) the manual process of article selection.

Only a small fraction of articles are PPI relevant when selecting a random collection from journals known to be annotation relevant in general. Although most participating teams did not adapt their systems to class imbalance they nonetheless could obtain sufficiently competitive results to make them useful as part of the annotation process. Compared to similar, previous tasks, it is possible to observe a tendency to go beyond simple bag-of-word approaches by integrating domain specific lexical resources, semantic labels and grammatical information to improve the document selection.

Methods

Result Structure

For each article, participants had to return a Boolean value (true/false) regarding its relevance for PPI curation (i.e., containing PPI with experimental evidence), a confidence score for this classification (in the range (0,1]), and the overall (unique) rank of the article in the whole set of articles with respect to its PPI relevance.

Evaluation

The evaluation is based on comparing automatically generated results with a manual annotated set of 6,000 PubMed records ("test set"), 900 of which were classified as "true". The same setup as for BioCreative II.5 was used: The overall set performances of the systems are evaluated using various measures, namely Accuracy, Sensitivity (Recall), Specificity, as well as Matthews' Correlation Coefficient (MCC score; the most stable of these evaluation function on unbalanced sets, as is the case for this task) for evaluating the pure classification performance (not taking into account rank or confidence). We also added the F-Scores for direct comparison to the BioCreative II results. The main utility measure of a system – i.e., the primary evaluation score for this task – is based on measuring a system's ability to provide the best possible ranked list of relevant abstracts, sorted from the most relevant (i.e., highest ranked article that is classified as true) to the most irrelevant article (i.e., highest ranked article classified as false). To this end, the area under the (interpolated) Precision/Recall curve is measured (AUC iP/R score) by using the results' ranking. Each team was allowed to submit five runs, and teams participating online (via the BCMS), could submit another five runs via their servers (i.e., independent of the number of "offline" runs), using the same setup as in BioCreative II.5. Additionally, we measured how much time it took to generate automatic predictions by the servers (online runs).

Authors' contributions

MK was the main organizer of the task, managed the teams, assembled the corpus and arranged for the Gold Standard annotations. MV contributed to corpus assembly, evaluation and the software development. FL wrote the software used for this task and ran the online challenge via the BCMS as well as the initial evaluation. AV, as a leading BioCreative organizer, was the responsible for this task.

Acknowledgements

We especially thank the BioGRID (Andrew Chatr-aryamontri) and MINT (Gianni Cesareni) databases and their curators for generating the data collections; without their enormous effort the ACT would not have been possible.

The work on BioCreative was supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and by grants BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovación and by the Spanish National Bioinformatics Institute (http://www.inab.org/).

References

1. Dowell K, Mcandrews-Hill M, Hill D, Drabkin H, Blake J: **Integrating text mining into the MGI biocuration workflow**. *Database* 2009, **2009**(0):bap019.
2. Krallinger M: **A Framework for BioCuration Workflows (part II)**. *Nature Precedings* 2009.
3. Cohen AM, Hersh WR, Peterson K, Yen P-Y: **Reducing workload in systematic review preparation using automated citation classification**. *J Am Med Inform Assoc* 2006, **13**(2):206-219.
4. Hersh W, Cohen A, Yang J, Bhupatiraju R: **TREC 2005 genomics track overview**. *Proceedings of the Fourteenth Text Retrieval Conference - TREC* 2005.
5. Cohen A: **An effective general purpose approach for automated biomedical document classification.** *AMIA Annu Symp Proc* 2006:161-165.
6. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II**. *Genome Biol* 2008, **9 Suppl 2**:S4.
7. Chatr-aryamontri A, Kerrien S, Khadake J, Orchard S, Ceol A, Licata L, Castagnoli L, Costa S, Derow C, Huntley R *et al*: **MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data**. *Genome Biol* 2008, **9 Suppl 2**:S5.
8. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An Overview of BioCreative II.5**. *IEEE/ACM Trans Comput Biol Bioinform* 2010, **7**(3):385-399.
9. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo C-J, Hsu C-N, Tsai RT-H, Hung H-C, Lau WW *et al*: **Introducing meta-services for biomedical information extraction**. *Genome Biol* 2008, **9 Suppl 2**:S6.

Overview of BioCreative III Gene Normalization

Zhiyong Lu[1], W. John Wilbur[1]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894 USA

Email addresses:

ZL: luzh@ncbi.nlm.nih.gov

WJW: wilbur@ncbi.nlm.nih.gov

Abstract

Background

The Gene Normalization (GN) task refers to the identification and linking of gene mentions in free text to standard gene database identifiers, an important task motivated by many real-world uses such as assisting literature curation for model organism databases. Here we report the GN challenge in BioCreative III where participating teams are asked to return a ranked list of gene ids of full-text articles. For training, we prepared 32 fully annotated articles and 500 partially annotated articles. A total of 507 articles were selected as the test set. We developed an EM algorithm approach for selecting 50 articles from the test set for obtaining gold-standard human annotations and used the same algorithm for inferring ground truth over the whole set of 507 articles based on team submissions. We report team performance by a newly proposed metric for measuring retrieval efficacy called Threshold Average Precision (TAP-k).

Results

We received a total of 37 runs from 14 different teams for the BioCreative III GN task. When evaluated using the gold-standard annotations of the 50 articles, the highest TAP-k scores are 0.3248 (k=5), 0.3469 (k=10), and 0.3466 (k=20), respectively. Higher TAP-k scores of 0.4581 (k=5, 10) and 0.4684 (k=20) are observed when evaluated using the inferred ground truth over the full test set.

Conclusions

Overall team results show that this year's GN task is more challenging than past events, which is likely due to the complexity of full text as well as species identification. By comparing team rankings with different evaluation data (gold

standard vs. inferred ground truth), we demonstrate that our approach succeeds in inferring ground truth adequate for effectively detecting good team performance.

Background

The gene normalization (GN) task in BioCreative III is similar to past GN tasks in BioCreative I and II (1-3) in that the goal is to link genes or gene products mentioned in the literature to standard database identifiers. This task has been inspired partly by a pressing need to assist model organism database (MOD) literature curation efforts, which typically involve identifying and normalizing genes being studied in an article. For instance, Mouse Genome Informatics (MGI) recently reported their search and evaluation of potential automatic tools for accelerating this gene finding process (4). Specifically, this year's GN task is to have participating systems return a list of gene database (Entrez Gene in this case) identifiers for a given article. There are two differences from past BioCreative GN challenges:

- Instead of using abstracts, full-length articles are provided.
- Instead of being species-specific, no species information is provided.

Both changes make this year's challenge event closer to the real literature curation task in MODs where humans are given full text articles without prior knowledge of organism information in the article.

Two additional new aspects of this year's GN task are the proposed evaluation metrics and the use of an EM algorithm for inferring ground truth based on team submissions. As many more genes are found in full text than in abstracts, returning genes by predicted confidence is preferred to a random order, as the former is more desirable in applications. Metrics used in past GN tasks such as Precision, Recall, and F-measure do not take ranking into consideration. Thus, we propose to use a new measure called

Threshold Average Precision (TAP-k), which is specifically designed for the measurement of retrieval efficacy in bioinformatics (13).

Finally, unlike in previous GN tasks where abstracts in the test set were completely hand annotated, the cost of manual curation on full text prevented us from obtaining human annotations for all 507 articles in the test set. Thus we resort to using team submissions for inferring ground truth. That is, given a labeling task and M independent labeling sources, it is possible to use these multiple sources to make estimates of the true labels which are generally more accurate than the labels from any single source alone. Perhaps the simplest approach to this is to use majority voting (5-7). On the other hand a number of methods have been developed using latent variables to represent in some way the quality of the labeling sources and based on the EM algorithm (8-12). There is evidence that such an approach can perform better than majority voting (8,11). We have chosen the most direct and transparent of the EM approaches (11) to apply to the GN task where we have multiple submissions as the multiple labeling sources. As far as we are aware this is the first attempt to base an evaluation of the performance of multiple computer algorithms on an EM algorithm for multiple independent data sources.

Methods

Data Preparation

For the purpose of obtaining full text articles in uniform formats and using them as a source for text analytics, all the articles selected for this task are published either by BioMed Central (BMC) or by Public Library of Science (PLoS), two PubMed Central (PMC) participating Open Access publishers. As a result, the text of each article was readily made available in both high-quality XML and PDF from PMC.

Participants were given a collection of training data to work with so that they could adjust their systems to optimal performance. The training set includes two sets of annotated full-length articles:

- 32 fully annotated articles by a group of invited professional MOD curators and by a group of bioinformaticians from the NCBI. Both groups were trained with detailed annotation guidelines (available as Appendix A) and a small number of example articles before producing gold-standard annotations. For each article in this set, a list of Entrez Gene ids is provided.
- A large number (500) of partially annotated articles. That is, not all genes that are mentioned in an article are annotated, but only the most important ones that within the scope of curation are annotated by human indexers at the National Library of Medicine (NLM). It is noted that most of the annotated genes are taken from the abstracts, though this is not 100%. This does not necessarily mean that the remainder of the text is useless. Presumably the full text can help to decide which genes are most important in the paper and determine the species to improve the prediction of the gene identifier.

For evaluating participating systems, we prepared a set of 507 articles as the test set. These articles were recently published and did not yet have any curated gene annotations. Due to the cost of manual curation, the same groups of curators were asked to produce human annotations only for a subset of 50 articles selected by the algorithm described below.

EM algorithm

In this scheme we assume there are M labeling sources and associate with the ith labeling source two numbers, the sensitivity as_i and the specificity bs_i. For the GN task we consider all the gene ids returned by the M sources as objects to be labeled.

Any given source produces a label for any such gene id which is the label “true” if the source returned that gene id or “false” if the source did not return that gene id. Then the sensitivity as_i is the probability that the ith source labels a correct gene id as true and the specificity bs_i is the probability that it labels an incorrect gene id as false. Assume there are N gene ids which require labeling. Then the model assumes a probability distribution $\left\{p_j\right\}_{j=1}^{N}$ where p_j is the probability that the jth gene id is correct. To begin the algorithm we initialize each p_j to be equal to the fraction of the M labels that are true for that gene id. The maximization step redefines the $\left\{as_i, bs_i\right\}_{i=1}^{M}$ in terms of the current $\left\{p_j\right\}_{j=1}^{N}$ by

$$\begin{aligned} as_i &= \left(1+\sum_{j=1}^{N}\delta_{ij}p_j\right)/\left(2+\sum_{j=1}^{N}p_j\right) \\ bs_i &= \left(1+\sum_{j=1}^{N}(1-\delta_{ij})(1-p_j)\right)/\left(2+\sum_{j=1}^{N}(1-p_j)\right) \end{aligned} \qquad (0.1)$$

where we have used typical Laplace smoothing and define δ_{ij} to be 1 if the ith source labels the jth gene id as true and 0 otherwise. The p_js are defined for the subsequent expectation step by

$$p_j = \frac{pr_j\prod_{i=1}^{M} as_i^{\delta_{ij}}\left(1-as_i\right)^{(1-\delta_{ij})}}{\left(pr_j\prod_{i=1}^{M} as_i^{\delta_{ij}}\left(1-as_i\right)^{(1-\delta_{ij})} + (1-pr_j)\prod_{i=1}^{M} bs_i^{(1-\delta_{ij})}\left(1-bs_i\right)^{\delta_{ij}}\right)} \qquad (0.2)$$

by Bayes’ theorem where for each j, pr_j is the prior for p_j. We initially took pr_j uniformly to be 0.5 and applied the algorithm to choose the 50 documents for hand labelling. Once we knew the correct annotations for the 50 document gold standard set we observed that only about 1% of gene ids returned by systems were correct. We subsequently have taken pr_j equal to 0.01 for all j in applying the algorithm to determine ground truth.

As mentioned above, our first use of this model was to find 50 documents among the 507 test documents which had the most variability in their labeling by different sources. For this purpose one submission from each team involved in the GN task was randomly selected and these submission were the 14 sources for application of the algorithm. When the algorithm was run to convergence we computed the entropy for the *j*th gene id by the formula

$$H_j = -p_j \log p_j - (1-p_j)\log(1-p_j) \qquad (0.3)$$

Each document was scored by the sum of the entropies for all the gene ids coming from that document. Thus a document score is a function of how many gene ids are reported for that document and how variably the gene ids are reported by the different sources. This sampling, running the model and scoring the documents, was repeated 100 times and the top 50 documents varied only a small amount from run to run. We chose the 50 documents with the highest average scores over the 100 trials for hand annotation to provide the *gold standard* evaluation.

The second use of the model was to apply it to the best submission from each team. The choice of the best submission itself is based on the gold standard, but we made no further use of the gold standard. From the converged model using these sources we obtained a set of probabilities $\left\{p_j\right\}_{j=1}^{N}$ and we accepted as correct all those gene ids for which $p_j \geq 0.5$ and considered all other gene ids to be incorrect. This labeling we refer to as the *silver standard*. We used it to evaluate all submissions on the whole set of 507 documents. A comparison of results as computed with the gold standard and the silver standard is given in Table 3.

Evaluation Metrics

We propose to use a new metric, Threshold Average Precision (TAP-k), for evaluating team performance. In short, TAP is Mean Average Precision (MAP) with a variable cutoff and terminal cutoff penalty. We refer interested readers to the original publication (13) and Appendix B for detailed description of the TAP-k metric. In our evaluation, we used three values of k: 5, 10 and 20.

Results

GN Annotation Data

As shown in Table 1, the average numbers (mean and median) of annotated genes per article in Set 1 are significantly lower than the ones in Set 2, while remaining relatively close to its counterparts in Set 3. This comparison suggests that the 50 selected articles are not representative of the articles in the training set. Instead, the entire test set seems akin to the training set in this respect.

Table 1: Statistics of annotated gene ids in the different data sets.

Set	Description	Min	Max	Mean	Median	St.dev.
1	Training Set (32 articles)	4	147	19	14	24
2	Test Set (50 articles – gold standard)	0	375	33	19	63
3	Test Set (507 articles – silver standard)	0	375	18	12	27

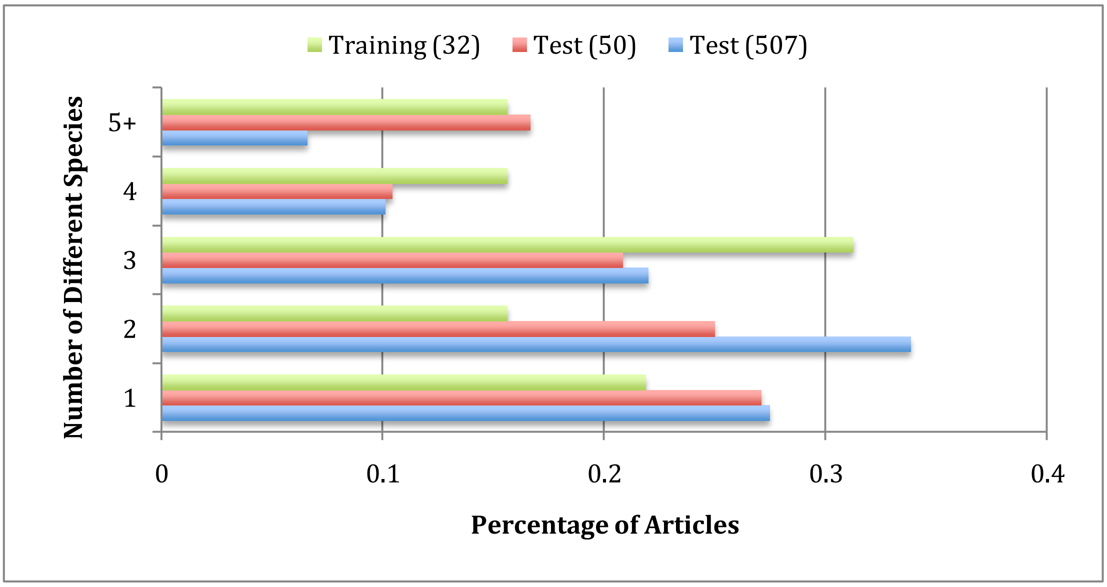
Table 2 shows that there are many different species involved in this year's GN task, which suggests that species identification and disambiguation may be critical in the process of finding the correct gene ids. We also show that the distributions of species among the genes in the three data sets look largely different. This indeed reflects the method of selecting the articles for training and evaluation: with some prior knowledge of a papers' species information, we were able to select the 32 articles as the training set to match the domain expertise of those invited professional MOD curators in order to obtain best possible human annotations. On the other hand, the articles in the test set were selected rather randomly as none was annotated prior to the evaluation.

Table 2: Statistics of species distribution in the different data sets.

#	Training Set (32 articles)	Test Set (50 articles)	Test Set (507 articles)
1	S. cereviaiae (27%)	Enterobacter sp. 638 (23%)	H. Sapiens (42%)
2	H. sapiens (20%)	M. musculus (14%)	M. musulus (24%)
3	M. musculus (12%)	H. Sapiens (11%)	D. melanogaster (6%)
4	D. melanogaster (10%)	S. pneumoniae TIGR4 (9%)	S. cerevisiae S228c (6%)
5	D. rerio (7%)	S. scrofa (5%)	Enterobacter sp. 638 (4%)
6	A. thaliana (5%)	M. oryzae 70-15 (4%)	R. norvegicus (4%)
7	C. elegans (3%)	D. melanogaster (4%)	A. thaliana (2%)
8	X. laevis (3%)	R. norvegicus (3%)	C. elegans (2%)
9	R. norvegicus (2%)	S. cerevisiae S228c(2%)	S. pneumoniae TIGR4 (2%)
10	G. gallus (2%)	E. histolytica HM-1 (2%)	S. scrofa (1%)
11+	Other 18 species (9%)	Other 65 species (23%)	Other 91 species (7%)

In addition to recognizing various species in free text, participating systems also needed to properly link them to the corresponding gene mentions in the articles. As shown in Figure 1 most articles (over 70%) in our data sets contain more than one species mention. In fact, it is not uncommon to see 5 or more species in an article. In cases where more than one species is found in an article, it can be challenging for systems to associate a gene mention with its correct species.

Figure 1: Percentage of articles annotated with different numbers of species in various data sets. Training (32) refers to the human annotations on the 32 articles in the training set. Test (50) and Test (507) refer to the gold standard and silver standard annotations on the 50 and 507 articles in the test set, respectively.



Team Results

Each team was allowed to submit up to 3 runs. Overall, we received a total of 37 runs from 14 teams. One team withdrew their late submission (one run) before the results were returned to the teams. Thus, per their request we do not report their system performance in the tables below. Nevertheless we included their withdrawn run when selecting 50 articles and computing the silver standard by our EM algorithm, as we believe more team submission data are preferable in this case.

We assessed each submitted run by comparing it to the gold and silver standard, respectively, and report their corresponding TAP scores (k = 5, 10, and 20) in Table 3. As highlighted in the table, the two runs from team 83 (T83_R1 and T83_R3) achieved highest TAP scores in almost all cases except when evaluated on the silver standard with k = 20 where the third run from Team 98 (T98_R3) was the best. However, we did not find a statistically significant difference between the results of the two teams (T83 and T98) when comparing their respective best runs (with different values of k) based on the Wilcoxon signed rank test.

Table 3: Team evaluation results on the 50 and 507 articles using gold and silver standard annotations, respectively. Results are sorted by team numbers.

Team_Runs	**Using gold standard (50 selected articles)**			**Using silver standard (All 507 articles)**		
	TAP (K=5)	**TAP K=10**	**TAP (K=20)**	**TAP (K = 5)**	**TAP (K = 10)**	**TAP (K = 20)**
T63_R1	0.0337	0.0484	0.0718	0.1567	0.1939	0.1954
T63_R2	0.0296	0.0454	0.0638	0.1368	0.1855	0.1942
T65_R1	0.0628	0.0958	0.1017	0.1487	0.1754	0.1938
T65_R2	0.0891	0.1073	0.1156	0.1533	0.1817	0.2024
T68_R1	0.1568	0.1817	0.1987	0.3398	0.3551	0.3516
T68_R2	0.1255	0.1431	0.1740	0.3257	0.3410	0.3375
T70_R1	0.0566	0.0566	0.0566	0.1146	0.1146	0.1146
T70_R2	0.0622	0.0622	0.0622	0.1243	0.1243	0.1243
T70_R3	0.0718	0.0718	0.0718	0.1512	0.1512	0.1512
T74_R1	0.2099	0.2447	0.2447	0.4518	0.4518	0.4518
T74_R2	0.2045	0.2417	0.2417	0.4514	0.4514	0.4514
T74_R3	0.2061	0.2432	0.2432	0.4555	0.4555	0.4555
T78_R1	0.0577	0.0726	0.1106	0.1245	0.1527	0.1877
T78_R2	0.0829	0.1161	0.1662	0.2495	0.2655	0.2655
T78_R3	0.0830	0.1091	0.1387	0.2219	0.2645	0.2762
T80_R1	0.1072	0.1556	0.1622	0.3983	0.3983	0.3983
T80_R2	0.0372	0.0507	0.0578	0.2165	0.2165	0.2165
T80_R3	0.0324	0.0432	0.0516	0.2224	0.2288	0.2288
T83_R1	0.3184	**0.3469**	**0.3466**	**0.4581**	**0.4581**	0.4581
T83_R2	0.3147	0.3366	0.3366	0.4293	0.4293	0.4293
T83_R3	**0.3228**	0.3445	0.3445	0.4303	0.4303	0.4303
T89_R1	0.1197	0.1197	0.1351	0.2681	0.2989	0.2989
T89_R2	0.1351	0.1521	0.1620	0.2624	0.2950	0.2950
T89_R3	0.1275	0.1522	0.1522	0.2873	0.2873	0.2873
T93_R1	0.1599	0.1842	0.2010	0.3916	0.3916	0.3916
T93_R2	0.1517	0.1804	0.2000	0.3602	0.3720	0.3720
T93_R3	0.1611	0.1856	0.2032	0.3946	0.3946	0.3946
T97_R1	0.0709	0.092	0.1001	0.1369	0.1620	0.1859
T97_R2	0.0630	0.0849	0.0945	0.1304	0.1563	0.1770
T97_R3	0.0709	0.092	0.1001	0.1369	0.1620	0.1859
T98_R1	0.2805	0.2971	0.3064	0.3720	0.3802	0.3779
T98_R2	0.2850	0.3033	0.3044	0.3682	0.3775	0.3767
T98_R3	0.2973	0.3125	0.3248	0.4086	0.4511	**0.4648**
T101_R1	0.1849	0.2235	0.2331	0.4128	0.4128	0.4128
T101_R2	0.1649	0.2102	0.2365	0.4097	0.4224	0.4224
T101_R3	0.1773	0.2096	0.2374	0.4351	0.4351	0.4351

To assess the quality of the silver standard, we show in Table 4 the results of team submissions against the silver standard on the 50 selected articles. Although the two best runs from Team 83 in Table 3 are still among the ones with the highest TAP scores, they no longer are the best runs. Instead, the top positions are replaced by T74_R3 (for k=5) and T98_R3 (for k=10 and 20), respectively.

Table 4: Team evaluation results on the 50 articles using the sliver standard annotations. Results are sorted by team numbers.

Team_Run	TAP (K=5)	TAP (K=10)	TAP (K=20)
T63_R1	0.0515	0.1045	0.142
T63_R2	0.0455	0.0978	0.1335
T65_R1	0.0996	0.1259	0.1473
T65_R2	0.109	0.1317	0.1522
T68_R1	0.2238	0.2719	0.3152
T68_R2	0.2098	0.2917	0.2917
T70_R1	0.053	0.053	0.053
T70_R2	0.0566	0.0566	0.0566
T70_R3	0.096	0.096	0.096
T74_R1	0.3677	0.3677	0.3677
T74_R2	0.3713	0.3713	0.3713
T74_R3	**0.3747**	0.3747	0.3747
T78_R1	0.0589	0.0793	0.1139
T78_R2	0.1048	0.1548	0.2114
T78_R3	0.0972	0.1394	0.1949
T80_R1	0.2464	0.2719	0.2719
T80_R2	0.0663	0.1107	0.1177
T80_R3	0.0749	0.1231	0.1291
T83_R1	0.3498	0.3531	0.3531
T83_R2	0.3222	0.3222	0.3222
T83_R3	0.3313	0.3313	0.3313
T89_R1	0.1714	0.217	0.217
T89_R2	0.2141	0.2581	0.2949
T89_R3	0.2054	0.2054	0.2054
T93_R1	0.2518	0.2979	0.2979
T93_R2	0.2011	0.2514	0.2854
T93_R3	0.2487	0.293	0.293
T97_R1	0.1066	0.1307	0.149
T97_R2	0.09	0.1126	0.1323
T97_R3	0.1066	0.1307	0.149
T98_R1	0.3218	0.3388	0.3494
T98_R2	0.3217	0.3391	0.3496
T98_R3	0.3576	**0.3953**	**0.4499**
T101_R1	0.3504	0.374	0.374
T101_R2	0.3068	0.379	0.3976
T101_R3	0.3077	0.3942	0.3942

Discussion

Team Results and Quality of Silver Standard

Although we are unable to directly compare this year's GN team results against the ones in previous GN challenges due to different evaluation metrics, results in Table 3 led us to believe that this year's GN task is more challenging, potentially due to the complexity of full text processing and species identification (14,15).

Using the silver standard allowed us to assess team submissions on the entire set of test articles without having human annotations for all articles. As can be seen from results in Tables 3 and 4, TAP scores are consistently higher when evaluated on the silver standard compared to the gold standard. Furthermore, individual team rankings may be affected. For instance, as mentioned earlier the best performing run was T83_R3 using gold standard but T74_R3 using silver standard. Nevertheless, it is evident that relative rankings tend to be largely preserved in this comparison. For instance, teams 83, 74, 98 and 101 consistently remain as the top tier group in all evaluations. This provides some justification for the silver standard and suggests that this approach to evaluation has some merit.

As just noted, TAP scores in Table 3 show that overall team performance is lower on the 50 articles than on the entire set of 507 articles. The reasons for this are two fold. First, the 50 articles are the most difficult ones for gene normalization (as shown by comparing the silver results for the 50 and the 507) and this supports our rationale for their choice. Second, by comparing the gold and silver results for the 50 in Tables 3 and 4, we can see that team results are always higher when evaluated using the silver standard. Taken together, this suggests that the true TAP scores on the entire test set should be slightly lower than what is currently reported using the silver standard in Table 3.

Team Methods

Each team was required to submit a system description before receiving the gold standard annotations on the 50 articles and their scores. Based on reading those submitted descriptions, we found the general framework for the gene normalization task comprises the following major steps:

1) Identifying gene mentions
2) Identifying species information and linking such information to gene mentions
3) Retrieving a list of candidate gene ids for a given gene mention
4) Selecting gene ids through disambiguation.

Conclusions

We have successfully organized a community-wide challenge event for the gene normalization task. There were a total of 37 submissions by 14 different teams from Asia, Europe, and North America. The highest TAP-k scores obtained on the gold-standard annotations of the 50 test articles are 0.3248 (k=5), 0.3469 (k=10), and 0.3466 (k=20), respectively. In addition, TAP-k scores of 0.4581 (k=5, 10) and 0.4684 (k=20) are observed when using the silver standard of the 507 test articles. In comparison with past BioCreative GN tasks, this year's task bears more resemblance to real-world tasks in which curators are given full text without knowing species information. As a consequence, this year's task has proved more difficult than the ones in the past, which is evident from the overall lower team performance.

Finally, we believe the TAP-k metric and EM algorithm proved to be adequate for evaluating retrieval efficacy and for inferring ground truth based on team submissions. In particular, the proposed pooling method allowed us to effectively detect good team performance without having to relying on human annotations. Future work should include conducting a more detailed analysis of various techniques and tools used by different participating teams, as this may provide valuable direction for future research on the GN problem. Also, we plan to combine results from different teams as an ensemble system to test maximal aggregate performance, as in various previous studies (1,16,17). Finally, we would like to investigate how systems developed for the GN task may be used in real-world applications.

Additional material

Additional file 1: GN annotation guidelines

Additional file 2: Introduction to TAP-k

Acknowledgements

This research is supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank all the annotators who produced the gold-standard annotations.

References

1. Morgan, A.A., Lu, Z., Wang, X., *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol*, 9 Suppl 2, S3.
2. Hirschman, L., Colosimo, M., Morgan, A., Yeh, A. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1, S11.
3. Colosimo, M.E., Morgan, A.A., Yeh, A.S., Colombe, J.B., Hirschman, L. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, 6 Suppl 1, S12.
4. Dowell, K.G., McAndrews-Hill, M.S., Hill, D.P., Drabkin, H.J., Blake, J.A. (2009) Integrating text mining into the MGI biocuration workflow. *Database (Oxford)*, 2009, bap019.
5. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y. (2008) Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii.
6. Sheng, V.S., Provost, F., Ipeirotis, P.G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Las Vegas, Nevada, USA.
7. Donmez, P., Carbonell, J.G., Schneider, J. (2009) Efficiently learning the accuracy of labeling sources for selective sampling. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Paris, France.
8. Whitechill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J. (2009) Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 2035-3043.
9. Welinder, P., Perona, P. (2010) Online crowdsourcing: rating annotators and obtaining cost-effective labels. *Workshop on Advancing Computer Vision with Humans in the Loop at CVPR'10*.

10. Smyth, P., Fayyad, U., Burl, M., Perona, P., Baldi, P. (1995) Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, 7.

11. Raykar, V.C., Yu, S., Zhao, L.H., *et al.* (2010) Learning From Crowds. *Journal of Machine Learning Research*, 11, 1297-1322.

12. Dawid, A.P., Skene, A.M. (1979) Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 20-28.

13. Carroll, H.D., Kann, M.G., Sheetlin, S.L., Spouge, J.L. (2010) Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics. *Bioinformatics*, 26, 1708-1713.

14. Kappeler, T., Kaljurand, K., Rinaldi, F. (2009) TX task: automatic detection of focus organisms in biomedical publications. *Proceedings of the Workshop on BioNLP*. Association for Computational Linguistics, Boulder, Colorado.

15. Wang, X., Tsujii, J., Ananiadou, S. (2010) Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26, 661-7.

16. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A., Valencia, A. (2010) An Overview of BioCreative II.5. *IEEE/ACM Trans Comput Biol Bioinform*, 7, 385-399.

17. Smith, L., Tanabe, L.K., Ando, R.J., *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol*, 9 Suppl 2, S2.

BioCreative GN Task 2010

Gene/Protein Annotation Guidelines

What to annotate and normalize:

1. Find gene/protein mentions in the full-length article including figure and table legends and map them to unique Entrez Gene identifiers (http://www.ncbi.nlm.nih.gov/gene/).
2. Entrez Gene Ids are required. (UniProt Ids or Model Organism Database Ids are optional).
3. Annotate all genes mentioned in the article including those genes mentioned in passing or only mentioned once in the article. However, there is no need to rank or group genes for this assignment.
4. When there is no explicit mention of a gene's organism of origin in surrounding text, try to use the article context to help determine its species. Annotate the gene only when the species information can be determined. Some helpful clues for determining species include details in the methods/materials section such as cell lines, organism-specific gene nomenclature conventions, etc.
5. You may also use your domain knowledge for determining which organism a gene belongs to when no explicit species information is given in the text. If there is absolutely no clue about the species, or in situations where the species information is ambiguous (e.g. the authors use one gene as a representative of its homologs), do not annotate the gene.
6. When cell lines from different species are used to study a gene, determine and use the gene's *species of origin* instead of a cell lines' *species of origin* for annotation.

What NOT to annotate:

1. Do not annotate references sections. But this section may be useful for species identification. However, do not go beyond reading reference titles. That is, don’t read the referenced articles.
2. Do not use or annotate supplementary material or supporting information.
3. Annotate target proteins but do not annotate antibodies/reagents that are used to study target proteins.
4. Do not annotate the Methods/Materials section for genes/proteins. But this section may be useful for species identification. (Our reasoning is that the Methods/Materials section often contains information about reagents or antibodies that are themselves proteins but are not *curatable* objects; if *curatable* genes/proteins are mentioned in such a section, then they will almost certainly be mentioned elsewhere in the article).
5. Do not annotate genes where no unique ids can be identified in Entrez Gene. For example, if you find a gene mention “x-tsk” in a paper and subsequently search it in Entrez Gene, you may be presented with two separate Entrez gene records (x-tsk-b1 & x-tsk-b2). In this case, if you can’t tell which specific gene is used in the paper based on your domain knowledge, do not annotate this gene.
6. Do not annotate a protein complex (e.g. TFTC complex). But if its members are explicitly given (NFKB-IKB complex) they should be annotated.
7. Do not annotate a protein family (e.g. cytokines; ring-h2 finger proteins) because no unique Entrez Gene id can be assigned to it.
8. Do not annotate a gene/protein with only non-species taxonomic information (e.g. mammalian p53) for the same reason above.

What is *TAP-k*?

Here we refer to the measure defined by Carroll, H. D., Kann, M. G., Sheetlin, S. L., and Spouge, J. L., Threshold Average Precision (TAP-k): A Measure of Retrieval Designed for Bioinformatics, *Bioinformatics Advanced Access published on May 26, 2010*.

The Threshold Average Precision (*TAP-k*) is *MAP* with a variable cutoff and terminal cutoff penalty.

For a single query the average precision (*AP*) is computed by summing the precision at each rank that contains a true positive item and then dividing this sum by the number of positives for that query. If the retrieval system assigns to each retrieved item a score and the retrieved items are ranked in decreasing order of score, then it may be useful to cut off the retrieval at some fixed score level x. We can compute the average precision with cutoff x (APC_x). This is the sum of the precision at each rank with a true positive item and a score >=x, divided by the total number of positives for the query. Finally, suppose that $y>x$ and further suppose there are no true positive items in the sum for APC_x that are below y. Then $APC_y=APC_x$. But clearly it would make more sense to choose the cutoff y than the cutoff x. To distinguish between these two cases we define the average precision with cutoff x and terminal penalty ($APCP_x$). Let P_x be the precision at the last rank with score >= x and let P be the total number of positives. Then define

$$APCP_x = \frac{TP * APC_x + 1 * P_x}{TP+1}. \tag{1.1}$$

$APCP_x$ is just the weighted average of APC_x and P_x with most of the weight applied to APC_x, but P_x supplying the terminal penalty. In our hypothetical case P_y will be greater than P_x so that $APCP_y$ is also greater than $APCP_x$ and the score rewards the better choice of cutoff or equally penalizes the poorer choice. Whereas *MAP* is the average of *AP* over all the queries, *TAP-k* is the average of $APCP_x$ over all the queries where x is chosen as the largest score that produces a median of k false positive retrievals over all the queries. The median is used here instead of the mean because it is more robust against noise and outliers.

There are some practical considerations when applying *TAP-k*. First, retrieval systems must produce scores commensurate with their rankings and these scores must be interpretable across different queries. Since most systems generate their retrieval by scoring this should not make the task any more difficult than usual. On the other hand some kind of score normalization may be necessary for some systems, depending on how the scores are constructed. An ideal score would be a probability estimate that the retrieved item is a true positive, but a score need not be a probability estimate for good performance. The score that is reported simply has to have the same implications for relevance of the item regardless of the query, for the best performance. Another important issue is the length of the retrieved lists returned by a system. If many of the retrieved lists are too short to have k false positives appear, then no cutoff score may produce a median number of *k* false positive retrievals for the set of queries. In that case we will take the cutoff score *x* to be the lowest score over all the retrieval lists for all the queries.

Example 1. Data for five queries, Q1-Q5 are presented in the table. The numbers in parentheses following the query numbers are the number of correct or relevant items for each query. This data was generated randomly based on the scores. Each score is the probability that the corresponding retrieved item would be relevant (relevance is shown by a 1 in the rel column for each query). The scores themselves are parts of power series which are convenient for generating realistic scores. Retrieval is cut off at 15 items for each query to keep the data easily manageable and as a consequence not all relevant items are necessarily retrieved.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.900	0	0.500	0	0.500	0	0.2	1	0.980
2	1	0.738	0	0.475	1	0.475	0	0.187	0	0.788
3	0	0.605	1	0.451	0	0.451	0	0.174	0	0.633
4	1	0.496	0	0.429	0	0.429	0	0.163	1	0.509
5	1	0.407	1	0.407	0	0.407	0	0.152	1	0.409
6	0	0.334	0	0.387	0	0.387	0	0.142	0	0.329
7	0	0.274	0	0.367	0	0.367	0	0.132	0	0.265
8	0	0.224	0	0.349	1	0.349	0	0.123	0	0.213
9	1	0.184	0	0.332	0	0.332	0	0.115	0	0.171
10	0	0.151	1	0.315	1	0.315	0	0.107	1	0.138
11	0	0.124	0	0.299	0	0.299	0	0.100	0	0.111
12	0	0.101	0	0.284	0	0.284	0	0.094	0	0.089
13	0	0.083	0	0.270	0	0.270	0	0.087	0	0.071
14	0	0.068	0	0.257	0	0.257	0	0.082	0	0.057
15	0	0.056	0	0.244	1	0.244	0	0.076	0	0.046

Here the score cutoff for *TAP*-5 is 0.213 and the values of *APCP*$_5$ are 0.675, 0.206, 0.264, 0, 0.413 and the average of these numbers, *TAP*-5, is 0.312. The blue background shows what parts of the retrieval were included in the scoring (likewise for subsequent examples).

Example 2. Example 1 output, but the system has limited its retrieval to the top 4 ranks for each query.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.900	0	0.500	0	0.500	0	0.2	1	0.980
2	1	0.738	0	0.475	1	0.475	0	0.187	0	0.788
3	0	0.605	1	0.451	0	0.451	0	0.174	0	0.633
4	1	0.496	0	0.429	0	0.429	0	0.163	1	0.509
5										
6										
7										
8										
9										
10										

11										
12										
13										
14										
15										

Here the cutoff score is 0.163 (the lowest score possible) and the $APCP_5$ values are 0.583, 0.097, 0.125, 0, 0.333 and the average, *TAP*-5, of these numbers is 0.228. Here the *TAP*-5 is lower than for example 1 because the system cut the retrieval off prematurely and this decreased the recall and thus the *TAP* -5 score.

Example 3. Example 1 output again, but scores changed so they only reflect the rank and not the quality of the retrieved material.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.9	0	0.9	0	0.9	0	0.9	1	0.9
2	1	0.85	0	0.85	1	0.85	0	0.85	0	0.85
3	0	0.8	1	0.8	0	0.8	0	0.8	0	0.8
4	1	0.75	0	0.75	0	0.75	0	0.75	1	0.75
5	1	0.7	1	0.7	0	0.7	0	0.7	1	0.7
6	0	0.65	0	0.65	0	0.65	0	0.65	0	0.65
7	0	0.6	0	0.6	0	0.6	0	0.6	0	0.6
8	0	0.55	0	0.55	1	0.55	0	0.55	0	0.55
9	1	0.5	0	0.5	0	0.5	0	0.5	0	0.5
10	0	0.45	1	0.45	1	0.45	0	0.45	1	0.45
11	0	0.4	0	0.4	0	0.4	0	0.4	0	0.4
12	0	0.35	0	0.35	0	0.35	0	0.35	0	0.35
13	0	0.3	0	0.3	0	0.3	0	0.3	0	0.3
14	0	0.25	0	0.25	0	0.25	0	0.25	0	0.25
15	0	0.2	0	0.2	1	0.2	0	0.2	0	0.2

Here the scores no longer reflect quality and thus they do not give an accurate idea of where to cut off retrieval to obtain maximal efficiency. As a result there is a drop in *TAP*-5 as compared with example 1. The cutoff score is 0.6 and the $APCP_5$ values are 0.687, 0.170, 0.107, 0, 0.421 and the average, *TAP*-5, is 0.277.

Systems Descriptions

Machine learning-based approaches for BioCreative III tasks

Shashank Agarwal*[1], Feifan Liu[2] , Zuofeng Li[2] and Hong Yu[1,2,3]

[1]Medical Informatics, College of Engineering and Applied Sciences, University of Wisconsin-Milwaukee,Milwaukee, WI, USA
[2]Department of Health Sciences, College of Health Science, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[3]Department of Computer Science and Electrical Engineering, College of Engineering and Applied Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Email: Shashank Agarwal*- agarwal@uwm.edu; Feifan Liu - liuf@uwm.edu; Zuofeng Li - lizuofeng@gmail.com; Hong Yu - hongyu@uwm.edu;

*Corresponding author

Abstract

Background: The BioCreative III challenge was conducted to evaluate text mining and information retrieval applications on three tasks: gene normalization (GN), interactive task (IAT), and protein-protein interaction (PPI), which comprised two sub-tasks - article classification task (ACT) and interaction methods task (IMT). We participated in all three tasks.

Methods: We developed machine learning-based approaches to explore diverse textual and non-textual features for each task. We also explored feature selection strategies to improve performance. We trained our systems on the training and development data provided by the organizers of BioCreative III and evaluated it on the test data provided by the organizers.

Results: For the GN Task, evaluation was conducted on the 50 most difficult articles from the test data of 507 articles. Our system obtained TAP-5, TAP-10 and TAP-20 scores of 0.14, 0.15 and 0.16, respectively. For ACT and IMT, our systems obtained 88% accuracy and 54% F1-score, respectively.

Gene Normalization Task (GN)

We developed a three-tiered GeneNorm system. GeneNorm first identifies gene mentions in articles and then searches for candidate Entrez Gene entries. Then, GeneNorm applies a disambiguation module that was built upon supervised machine learning. To that end, we explored diverse learning features.

Tier 1: Identifying Gene Mentions

In this tier, our strategy was to maximize recall for gene mention identification. We identified gene mentions using two methods. In the first method, we modified an existing Conditional Random Fields (CRF) based gene named entity recognizer, BANNER, by adding dictionary lookup as a binary feature. The gene symbols and synonyms in the Entrez Gene database were used as the dictionary of possible genes. For each token, we first calculated its frequency inside and outside gene names using the BioCreative II gene mention training data. If the frequency of the token inside gene names was greater than or equal to the frequency outside gene names and the token appeared in the gene dictionary as well, then the lookup feature was assigned as true, otherwise it was assigned as false. We trained a model for gene named entity recognition using the modified version of BANNER on BioCreative II's gene mention training data. On evaluating the performance of this model on BioCreative II's gene mention test data, we obtained a precision, recall and f-score of 83.9%, 86.3% and 85.1% respectively. The f-score of the modified system is similar to the original BANNER system, but the recall is improved (86.3%) compared with the original system (83.1%).

The second method to identify gene mentions was based on italics markup tags in the document. The data provided for BioCreative III was in XML format, with italics markup in text available for some articles. Italics markups are usually used for gene/protein names and species name. Since our goal was to increase recall, we added the terms marked in italics as gene mentions.

Tier 2: Identifying candidate genes from identified gene mentions

We built an index of all gene symbols and names in the Entrez Gene database and linked them to the corresponding gene ids using Apache Lucene. We also added synonyms from the corresponding SwissProt entries to the index. Each identified gene mention in Step 1 was then expanded by a rule-based gene name variation generator and the expansions were used to query the Entrez Gene index. Top 100 genes returned as a result to each query were considered as the candidate genes in the article. We ran steps 1 and 2 on the training articles for BioCreative III Gene Normalization task. We were provided with a total of 525 articles; 500 of these articles were annotated for important genes only, whereas 32 articles were annotated for all genes. We obtained a precision, recall and F1-score of 0.02%, 92.08% and 0.05% respectively for important genes and 0.26%, 87.15% and 0.51% for all genes. The poor precision indicated that there were a lot of false positive candidate genes at this stage.

Tier 3: Learning framework for disambiguation

We explored and evaluated several learning algorithms for further disambiguation among those candidate genes. To do that, we identified 26 features for each candidate gene as described in Table 1.

Table 1 - Features used for Gene Normalization Task

Feature name	**Feature type**	**Description**
Sequence	Binary	If the article has a genetic sequence, checks if the genetic sequence belongs to this gene
BC II.5	Binary	If the gene was identified by the GN system developed for BioCreative II.5
GO Score (4 features)	Continuous	Similarity* between this gene's Gene Ontology (GO) annotations' concatenated names and descriptions, and (1) all the text of the article, (2) title of the article, (3) abstract of the article and (4) gene mention containing sentences in the article
GeneRIF Score (4 features)	Continuous	Similarity* between this gene's GeneRIF annotations, and (1) all the text of the article, (2) title of the article, (3) abstract of the article and (4) gene mention containing sentences in the article
Lookup (5 features)	Binary	Look up the presence of (1) gene's species in article (uses LINNAEUS and regular expression), (2) gene's map location term in article, (3, 4, 5) gene's mention in (3) title, (4) abstract and (5) figure text extracted using an optical character recognition software
Count (7 features)	Integer	Counts the number of (1) gene's GO annotations, (2) gene's GeneRIF annotations, (3) gene's interacting genes, (4) gene's interacting genes mentioned in article, (5) times gene was mentioned, and (6, 7) gene's species' annotation in (6) BioGRID and (7) GeneRIF
String Similarity (4 features)	Continuous	Calculates the string similarity between the gene's mention in article and gene's official symbol and any synonym using edit distance and Jaro-Winkler measure

* Similarity between texts was calculated using LingPipe

The organizers of BioCreative III released gold annotation for 50 most difficult articles from the 507 articles test data. On these 50 articles, our system obtained a TAP-5, TAP-10 and TAP-20 scores of 0.14, 0.15 and 0.16, respectively.

Interactive Task (IAT)

We developed a demonstration system-GeneIR, that performs both gene indexing and gene oriented document retrieval for the IAT in BioCreative III.

Gene Indexing

We first identified and normalized all gene mentions in a given article provided by the user. We used the gene normalization system we developed for the GN task, GeneNorm, to return all gene IDs (Entrez Gene ID) mentioned in an article. We extracted the frequency with which the gene was mentioned in the article, and checked if the gene was mentioned in the title or the abstract of the article.

To score the centrality of each gene, we trained a machine learning classifier. To train this classifier, we used the important genes in the GN training data as positive instances, and the remaining genes as negative instances. As features for the classifier, we determined if the gene appears in the title or abstract of the article, the number of times it appears in the article, the number of Gene Ontology (GO) and GeneRIF annotations associated with the gene, and the GeneRIF species popularity of the gene (see Table 1 for species popularity).

Retrieving

For the retrieval sub-task, we indexed all articles in the data source. We indexed the title, abstract, full-text, figure legend and references' text separately. If the user enters a gene name, it is treated as a query to search the article index. To account for gene name variations (for example, BRCA1 vs BRCA-1), a gene name variation generator was implemented to expand the gene name query. If the user enters a gene id, the system obtains the gene's symbol, synonyms and their variations as query to retrieve relevant documents.

User Interface

A user interface for our system is available at http://autumn.ims.uwm.edu:8080/biocreative3iat/. We provide two search boxes, one to obtain articles based on gene name or gene's Entrez ID, the other to obtain all genes normalized for an article of a given PMC ID. From the gene results or article results, one can view other genes in an article or other articles containing a gene, respectively. When viewing the gene normalizations for an article, the genes can be sorted by centrality (default), presence in title and abstract, or the frequency with which they appear in the article. Users can view all genes or an individual gene highlighted in the article. Also, genes can be added or deleted for a given article.

Protein-Protein Interaction Task (PPI)

Article Classification Task (ACT)

For ACT we trained supervised machine learning algorithms Support Vector Machines (SVMs) and multinomial Naive Bayes (NB). All text was normalized by lowercasing, removing punctuations, stemming words and removing numbers. We used unigrams (individual words) and bigrams (two consecutive words) as features for the machine learning classifiers. We sorted features by their mutual information score and trained the classifiers by using either the top 400 or the top 1000 features.

We were provided with a 2280 articles training data and a 4000 articles development data by the BioCreative III organizers. Positive and negative instances were evenly distributed in the training data whereas in the development data, there were 682 positive instances and 3318 negative instances. The distribution of positive and negative instances in the development data was similar to the distribution in the test data. For our submission, we trained the classifier on development data only or a combination of training and development data. Our hypothesis was that training on development data would allow the classifier to learn the distribution of instances in the test data, whereas adding the training data would provide more instances for learning, albeit at cost of a slightly skewed distribution.

On evaluating our ACT system on the test data, we found that the best accuracy of 88% was obtained by a SVM classifier trainined on development data only using the top 400 unigrams and bigrams features. The best AUC value of 62% was attained by a NB classifier trained on the combination of training and development data using the top 400 unigram and bigram features.

Interaction Methods Task (IMT)

The IMT involved mapping nodes in PSI-MI ontology to articles. For each ontology node, we obtained the concept name and its synonyms. We manually added synonyms for some ontology nodes, such as “anti bait immunoprecipitation” for “anti bait coimmunoprecipitation” and “radioligand binding” for “saturation binding”. A keyword for each ontology node was manually extracted by the first author, for example, “coimmunoprecipitation” for “anti bait coimmunoprecipitation”. We extracted unigrams and bigrams from each node’s concept name and synonyms. For each unigram and bigram, we calculated the mutual information score and chi-square value using the training data.

We approached IMT as a classification problem, where we try to determine if an article-ontology node pair is positive or negative. We identified 21 features (as listed in Table 2) and scored those features for each article-ontology node pair. We then trained machine learning classifiers Random Forest, Random Committee, Naive Bayes Tree and J48 to predict the label for each article-ontology node pair.

Evaluation on the test data indicated the best F1-score of 54% was attained by a Random Forest classifier.

Table 2 - Features used for IMT

Feature	**Feature type**	**Description**
Perfect match (2 features)	Binary	For each node, checks if (1) the concept name or (2) any synonym name appears in the article
Term match (4 features)	Binary	For each node, checks if any unigram/bigram in the node's (1, 2) concept name or (3, 4) synonyms appears in the article
Term match ratio (4 features)	Continuous	For each node, the ratio unigram/bigram in the node's (1, 2) concept name or (3, 4) synonyms that appears in the article
Matched terms mutual information sum (4 features)	Continuous	Sum of mutual information score of each matching unigram/bigram in the node's (1, 2) concept name or (3, 4) any synonym.
Matched term chi-squared sum (4 features)	Continuous	Sum of chi-squared value of each matching unigram/bigram in the node's (1, 2) concept name or (3, 4) any synonym.
Node popularity	Integer	The number of times this node is annotated in the training data
Regex annotation	Binary	Checks if the regular expression-based annotator that was provided by the organizers of BioCreative III annotates the current article-ontology node pair
Keyword presence	Binary	Checks if the keyword for the ontology node appears in the article

Authors contributions

FL conducted machine learning for GN. ZL implemented the sequence based normalizer. SA conducted PPI, IAT and feature identification and scoring for GN. HY and FL provided guidance.

Acknowledgements

We acknowledge the support from the National Library of Medicine, grant numbers 5R01LM009836 to Hong Yu and 5R01LM010125 to Isaac Kohane. We thank DaeHyun Kim and Balaji Ramesh for developing the figure text extraction system.

Online Gene Indexing and Retrieval for BioCreative III at the University of Iowa

Sanmitra Bhattacharya*[1], Aditya K Sehgal[2] and Padmini Srinivasan[1,3]

[1]Department of Computer Science, The University of Iowa, 14 MacLean Hall, Iowa City, Iowa 52242, USA
[2]Parity Computing Inc., 6160 Lusk Blvd. Suite C205, San Diego, CA 92121, USA
[3]Department of Management Sciences, The University of Iowa, S210 PBB, Iowa City, Iowa 52242, USA

Email: Sanmitra Bhattacharya*- sanmitra-bhattacharya@uiowa.edu; Aditya K Sehgal - a.sehgal@paritycomputing.com; Padmini Srinivasan - padmini-srinivasan@uiowa.edu;

*Corresponding author

Abstract

Background: The interactive demonstration task (IAT) for the BioCreative III Challenge focused on gene indexing and retrieval using full text articles. The goal of the indexing subtask was to identify the unique identifiers for gene mentions in selected PMC articles while the goal of the retrieval subtask was to identify relevant PMC documents for a selected gene.

Results: The IAT task of gene indexing and retrieval using full text articles was based on our BioCreative III gene normalization system. For the indexing subtask, several features that would assist a curation workflow were implemented. This includes display of full text highlighted with selected gene and species names, gene names normalized to single Entrez Gene identifiers, links from identifiers to standard databases and a ranking of genes based on the frequency of occurrence of a gene mention in a document. For the retrieval subtask a ranked list of PMCIDs based on the frequency of a selected gene mention in an article is returned.

Conclusions: Evaluations for this task was based on the feedback provided by the BioCreative User Advisory Group (UAG). The system for the interactive demonstration task for gene indexing and retrieval is available online at: http://siena.cs.uiowa.edu/~sbhttcha/.

Background

Previous BioCreative challenges [1,2] included tasks like gene normalization which represents a simplification of the real curation task. In the real process, the curator generally works from the full text of the articles, and identifies only particular kinds of genes of interest (for example, only genes for a specific organisms or only genes that have experimental evidence in the article). The gene indexing and retrieval of full-text articles in BioCreative III challenge [3] was closer to real curation pipeline. Our IAT system, which was based on our BioCreative III gene normalization system does cross-species gene indexing and retrieval. Besides the ambiguity in gene names, cross-species indexing and retrieval task is a challenging task as different species follow different naming conventions and are used in varied ways across the literature. Hence accurate identification and normalization of species names to unique taxonomy identifiers is pertinent to the cross-species gene normalization. Species names also have inherent ambiguities when used in abbreviated forms (as in C. elegans which refers to 41 different species in the NCBI Taxonomy) and are often referred to indirectly (as in patient or women for human).

Results and Discussion

The IAT task had two subtasks, namely, indexing and retrieval. The home page of our online system gives users a choice between these two tasks. For the Indexing task a user can either enter a PMCID (from the BioCreative training set) or select from a list of PMCIDs. For the Retrieval task an user can either enter a gene name or select from a list of gene names which are displayed as an alphabetically sorted list.

Indexing

For the indexing subtask we designed an interface where the full-text of an user-selected article is displayed in the left frame of the web page. In the right frame the gene names, species names, normalized Taxonomy IDs [4], normalized Entrez Gene IDs [5] and frequency count of the gene names corresponding to the article are displayed. The frequency count is based on the count of the gene names as identified by the gene name taggers. The results are initially sorted based on the gene mention frequencies. However, the user can sort the results on individual fields. Gene and species names are highlighted in yellow by selecting individual gene and species names from the right frame. The species identifiers and normalized gene identifiers are linked to the corresponding records in the NCBI Taxonomy database and Entrez Gene, respectively. Figure 1 shows a screen-shot of the indexing system.

Retrieval

For the retrieval subtask we designed an interface which displays a list of relevant PMCIDs and frequency per article for a selected gene name. The PMCIDs and frequencies can be sorted in ascending or descending orders. Clicking on a PMCID displays the full text of an article. Figure 2 shows a screenshot of the retrieval system.

Conclusions

In this paper we have presented an interactive demonstration task for gene indexing and retrieval. The system is based on our BioCreative III gene normalization system. Improving the gene normalization strategy is inherent to improving the IAT system performance. In addition to this, curation process can be facilitated by the integration of certain features to the system like specifying article sections corresponding to a gene/protein name since curators are mostly interested in gene mentions appearing in certain sections of an article where experimental evidence is available.

Methods

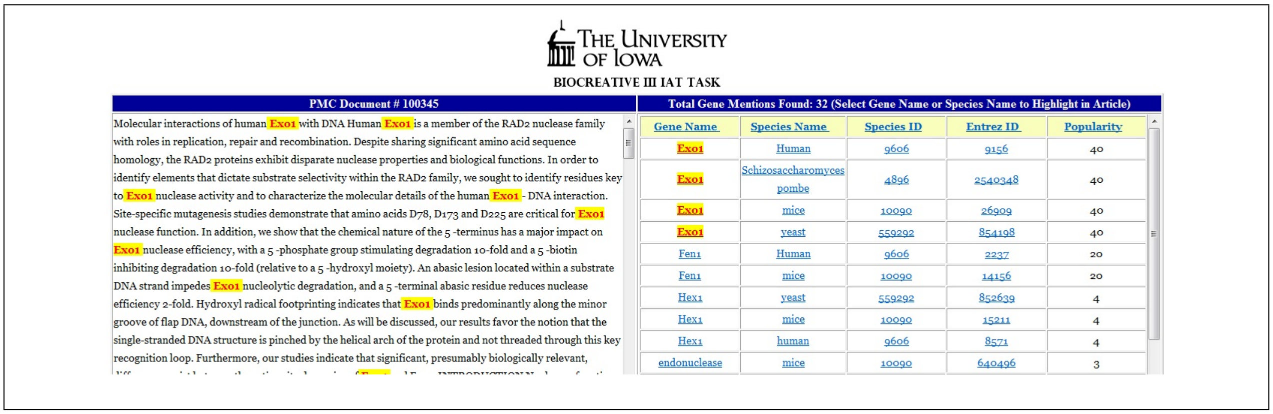
The IAT task consisted of a single large dataset of over 17,000 PMC articles. This served as both the training and testing set. These file were provided in XML format. The following steps follow directly from our BioCreative III gene normalization system. The XML files were stripped of XML tags and XML codes for Greek symbols were replaced with corresponding Greek names. Since no species information was provided we used LINNAEUS [6], a species name identification system for biomedical articles. Modifications were made to the LINNAEUS species dictionary to include the first names of model organisms (which are often referred to in the literature on a first name basis) for the respective taxon identifiers of those organism. For example, *Arabidopsis* (for model organism *Arabidopsis thaliana*) was added to species identifier 3702, etc. The gene/protein names were identified using ABNER [7] (trained on NLPBA corpus) and LingPipe [8] (trained on GENIA corpus) gene name taggers. Gene mentions containing words like antibody, Ab, antigen, etc. were removed from the final list. The species taxon identifiers and gene mention pairs were associated based on their proximity and these pairs were used to query Entrez gene database (querying being limited to official gene name, official symbol and synonyms). The first Entrez Gene identifier, if found, was considered to be the unique for that gene mention. Several variations of window size for proximity and gene-species associations were tried to get the optimal results on the training set.

The input forms were designed in HTML/CSS/JavaScript and PostgreSQL was used as the backend database. Requests from the input form are processed on the web server using CGI scripts.

References

1. Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists**. *BMC Bioinformatics* 2005, **6 Suppl 1**:S11.
2. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu HH, Torres R, Krauthammer M, Lau WW, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L: **Overview of BioCreative II gene normalization**. *Genome Biol.* 2008, **9 Suppl 2**:S3.
3. **BioCreAtIvE: Critical Assessment of Information Extraction in Biology** [http://www.biocreative.org/].
4. **The NCBI Entrez Taxonomy Homepage** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy].
5. **Entrez Gene** [http://www.ncbi.nlm.nih.gov/gene].
6. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature**. *BMC Bioinformatics* 2010, **11**:85.
7. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text**. *Bioinformatics* 2005, **21**:3191–3192.
8. **Alias-i. 2008. LingPipe 4.0.0.** [http://alias-i.com/lingpipe(accessedMay24,2010)].

Figures



THE UNIVERSITY OF IOWA

BIOCREATIVE III IAT TASK

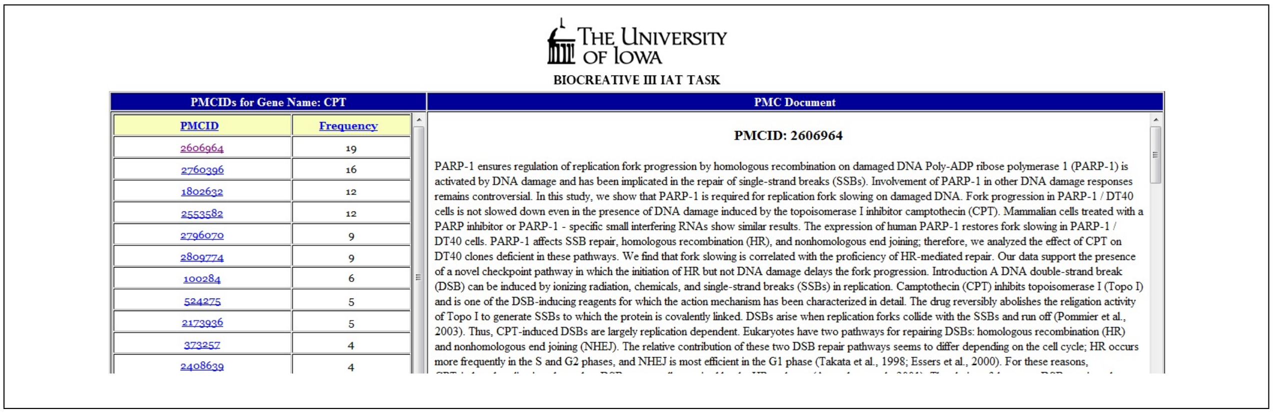
PMC Document # 100345

Molecular interactions of human Exo1 with DNA Human Exo1 is a member of the RAD2 nuclease family with roles in replication, repair and recombination. Despite sharing significant amino acid sequence homology, the RAD2 proteins exhibit disparate nuclease properties and biological functions. In order to identify elements that dictate substrate selectivity within the RAD2 family, we sought to identify residues key to Exo1 nuclease activity and to characterize the molecular details of the human Exo1 - DNA interaction. Site-specific mutagenesis studies demonstrate that amino acids D78, D173 and D225 are critical for Exo1 nuclease function. In addition, we show that the chemical nature of the 5 -terminus has a major impact on Exo1 nuclease efficiency, with a 5 -phosphate group stimulating degradation 10-fold and a 5 -biotin inhibiting degradation 10-fold (relative to a 5 -hydroxyl moiety). An abasic lesion located within a substrate DNA strand impedes Exo1 nucleolytic degradation, and a 5 -terminal abasic residue reduces nuclease efficiency 2-fold. Hydroxyl radical footprinting indicates that Exo1 binds predominantly along the minor groove of flap DNA, downstream of the junction. As will be discussed, our results favor the notion that the single-stranded DNA structure is pinched by the helical arch of the protein and not threaded through this key recognition loop. Furthermore, our studies indicate that significant, presumably biologically relevant,

Total Gene Mentions Found: 32 (Select Gene Name or Species Name to Highlight in Article)

Gene Name	Species Name	Species ID	Entrez ID	Popularity
Exo1	Human	9606	9156	40
Exo1	Schizosaccharomyces pombe	4896	2540348	40
Exo1	mice	10090	26909	40
Exo1	yeast	559292	854198	40
Fen1	Human	9606	2237	20
Fen1	mice	10090	14156	20
Hex1	yeast	559292	852639	4
Hex1	mice	10090	15211	4
Hex1	human	9606	8571	4
endonuclease	mice	10090	640496	3

Figure 1: IAT Indexing Task — The screen-shot shows a sample output for the indexing task. The full-text of the selected article is displayed in the left frame while the gene name, species name, species ID, Entrez Gene ID and Frequency are displayed in the right frame of the web page.



THE UNIVERSITY OF IOWA

BIOCREATIVE III IAT TASK

PMCIDs for Gene Name: CPT

PMCID	Frequency
2606964	19
2760396	16
1802632	12
2553582	12
2796070	9
2809774	9
100284	6
524275	5
2173936	5
373257	4
2408639	4

PMC Document

PMCID: 2606964

PARP-1 ensures regulation of replication fork progression by homologous recombination on damaged DNA Poly-ADP ribose polymerase 1 (PARP-1) is activated by DNA damage and has been implicated in the repair of single-strand breaks (SSBs). Involvement of PARP-1 in other DNA damage responses remains controversial. In this study, we show that PARP-1 is required for replication fork slowing on damaged DNA. Fork progression in PARP-1 / DT40 cells is not slowed down even in the presence of DNA damage induced by the topoisomerase I inhibitor camptothecin (CPT). Mammalian cells treated with a PARP inhibitor or PARP-1 - specific small interfering RNAs show similar results. The expression of human PARP-1 restores fork slowing in PARP-1 / DT40 cells. PARP-1 affects SSB repair, homologous recombination (HR), and nonhomologous end joining; therefore, we analyzed the effect of CPT on DT40 clones deficient in these pathways. We find that fork slowing is correlated with the proficiency of HR-mediated repair. Our data support the presence of a novel checkpoint pathway in which the initiation of HR but not DNA damage delays the fork progression. Introduction A DNA double-strand break (DSB) can be induced by ionizing radiation, chemicals, and single-strand breaks (SSBs) in replication. Camptothecin (CPT) inhibits topoisomerase I (Topo I) and is one of the DSB-inducing reagents for which the action mechanism has been characterized in detail. The drug reversibly abolishes the religation activity of Topo I to generate SSBs to which the protein is covalently linked. DSBs arise when replication forks collide with the SSBs and run off (Pommier et al., 2003). Thus, CPT-induced DSBs are largely replication dependent. Eukaryotes have two pathways for repairing DSBs: homologous recombination (HR) and nonhomologous end joining (NHEJ). The relative contribution of these two DSB repair pathways seems to differ depending on the cell cycle; HR occurs more frequently in the S and G2 phases, and NHEJ is most efficient in the G1 phase (Takata et al., 1998; Essers et al., 2000). For these reasons,

Figure 2: IAT Retrieval Task — The screen-shot shows a sample output for the retrieval task. The PMCIDs and Frequencies for a selected gene name are displayed in the left frame while the full-text of a selected PMCID is displayed in the right frame of the web page.

Cross-species Gene Normalization at the University of Iowa

Sanmitra Bhattacharya*[1], Aditya K Sehgal[2] and Padmini Srinivasan[1,3]

[1]Department of Computer Science, The University of Iowa, 14 MacLean Hall, Iowa City, Iowa 52242, USA
[2]Parity Computing Inc., 6160 Lusk Blvd. Suite C205, San Diego, CA 92121, USA
[3]Department of Management Sciences, The University of Iowa, S210 PBB, Iowa City, Iowa 52242, USA

Email: Sanmitra Bhattacharya*- sanmitra-bhattacharya@uiowa.edu; Aditya K Sehgal - a.sehgal@paritycomputing.com; Padmini Srinivasan - padmini-srinivasan@uiowa.edu;

*Corresponding author

Abstract

Background: With the increasing availability of full text articles through open access publishing, the scope of biomedical text mining is no longer limited to the abstracts of research literature. Cross-species gene normalization using full-text articles is an important step towards the use of full text articles in the area of biomedical text-mining research. This was one of the goals of the BioCreative III Challenge.

Results: In this paper, we present a gene normalization strategy based on the identification of gene and species entities in full text articles and their association. ABNER and LingPipe were used as gene name taggers and LINNAEUS was used for species identification. To associate a species name with a gene name, proximity of the gene and species names was considered. Various window sizes for character boundaries were chosen for this proximity-based association. Based on these associations, a unique Entrez gene identifier, if found, was returned for each gene mentioned in an article.

Conclusions: For the test set, our estimation shows best results with a strategy that considers only the Entrez Gene identifiers found in common by separate runs using ABNER and LingPipe as gene name taggers. This strategy used a 1000 character boundary window for gene-species name association. The highest TAP-*k* ($k = 20$) score returned by our system was 0.1662 for this strategy (Run 2).

Background

The volume of biomedical literature is expanding exponentially. Text mining in the biomedical literature has been widely applied to several biological problems including gene mention identification, gene name normalization, protein-protein interaction, gene-protein interaction, gene-drug interaction and so on. However, little research is available on biomedical text-mining using full-text articles as opposed to article abstracts [1]. The 3^{rd} Critical Assessment for Information Extraction in Biology challenge, BioCreative III [2] took the research in that direction with the use of full-text articles in the cross-species gene normalization (GN) task.

The goal of the gene normalization (GN) task is to determine the unique identifiers like Entrez Gene IDs [3, 4] of genes and gene products mentioned in scientific literature. There are several challenges to gene normalization. First, genes are often referred to in descriptive terms instead of precise gene names or symbols as in 'v-rel reticuloendotheliosis viral oncogene homolog A (avian)' for the gene name 'Rela' (Entrez Gene ID: 19697). This makes the association of a gene mention with a correct identifier difficult. Secondly, gene mentions are highly ambiguous. For example, the gene name 'RRM1' returns 44 results in Entrez Gene database. Out of these results, 'RRM1' gene of Human has an Entrez Gene ID of 6240 while 'RRM1' gene of Mouse has an Entrez Gene ID of 20133. It is to be noted here that the specific combination of a gene/protein mention and a species identifier returns an unique identifier. Thus in an article the choice of correct identifier (Entrez Gene ID) depends on the context i.e. the species indicated (sometimes implicitly) when the gene is discussed. A third problem is that species names can also be referred to in an article indirectly like patient, women, etc. which actually refer to the same species identifier as human (NCBI Taxonomy ID: 9606). Also shorthand species names like *C. elegans* refer to 41 species in the NCBI Taxonomy. These problems complicate the challenge of determining the correct identifier for a given mention of a gene in an article.

Related Work

Hakenberg et al. [5] proposed the first publicly available inter-species gene normalization technique called GNAT. For gene name recognition they used a set of gene dictionaries for all candidate species and coupled that with gene name recognition by BANNER [6], a CRF-based tagger. Species names were identified either by using AliBaba [7] or from the cell-line information. Gene and species names were associated using using various criteria such as compound nouns or phrases having both gene and species terms, etc. Background information from text was used to select a unique identifier from a list of candidate identifiers. Neves et al. [8] proposed an open-source Java based gene/protein tagger and normalization system, Moara, which uses a trainable CBR-Tagger for gene/protein identification and ML-Normalization for the normalization task. Recently, a species-based gene normalization strategy had been proposed by Verspoor et al. [9], which is similar to our approach. Their system performs a dictionary-based gene/protein and species recognition followed by gene/protein name ambiguity resolution. Proteins are associated to species using several strategies for ambiguity resolution. A confidence score is given to these normalized proteins based on the method used for species association.

Results and Discussion

BioCreative III introduced a new evaluation metric called TAP-*k* for the GN evaluations. TAP-*k* is closely related to the widely used average precision measure in information retrieval. The idea behind the method is that, if a retrieval system generates outputs which are ranked in the descending order of some confidence score, then it might be useful to cut-off the retrieval at some fixed threshold score. In case there are true positives below a certain threshold, a precision of 0 is assigned while calculating the average precision over all relevant records. The threshold is chosen as the largest score that produces a median of *k* false positive retrievals over the set of queries. A more detailed description of the method is available here [10]. The values of *k* chosen for BioCreative III evaluations were 5, 10 and 20.

GN task evaluations

For the GN task we experimented with two widely-used gene name taggers namely ABNER [11] and LingPipe [12] while LINNAEUS [13] was used for species name identification. Various combinations of these taggers and different confidence scores were used to set up various experiments. Here we present a brief description and evaluation of those experiments which were used in our BioCreative test submissions.

Run 1: Conditional assignment of majority species

In this strategy, we consider the count of the species mentions in an article. A gene is associated with a species if it is found within the specified character window. Here LingPipe was used as the gene name tagger. If an association is not found in the given window then the species occurring most frequently is associated with the gene name. Confidence was calculated using the count of majority species mention divided by the total count of all species names appearing in that article.

Run 2: Run 1 strategy with LingPipe/ABNER intersection

For this run we executed separate runs using ABNER and LingPipe with a similar strategy as in Run 1. An intersection of the associations produced by these two systems was considered for the test set submission.

Run 3: ABNER 'intersection' LingPipe run

For this run, gene mentions were tagged by ABNER and LingPipe separately. For each gene name the system searches for species names within a specified character window. A confidence score is calculated based on the proximity of the associated gene and species names. This confidence can in turn be used as a cut-off threshold score for the associations. Here we consider only the gene-species associations that are identified by both systems. For such overlapping associations, the higher confidence score of the two was retained.

According to our estimates, the highest TAP-k ($k = 20$) score achieved was 0.166 for the Run 2 strategy. The scores for TAP-k (k = 5, 10) for this run were 0.0829 and 0.1161 respectively. For the Run 1, the TAP-k (k = 5, 10, 20) scores were 0.0577, 0.0726 and 0.1106 respectively. For the Run 3 run the scores were 0.0830, 0.1091 and 0.1387 respectively. These results are summarized in Table 1. Besides the strategies described above we experimented with a some other strategies where ABNER and LingPipe were used separately or together (union) as gene taggers. However, the performance of these strategies was worse compared to the three strategies chosen for test set submission.

Conclusions

In this paper we have presented a cross-species gene normalization system from full-text PMC articles. We have experimented with various strategies for gene name identification and gene-species name associations under varying conditions. The performance of the different strategies suggest that the idea of identification of gene and species entities and association of these entities are valuable for gene name normalization. Our experiments showed that associations that considered larger character boundary windows proved to be more effective than smaller character boundary windows. In case of lack of association in a smaller window, the assignment of majority species to gene names also showed performance improvement. Based on the results obtained from our study, we believe that the performance of the gene normalization system can be improved further by the inclusion of external knowledge resources. The gene mention step can be broken down into a two step process – a dictionary based gene/protein identification followed by gene mention identification by gene name taggers. This would most likely improve both the precision and recall of our system. Post-submission we found that a stricter (i.e. field limited) querying of Entrez Gene for the gene-species associations reduces the number of false positives. Also, a better selection of terms for the stop list and filtering of gene names to remove potential tagging errors will help in improving the system performance.

Methods

Data

The training data for BioCreative III gene normalization task consisted of two types of data. One set consisted of a small number (32) of PubMed Central (PMC) articles fully annotated by a group of trained and experienced curators. The second set consisted of a larger number (500) of partially annotated articles. For each of these training sets, a list of Entrez Gene IDs corresponding to each document were provided. The test set comprised of over 500 full-text articles, 50 of which were chosen for evaluations.

Conversion of XML files to text files

The full-text XML files provided for the GN task were stripped off the XML tags and converted to plain-text articles. Following the annotation guidelines from BioCreative III organizers, gene mentions from certain sections of the text had to be omitted from the normalization process. Hence, sections like References/Bibliography, Supplementary Materials/Supporting Information and Methods/Materials sections were removed from the text used for further processing. Also XML codes for Greek symbols like alpha, beta, etc. were replaced with corresponding Greek names. The resulting text

Strategy	Gene Tagger	Character Boundary	TAP-5	TAP-10	TAP-20
Run 1	LingPipe	1000	0.0577	0.0726	0.1106
Run 2	ABNER/LingPipe	1000	0.0829	0.1161	0.1662
Run 3	ABNER/LingPipe	10000	0.0830	0.1091	0.1387

Table 1: Evaluations of Test Set runs

was used in the subsequent steps.

Gene/Protein identification

In this step we needed to identify the gene/protein mentions in the text. According to the annotation guidelines, all gene names including those mentioned in passing or only once in an article had to be identified. Two CRF-based taggers, ABNER and LingPipe were used for this purpose. ABNER (trained on NLPBA corpus) and LingPipe (trained on GENIA corpus) have been used extensively for gene mention recognition in previous BioCreative tasks and have shown consistently high performance. The gene mentions identified by ABNER and LingPipe were filtered using a stop list. The stop list is also governed by the annotation guidelines from BioCreative organizers. Following the guidelines, we remove the gene mentions which contains words like antibody, *Ab*, antigen, immunoglobulin, *IgG*, reagent, substrate, enzyme, complex, family, super-family, transcription factors, etc. However, certain condensed gene names mentioned as a range had to be expanded to include constituent names. For example, *Xnr1-Xnr6* (or *Xnr1-6*) and *Cdk1/2* were expanded to *Xnr1, Xnr2, ..., Xnr6* and *Cdk1* and *Cdk2*, respectively.

Species name identification

Species name identification was very important for our system. We used an open-source species name recognition and normalization software system, LINNAEUS, for this purpose. LINNAEUS returns a normalized list of species identifiers for species names identified in the article. The LINNAEUS species dictionary was modified to include first names of model organisms. For example we added *Arabidopsis* (for model organism *Arabidopsis thaliana*) to species identifier 3702, *Xenopus* (for model organism *Xenopus laevis*) to species identifier 8355, etc. We also added new entries to the species dictionary for widely used strains of organisms such as *Escherichia coli K-12, Saccharomyces cerevisiae S288c*, etc.

Gene-species name association

The gene mentions and the species names identified were associated based on proximity and character boundary windows were used for their association (e.g. gene and species mention occurring within 1000 characters of each other in Run1 and Run 2). Three types of gene-species associations (discussed earlier) were considered. A confidence score, measured on a scale of 0.0 − 1.0 based on the distance between the gene and the species names was calculated depending on the selected strategy.

Entrez Gene ID retrieval

For each of the gene species associations, we search Entrez Gene for a unique identifier for that gene mention. The first Entrez Gene ID retrieved from this search was returned as the unique identifier for that gene name. Further analysis after submission revealed that limiting the search on Entrez Gene to fields like official gene/protein name, official symbol and synonyms proved to be beneficial.

References

1. Schuemie MJ, Weeber M, Schijvenaars BJ, van Mulligen EM, van der Eijk CC, Jelier R, Mons B, Kors JA: **Distribution of information in biomedical abstracts and full-text publications**. *Bioinformatics* 2004, **20**:2597–2604.
2. **BioCreAtIvE: Critical Assessment of Information Extraction in Biology** [http://www.biocreative.org/].
3. **Entrez Gene** [http://www.ncbi.nlm.nih.gov/gene].
4. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res.* 2007, **35**:26–31.
5. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT**. *Bioinformatics* 2008, **24**:126–132.
6. Leaman R, Gonzalez G: **BANNER: an executable survey of advances in biomedical named entity recognition**. *Pac Symp Biocomput* 2008, :652–663.

7. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph**. *Bioinformatics* 2006, **22**:2444–2445.

8. Neves ML, Carazo JM, Pascual-Montano A: **Moara: a Java library for extracting and normalizing gene and protein mentions**. *BMC Bioinformatics* 2010, **11**:157.

9. Verspoor K, Roeder C, Johnson HL, Cohen KB, Baumgartner WA, Hunter LE: **Exploring species-based strategies for gene normalization**. *IEEE/ACM Trans Comput Biol Bioinform* 2010, **7**:462–471.

10. Carroll HD, Kann MG, Sheetlin SL, Spouge JL: **Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics**. *Bioinformatics* 2010, **26**:1708–1713.

11. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text**. *Bioinformatics* 2005, **21**:3191–3192.

12. **Alias-i. 2008. LingPipe 4.0.0.** [http://alias-i.com/lingpipe(accessedMay24,2010)].

13. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature**. *BMC Bioinformatics* 2010, **11**:85.

Identifying protein-protein interactions in biomedical text articles

Rezarta Islamaj Doğan[1], Yi Yang[2], Aurélie Névéol[1], Minlie Huang[2], Zhiyong Lu[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China

Abstract

In this paper we present our approaches in identifying protein-protein interaction (PPI) articles and specific protein-protein interaction methods in biomedical text articles during our participation in the PPI task of BioCreative Challenge 2010. The first PPI task required classifying whether a given article contains relevant information for protein-protein interaction study purposes. Our approach to this task included exploitation of different textual feature types in a wide margin classifier setting. The second PPI task required identifying which protein-protein interaction methods were relevant for a given article. Our approach to this task included manually perfected rules, regular expressions and heuristics, custom trained supervised models, and learning to rank methods.

Article classification task

The article classification task (ACT) starts with a list of PubMed records, and aims at indentifying which records contain descriptions indicating that that article is relevant for protein-protein interaction curation purposes.

Data

Training data for BioCreative III ACT task contained an equal number of positive and negative articles, but the development and test datasets were imbalanced. In addition, the negative articles provided in the training dataset were not sufficiently representative of the non-PPI relevant articles. This was verified when the model learned on the training dataset did not exhibit the same performance when tested on the development dataset, and vice-versa. In response to these issues, we enriched our training dataset by collecting all data provided for this task from BioCreative II and BioCreative II.5 challenges. Furthermore, we used the related articles function of PubMed to extract three closest articles to each article labeled as negative in our dataset. After removing duplicates, this process provided us with 5,138 positive and 7,754 negative articles.

Methods

As our principal learner, we selected a wide margin classifier similar to a linear support vector machine, which have been proven to handle sparse high dimension data, and have been shown to work for various text classification tasks [6]. Different from SVM, our learner was based on the modified Huber loss function [7]. In comparison, this function is differentiable and the speed of optimization is faster. Here we describe the different feature models we explored for classifying an article that contained protein-protein interaction mentions.

Bag-of-words features model. Each article was represented as a vector of words found in its title and abstract. We did not perform stemming, because un-stemmed features provided better results for this task. With respect to class imbalance, we bootstrapped the negative documents by sampling, with replacement, so that each negative article was counted at least six times.

Bigram features model. Each article was represented as a vector of two consecutive words found in its title and abstract in the same sentence. Stop-words were included.

Co-occurring features model. Each article was represented as a vector of two co-occurring words. This model extended the bigram model to including all the words which co-occurred in the same sentence. Word position within the sentence was not taken into consideration. Stop-words were included. Stemming was performed in order to reduce the resulting number of different features.

String features model. Each article was represented as a vector of character strings of length 8. A sliding window of eight characters within each sentence generated all unique strings.

K-nearest neighbors method. For a given article in the test set, we computed the similarity to articles in the training dataset using a vector space model with TF-IDF scheme [5]. The test article was scored based on the 10 nearest neighbors' classification labels, and their collective pair-wise similarity score to the test article.

String matching rules. This method used a set of hand-crafted rules to check whether any signature protein-protein interaction terms could be found in the title or abstract of a given article. This method simply counted the number of matched rules.

Feature selection method. An iterative feature selection algorithm was applied to each set of features described above. For each set of features, we trained the wide margin classifier and noted their weights. We averaged all the features weights using five-fold cross-validation and eliminated 1000 features whose weights were closest to zero. The classifier was trained again on the remaining features. After every step of the feature selection process, we applied the learned model to the development dataset and noted the performance. Finally, we selected 15,000 features for the bag-of-words features model, and 10,000 features for each of the bigram, co-occurring and string feature models, respectively.

Merging decisions method. Two different decision methods were developed to combine the different models applied to the ACT test dataset. First, the strict decision making schema assigned an article to the positive class only if all the models participating in the decision making had done so. Second, the log-linear classifier schema combined the individual score outputs for different models participating the decision making in a weighted fashion. Finally, an article was assigned to the positive class, only if the combined score exceeded an empirically decided threshold.

Results

We submitted four runs to the ACT task based on the methods presented above. Our first run consisted of the bag of words model after feature selection. Our second run consisted of merging the bigram features model, the co-occurrence features model and the string features model, prior to feature selection, using the strict decision making schema. Our third run was produced similar to run two, but after feature selection was applied to each of the three models individually. Finally, our forth run used the log-linear classifier to combine the bag-of-words, bigram, string matching and kNN models. Our results are presented in Table 1 (development dataset) and Table 2 (test set, as reported by challenge organizers)

Table 1 Overall performance of our submitted runs, on the development set

	TP	FP	FN	Sensitivity	Specificity	Accuracy	Matthew's	AUC iP/R
Run1	310	95	372	0.455	0.971	0.883	0.531	0.673
Run2	333	169	349	0.488	0.949	0.871	0.497	0.612
Run3	646	439	36	0.947	0.868	0.881	0.689	0.887
Run4	491	880	191	0.720	0.735	0.732	0.360	0.549

Table 2 Overall performance of our submitted runs, on the test set, as reported by the challenge organizers

	TP	FP	FN	Sensitivity	Specificity	Accuracy	Matthew's	AUC iP/R
Run1	398	162	512	0.437	0.968	0.888	0.500	0.616
Run2	517	311	393	0.568	0.939	0.883	0.527	0.619
Run3	659	881	251	0.724	0.827	0.811	0.453	0.603
Run4	694	873	216	0.763	0.829	0.819	0.483	0.637

Interaction method task

The interaction method task (IMT) aims at determining the PPI technique(s) to support the interactions found in the article. Using the standardized terminology, a ranked list of protein-protein interaction methods taken from the PSI-MI ontology are selected and assigned to biomedical articles.

Data

Results from BioCreative II showed that processing full text articles, in particular Material and Method section, was useful for this task [3,4]. In accordance, for each document in the training dataset of BioCreative III, we extracted: title, abstract, methods/materials, and figure captions sections. Adding the articles of BioCreative II and additional data obtained from the MINT database [2] produced a set of 3,765 articles annotated with PSI-MI codes. We refer to this set as "the reference collection". Full text was available for a subset of the articles in the reference collection, while title and abstract text was available for all articles.

Methods

Problems involving automatic assignment of controlled vocabulary terms to biomedical articles, such as MeSH indexing, have been shown to benefit from combination of natural language processing and machine learning methods [1]. As such, we focused our efforts on developing text analysis and machine learning methods that could be combined for optimal performance.

Pattern matching method. A pattern matching algorithm relying on PSI-MI terms and synonyms was developed and applied to the different types of text obtained for each article: full text, material and methods text and figure caption text. Each code was extracted with a score based on the number of patterns that were found to identify it.

K nearest neighbours method. For a given article requiring assignment of PSI-MI codes, we computed the similarity to articles in the reference collection using a vector space model with TF-IDF scheme [5] and computed a score for each PSI-MI code assigned to at least one of the k nearest neighbours. The optimal value of k and a score threshold for selecting a given PSI-MI code were determined by empirical experiments on the BC3 training set.

Mapping MeSH to PSI-MI method. Biomedical articles of interest for protein-protein interaction are usually published in journals indexed in MEDLINE. As such, they are assigned MeSH indexing terms. Because there are no direct links between PSI-MI and MeSH, we manually reviewed the PSI-MI codes from the training set and developed MeSH to PSI-MI mappings. For example, PSI-MI code MI:0114 (x-ray crystallography) was mapped to the main heading "Crystallography, X-Ray" and code MI:0077 (nuclear magnetic resonance) was mapped to "Magnetic Resonance Spectroscopy". About 74% of PSI-MI codes did not have any MeSH equivalents e.g., MI:0415 (enzymatic study). This method provided a binary indication on whether a given code could be obtained from mapping MeSH indexing of the article to PSI-MI.

Merging methods. The analysis of results yielded by each of the methods described above showed that per-code performance varied significantly from code to code within a method and from method to method for a given code. As a result, decided to merge methods based on code performance on the training set. For instance, pattern matching on full text was used for code MI:0071, pattern matching on caption text was used for code MI:0096, k-NN was used for code MI:0018, MeSH mapping was used for code MI:0114 and so on. Overall, the results reflected the class imbalance observed in the training dataset, so we did not address this issue further.

Learning-to-rank method. This approach consists of three steps: finding nearest neighbors, extracting features for each annotation and scoring each annotation with the ranking model.

Finding nearest neighbours. First, for each article in the training set, we retrieved 50 nearest neighboring documents using a different local implementation of the standard algorithm. We represented each document with a vector of top 1000 TF-IDF-weighted words per document, and used cosine similarity to determine the similarity between documents. We experimented with bag-of-words vectors extracted from full text, title and abstract, and methods and materials. For each representation, we evaluated the resulting model on the development dataset, and decided to use the methods and materials section to represent each document.

Extracting features. Second, we collected the PSI-MI annotations for each of the 50 neighbor documents. This provided us with a list of PSI-MI codes, for which, we extracted these features:

a) *Query-likelihood features:* The experiment code name and its synonyms are viewed as query. We used BM25 model to compute likelihood scores between query annotation and the document. The highest score is chosen as the feature value.
b) *Neighborhood features:* There are two neighborhood features. The first counts the neighboring documents to which a candidate code was assigned. The second sums up the document similarity scores for each neighboring document to which the candidate code was attached.
c) *Synonym features:* There are two binary features: the first indicates whether the name or synonyms of a code can be exactly found in the document; the second notes whether there exists a code or synonym whose individual words have all been observed in the document.
d) *Tf-idf features*: For each code, we chose two lists of signature terms from its name and definition text respectively. For each list, we designed two binary features. The two features note whether one or more than 1/3 words in a signature term list occur in the document.

e) *Key-imt- signature feature:* We manually computed a set of keywords for the most frequent 7 codes that cover over 80% all annotations. This binary feature indicates whether one word in the keywords can be found in the document

Ranking Method. Third, we assigned a score to each annotation in this list, using the ranking model. We used a list-wise learning-to-rank algorithm to ranking the codes. We choose ListNet as our ranking algorithm [9]. We submitted the top N ranked codes for each document and performed several experiments to find the best N.

Extracting Evidence Text

For each PSI-MI code assigned to the articles, evidence text had to be provided. We used the pattern matching method to select the evidence text among the full text sentences that contained a relevant pattern for the codes.

Table 3 Overall performance of our submitted runs, on the development set

	TP	FP	FN	Micro precision	Micro recall	Micro f-measure	Micro AUC iP/R	Macro precision	Macro recall	Macro f-measure	Macro AUC iP/R
Run1	811	754	568	0.518	0.588	0.551	0.331	0.513	0.619	0.520	0.517
Run2	691	680	688	0.504	0.501	0.503	0.270	0.458	0.520	0.456	0.447
Run3	584	907	795	0.392	0.423	0.407	0.185	0.320	0.455	0.351	0.361
Run4	583	904	796	0.392	0.423	0.407	0.184	0.320	0.455	0.351	0.360
Run5	460	344	919	0.572	0.334	0.421	0.192	0.370	0.340	0.330	0.293

Table 4 Overall performance of our submitted runs, on the test set, as reported by the challenge organizers

	TP	FP	FN	Micro precision	Micro recall	Micro f-measure	Micro AUC iP/R	Macro precision	Macro recall	Macro f-measure	Macro AUC iP/R
Run1	272	338	234	0.446	0.538	0.487	0.271	0.473	0.550	0.471	0.433
Run2	289	436	238	0.399	0.548	0.462	0.270	0.412	0.546	0.442	0.432
Run3	235	431	292	0.353	0.446	0.394	0.157	0.353	0.455	0.375	0.325
Run4	235	430	292	0.353	0.446	0.394	0.158	0.353	0.455	0.375	0.325
Run5	96	79	203	0.549	0.321	0.405	0.196	0.564	0.307	0.370	0.294

Results

We submitted five runs to the IMT task based on the methods presented above. Our first run consisted of merging text analysis, kNN and Mapping MeSH to PSI-MI. Our second run consisted of kNN code assignments selected only if text evidence was available (in this run no evidence text was selected randomly). Our third and four runs were produced using the Ranking Method. Both of these runs use features extracted from methods/materials sections of the documents. The difference between run 3 and run 4 lies in the number of annotations that they produced. Run 3 returned the top 3 annotations for each test document. Run 4 instead used a score threshold to decide on the number of PSI-MI codes to assign to each test document, and returned up to three annotations for each test document. Finally, our run 5 was optimized for precision by combining the results of run 2 and run 3. In this setting, for each test document, only the PSI-MI codes which were predicted by both methods were reported.

Conclusions and Discussion

Our approaches presented here are applicable to many different text categorization tasks. For the ACT task, we presented a collection of feature construction methods which were able to capture sufficiently the PPI-relevancy of a given biomedical record. We were surprised by the fact that stop words were important features for all the different models, however expressions such as "interacts with" or "pull down" may indeed indicate a protein-protein interaction relevant record. This reveals that the use of sophisticated hand-crafted string matching rules as well as specialized name entity recognition systems should be particularly helpful in increasing the accuracy performance of this task. On the other hand, a higher performance can also be obtained with the supervised methods that we presented in this work. However, a well-represented, larger dataset is required for training. The biggest challenge we encountered was the fact that the irrelevant articles were not sufficiently representative for the entire sample space.

For the IMT task, the problem has other pre-requisites. A specific PPI interaction method is generally not the main topic or focus of a research article. As such, those specific terms are rarely mentioned in an article's title and abstract. Rather, these details usually appear in figure captions and in methods/materials sections. So full text processing is a must. Another difficulty is the fact that the section titles which contain this useful information are not standard among different publication venues. Finally, authors often use different terms when describing their methods than those found in the compiled controlled vocabulary lists. Mappings between the PPI-methods mentions in the text and their definitions in the ontology entries, as a result, require further study.

References

1. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. **The NLM Indexing Initiative's Medical Text Indexer.** *Stud Health Technol Inform*. 2004;107(Pt1):268-72.
2. MINT, the Molecular INTeraction database. http://mint.bio.uniroma2.it/mint/
3. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. **Overview of the protein-protein interaction annotation extraction task of BioCreative II**. *Genome Biol.* 2008;9 Suppl 2:S4.
4. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Clematide S, Hess M, von Allmen JM, Parisot P, Romacker M, Vachon T. **OntoGene in BioCreative II**.*Genome Biol.* 2008;9 Suppl 2:S13.
5. Manning C, Schütze H. **Foundations of Statistical Natural Language Processing**, MIT Press. Cambridge, MA: May 1999:544
6. Medical Subject Headings. http://www.nlm.nih.gov/mesh/
7. Thorsten, J. **Text categorization with support vector machines: learning with many relevant features.** *Proceedings of ECML-98, 10th European Conference on Machine Learning.*1998; 1398:137-142.
8. Zhang, T. **Solving large scale linear prediction problems using stochastic gradient descent algorithms.** *ICML '04: Proceedings of the twenty-first international conference on Machine learning.* 2004; 116.
9. Cao Z, Qin T, Liu TY, Tsai MF, Li H. **Learning to rank: from pairwise approach to listwise approach.** *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 129–136, New York, NY, USA, 2007. ACM Press.

Fast classification of scientific abstracts related to protein-protein interaction using a naïve Bayesian linear classifier

Jean-Fred Fontaine[1§], Miguel A. Andrade-Navarro[1]

[1] Computational Biology and Data Mining Group, Max-Delbrück-Centrum für

Molekulare Medizin, Robert-Rössle-Str. 10, 13125 Berlin, Germany

[§]Corresponding author

Email addresses:

JFF: jean-fred.fontaine@mdc-berlin.de

MAAN: miguel.andrade@mdc-berlin.de

Abstract

Medline Ranker is a fast and accurate generalist document retrieval tool based on a naïve Bayesian linear classifier that allows the scan of the biomedical literature for any selected topic. We have used this tool to classify scientific abstracts related to protein-protein interactions, using nouns as features, from the article classification task of the BioCreative 3 challenge. Even if not specialised in this topic, Medline Ranker showed a good performance and its outstanding speed allows its use on millions of abstracts within a few minutes. Availability: http://cbdm.mdc-berlin.de/tools/medlineranker/.

Background

Medline Ranker is a fast document retrieval tool that allows the scan of the recent biomedical bibliography for any selected topic. Its algorithm based on naïve Bayesian statistics was shown to be very precise in various benchmarks, classifying for instance abstracts related to radiology or Alzheimer's disease [1]. The Medline Ranker web server has a simple, flexible and powerful user interface that allows the selection of a training set, a background set, and a test set represented as PubMed identifier (PMID) lists. Thus, it can be used directly to process BioCreative 3 datasets using default parameters, though careful parameter selection would positively impact performance.

Medline Ranker applies a linear naïve Bayesian classifier (LNBC) using nouns as features with a high processing speed (approximately 18000 abstracts per seconds). Even if not specialised in the topic of protein-protein interactions (PPIs), for example by extracting gene mentions, our tool may be of interest in this task because it can scan the ever growing scientific literature, already composed of millions of references, in a few minutes. This article describes our participation using Medline Ranker to the BioCreative 3 abstract classification task (ACT) on PPI related abstracts (http://www.biocreative.org).

Results

Organisers of the BioCreative 3 ACT provided each registered team with a balanced training set composed of 2280 labelled abstracts (1140 positives and 1140 negatives), a development set of 4000 labelled abstracts (682 positives and 3318 negatives), and a test set of 6000 unlabelled abstracts. The training set was used to estimate the parameters, the development set for testing, and results of five runs of Medline Ranker on the test set were sent to the ACT organisers who evaluated classification performances.

Parameter estimation

Two parameters of Medline Ranker are relevant to the ACT: the minimal frequency of a selected noun into the training set, and the P-value cutoff to decide positive and negative predictions. Parameters were defined by observing classification performance on leave-one-out cross validations of the training set. Nouns not found often enough may introduce noise in the trained statistical model [2]. To be used for classification, a noun was required to occur at least 9 times in the training set because this value maximized the area under the receiver operating characteristic curve (AROC=0.898). For each of the five possible runs on the test set, we selected a P-value cutoff to have a high Matthew's correlation coefficient (MCC), a high recall, or a high specificity (Table 1). Maximal values were: MCC=0.644, recall=0.950 and precision=0.940.

Training and results on the test set

Five runs were done on the test set by training the algorithm with the defined parameters. The training defined 833 nouns including these top discriminative terms: ubiquitination, hybrid, coimmunoprecipitation, and coactivator. We observed that terms related to protein post-translational modification (PTM) were highly ranked (ubiquitination, phosphorylation and sumoylation had a rank equal to 1, 8, and 48 respectively). From preliminary results of the five runs on the test set, the Medline Ranker tool showed a maximal MCC of 0.464, a maximal recall of 0.933 and a maximal precision of 0.678 (Table 2). The mean run total duration was 1.294 seconds with a standard deviation of 0.08 seconds.

Discussion

The Medline Ranker tool was used in the BioCreative 3 ACT to classify abstracts related to PPI. The machine learning algorithm was trained with the provided training set only and various runs on the test set showed sensitive, precise, or balanced results. On the one hand, the training set was composed of an equal number of positive and negative abstracts (total=2280); on the other hand, the test set was unbalanced (910 positives and 5090 negatives). That led to differences in observed classification performances of the two sets (Table 1 and Table 2). Medline Ranker applies a class imbalance correction on the training set [1, 3], but not on the test set where labels are supposed to be unknown. Discriminative words were biased to PTM reflecting involvement of these mechanisms in PPI [4]. As the training set was relatively small, it may not properly sample the literature of PPI-relevant journals as expected.

The running time in processing BioCreative 3 datasets was very short (8280 abstracts in total for an average duration of 1.294 seconds). This is mostly explained by extensive pre processing and storage of the whole MEDLINE. Moreover, even if not as accurate as support vector machine classifiers, training a LNBC is significantly faster [5] and it allows our tool to process millions of abstracts with good performance in a practical time.

Methods

Data and algorithm for abstract classification were described previously [1]. Briefly, MEDLINE XML data are downloaded weekly and stored in a MySQL database. Only records having an English abstract are stored. After part-of-speech processing of each abstract, nouns are retained and stored in the database as abstract profiles. A stop word list is used to remove common and non meaningful terms. For classification, a LNBC is trained on abstract profiles and multiple occurrences of nouns in a single abstract are not counted [1, 5]. The algorithm returns a P-value representing the confidence in classification for an abstract. P-values are extrapolated from a simulation on 10 000 randomly chosen abstracts in MEDLINE. Two equally scored abstracts are randomly assigned consecutive ranks. Scores for BioCreative 3 are obtained by subtracting the P-value to 1, after truncation of the P-value to]0,1[.

Acknowledgements

This project is funded within the framework of the Medical Genome Research Programme NGFN-Plus by the German Ministry of Education and Research (reference number: 01GS08170), and by the Helmholtz Alliance in Systems Biology (Germany).

References

1. Fontaine J, Barbosa-Silva A, Schaefer M, et al.: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic acids research* 2009, **37**:W141-6.

2. Suomela BP, Andrade MA: **Ranking the whole MEDLINE database according to a large training set using text indexing.** *BMC bioinformatics* 2005, **6**:75.

3. Poulter GL, Rubin DL, Altman RB, Seoighe C: **MScanner: a classifier for retrieving Medline citations.** *BMC bioinformatics* 2008, **9**:108.

4. Deribe YL, Pawson T, Dikic I: **Post-translational modifications in signal integration**. *Nature Structural & Molecular Biology* 2010, **17**:666-672.

5. Wilbur WJ, Kim W: **The Ineffectiveness of Within - Document Term Frequency in Text Classification.** *Information retrieval* 2009, **12**:509-525.

Tables

Table 1 – Training set cross validations

Run	Recall	Precision	MCC	P-value cutoff
1	0.837	0.813	0.644	1.92E-01
2	0.839	0.808	0.640	1.98E-01
3	0.950	0.701	0.582	4.05E-01
4	0.929	0.726	0.603	3.49E-01
5	0.263	0.940	0.355	1.58E-02

Table 2 – Preliminary classification results on the test set

Run	Recall	Precision	MCC
1	0.769	0.416	0.460
2	0.778	0.416	0.464
3	0.933	0.294	0.382
4	0.896	0.322	0.405
5	0.180	0.678	0.301

Inference network method on cross species gene normalization in full-text articles

Hung-Yu Kao *[1], Chih-Hsuan Wei *[1]

[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.

Email addresses:

HYK: hykao@mail.ncku.edu.tw

CHW: p7896116@mail.ncku.edu.tw

Abstract

Background

In order to access and utilize the rich biological information in biomedical literatures, the recognition and normalization of name entities in literatures are necessary and crucial processes. In this paper, we focus on the accuracy improvement of normalization task. Cross species gene normalization is an important and difficult challenge because of the name ambiguity and variation in biological literatures.

Results

We propose a new approach that employs an inference network method to handle these issues. The proposed model utilizes the Term Frequency-Inverse Document Frequency (TF-IDF) weighting strategy to calculate similarity scores among tagged entities and database identifiers.

Conclusions

In conducted experiments, the proposed model attains 45.5% in F-measure and 34.69% in TAP score on the selected 50 articles that received the most different results from pooled team submissions and regarded as the most difficult 50 articles.

Background

Text mining on biomedical data sources has been noted in the last several years, and then scientists dedicated to extract the information automatically and precisely. Evaluating the assistance of text mining on biomedical data sources has been reported that text mining techniques for biomedical information extraction are not completely reliable [1] and remain challenging to increase the assistance in this domain. In the biomedical text mining issue, many researchers developed a lot of automatic information extraction methods for the biomedical literatures. This is mainly consisting of two tasks. Relation extraction (RE) is one of the tasks retrieving the biomedical information that identifies the relationships among biomedical entities in the literature. While extracting relations, each biomedical entity such as gene, protein or disease, etc. refers to map the biomedical entities to the database identifiers from articles. Cross-species gene normalization (GN), are needed in this task. GN has to map gene mention entities to identifiers.

The GN task decides on the correlation species but also normalizes the database identifier to the gene mention and produces a list of the EntrezGene [2] identifiers of all species including human for all the genes/proteins mentions in full text articles. Inter-species gene mention normalization is a particular challenge associated with high ambiguity of gene names, particularly with regards to orthologous genes.

Results

We evaluate our method with the provided full-text articles in the workshop of Biocreative III. The statistics of annotations of full-text articles could help us to understand the phenomena of the statistic of the annotations in Table 1.

We first evaluated our method by Threshold Average Precision (TAP-K) score[3]. As shown in Table 2, our proposed method achieves more than thirty percent on the Biocreative-III gene normalization task. Then, we presented the best result of the TAP-K experiment by F-measure in Table 3, and we obtained 45.53% of F-measure (53.85% precision, 39.44% recall).

Conclusions

In terms of the efficiency, we spent less than one minute to identify the EntrezIDs of a full-text article through the system on 3.4GHz server with 2GB RAM.

We proposed a gene name normalization system for mapping a biomedical entity to the correct EntrezID by applying the similarity-based inference network model. This method can be used to solve both the term variation and ambiguity problem.

Furthermore, the intersection filtering method is useful for the term ambiguity challenge. This method can filter some ambiguous candidate EntrezIDs. Some bag of words may not be an obvious gene name or a gene name, but they may be some of the same related words of the correct EntrezID. Combining these evidences can enhance the inference capability.

Experimental results show that our method archives a 34.69% TAP score, 53.85% precision, 39.44% recall, and 45.53% F-measure by the evaluation data provided by Biocreative III GN task.

Methods

For this gene normalization task, we develop an inference network model shown as the Figure 1. The four modules of this model would be discussed below.

Name entity recognition module (NER)

The gene name entity tagging tool used in our system is AIIA-GMT[4]. It is a XML-RPC client of a web-service server that provides the service to recognize named entities in the biomedical articles.

Due to the varied naming styles of gene names in the biomedical literatures, the tagged entity cannot always exactly match a gene name in the dictionary. To address this issue, we therefore proposed a post-processing module to enhance the ability of general-purpose recognition system. This module includes four translation rules, i.e., the number type, conjunctions, enumerations, and parentheses, are applied to tokenize gene names.

- The first rule is the number type. Numbers of different subunit type have to be unified, e.g., the Roman, Arabian and Latin.
- Secondly, entities with conjunctions needed to be split. Sometimes, two or more gene names were combined into one mention by several conjunctions.
- Then, an enumeration entity with the sequential numbers means several gene names which belong to the same family. We separated the entity to several different gene names by sharing their mutual family name
- In the last rule, abbreviations in the parentheses after a gene name, e.g.,

"Hypoxia inducible factor 1 (HIF1)", should be separately extracted.

Species name entity normalization module (SNEN)

We collect three different species name lexicons, such as NCBI taxonomy, Cell line list from Wikipedia and Linnaeus corpus [5] to construct the species name lexicon. Every species synonyms of the lexicon are used to detect the species name by the dictionary-based matching. To handle two missing cases of the matching result, we devise two robust partial matching strategies.

Firstly, some species entities are genus names. These entities always occur together with the original species name in articles, e.g., "Arabidopsis" used for "Arabidopsis thaliana" when the same article includes these two entities. Secondly, some species entities cannot be extracted exactly, because a species name may has a variety of sub types, e.g., "Escherichia coli strain k-12 substrain mg1655" is same to "E. coli str. k-12 substr. mg1655", "E. coli mg1655" and "E. coli k-12 mg1655". These synonyms are too highly varied to match. For this case, the leaf type, i.e., the last subtype of the species, is oriented to a key identifier of this species. After the species entity recognition is finished, the mention species entities are used to extract all candidate species sub types of this one. For example, "Escherichia coli" can find 45 key identifiers and "mg1655" indicates to Tax_id:511145.

Species assignation module (SA)

After the NER and SNEN modules, each gene entity is assigned the suitable species ID. We applied four species ID assignment rules for Species assignation. The design of these rules was originated from [6]**,** and modified in this module. The detail is as the following:

- Previous species entity: the species ID assigned to a gene entity if the species entity appears front the gene entity.
- Previous species letter: The first letter of the name also can be an abbreviation of its species, like "hZIP 2" when the original name of a human gene was "ZIP2". The gene entities "ZIP2" and "hZIP2" would be assigned the same

species ID.

- Species and gene entities in the same sentence: the species ID assigned to gene entity if the species entity appears in the same sentence of the gene entity.
- Majority voting: the most frequently mention species ID assigned to the gene entity if it cannot be assigned by previous rules.

Fast inferring module for gene name entity normalization module (NEN)

After species assignation, the proposed fast inferring module utilized to calculate the inference scores for candidate EntrezIDs from articles. The design of this method was inspired by our previous work[7]. We used two inference estimations, i.e., the entity inference and the bag-of-word inference to measure the inference confidence scores. The gene name entities are divided to two lists, Entity list and Bag of word list. Entity list is collected by SR output and Bag of word list is collected by all bag of word from Entity list. Each record of these two lists is used to obtain candidate EntrezID (*Cid*). Before inference estimations, the Cid would be filtered by our proposed intersection filtering method. The two inference estimations applied TF-IDF based inference network to determine the possible EntrezIDs for each article.

Acknowledgements

We thank GN organizers and Biocreative III that provide this chance to experiment our method in full text articles. Particularly, we thank Prof. Zhiyong Lu for introduction of the task information. This research was supported by IKMLAB of National Cheng Kung University, Taiwan, ROC.

References

[1] B. Alex, *et al.*, "Assisted curation: does text mining really help?," *Pac Symp Biocomput,* pp. 556-67, 2008.

[2] D. Maglott, *et al.*, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research,* vol. 00, pp. D1-D6, 2006.

[3] H. D. Carroll, *et al.*, "Threshold Average Precision (TAP-k): A Measure of Retrieval Designed for Bioinformatics," *Bioinformatics,* vol. 26, pp. 1708-1713, 2010.

[4] C.-N. Hsu, *et al.*, "Integrating High Dimensional Bi-directional Parsing Models for Gene Mention Tagging," *BIOINFORMATICS,* vol. 24, pp. i286-i294, 2008.

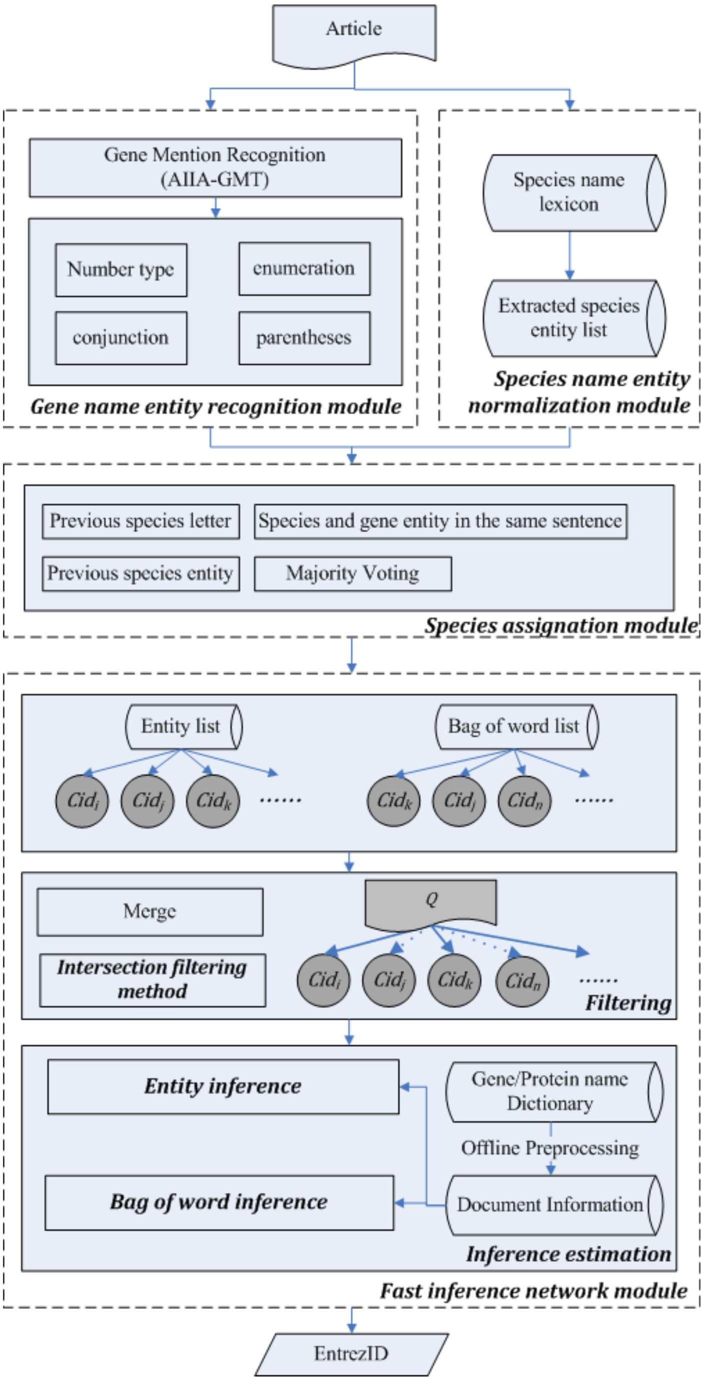
[5] M. Gerner, *et al.*, "LINNAEUS: A species name identification system for biomedical literature," *BMC Bioinformatics*, vol. 11, 2010.

[6] X. Wang*, et al.*, "Disambiguating the Species of Biomedical Named Entities using Natural Language Parsers," *Bioinformatics (Advance Access published),* vol. January 6, 2010.

[7] C.-H. Wei, *et al.*, "Normalizing Biomedical Name Entities by Similarity-Based Inference Network and De-ambiguity Mining," in *Ninth IEEE International Conference on Bioinformatics and Bioengineering Workshop: Semantic Biomedical Computing*, Taichung, Taiwan, 2009, pp. 461-466.

Figures

Figure 1 - Architecture of the gene name normalization method



Tables

Table 1 - Statistics of annotations of full-text articles

Number of total full-text articles	507
Number of annotation released articles	50
Total Gene IDs of articles of annotation released articles	1666
Avg. Gene IDs of articles of annotation released articles	33.32

Table 2. Statistical of the TAP-K

	TAP_5 (%)	TAP_10 (%)	TAP_20 (%)
1st run	31.84	34.69	34.66
2rd run	31.47	33.66	33.66
3nd run	32.28	34.45	34.45

Table 3. Statistical of the F-measure

Total EntrezIDs	TP	FP	FN	Precision(%)	Recall(%)	F-measure(%)
1666	657	563	1009	53.85	39.44	45.53

Improving Protein-Protein Interaction Article Classification Performance by Utilizing Grammatical Relations

Sun Kim[1] and W. John Wilbur*[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Email: Sun Kim - sun.kim@nih.gov; W. John Wilbur*- wilbur@ncbi.nlm.nih.gov;

*Corresponding author

Abstract

Background: Protein-protein interactions (PPIs) are important for understanding fundamental biological processes. However, most of the information still remains in research papers. To capture the hidden PPIs, statistical or machine learning (ML) approaches have been proposed so far. Each approach has pros and cons, and previous work has suggested that there are certain patterns in PPI sentences.

Results: We present a PPI article classification method that automatically learns grammatical patterns on the training corpus, and predicts unknown text based on the data-driven model. More specifically, a dependency parser with gene mention tagging is utilized along with common term-based features. A large margin classifier with Huber loss function is used for the core learning system. The experimental results show that our approach outperforms ML methods using a bag-of-words (BOW) representation. Moreover, the performance changes by selected features are analyzed.

Conclusions: We found that PPI and non-PPI articles can be more easily distinguished by using their grammatical patterns. Also, heuristic knowledge such as gene mention detection can help improve system performance with a limited training corpus. The proposed method stands out among ML-based methods because it shows a way of using grammatical relations for PPI article filtering, not words- nor fixed rule-matching only.

Table 1: The corpus information used in our experiments.

Corpus Name	Positive Examples	Negative Examples	Total Examples
BioCreative II	3874	2298	6172
BioCreative II.5	124	1066	1190
BioCreative III Training Set	1140	1140	2280
Total Training Set	5138	4504	9642
BioCreative III Development Set	682	3318	4000

Background

A plethora of biomedical literature that describes protein-protein interaction experiments by specifying individual interacting proteins and the corresponding interaction types exists. While many efforts have been made to create protein interaction databases such as MINT, IntAct, and DIP, several constraints such as the problems of manual curation of a database, the rapid growth of biomedical literature, and newly discovered proteins make it difficult for database curators to keep up with the published information [1].

Among various approaches to mine protein-protein interaction (PPI) information, machine learning (ML) techniques have gained popularity in recent years. In contrast to rule-based approaches, ML methods can discover new patterns not captured in a known trigger word list. Several natural language processing (NLP) approaches also have been proposed for PPI extraction [2–5], where PPI sentences are assumed to have unique grammatical structures. However, the effectiveness of using parsing information has been hardly investigated at the article classification level.

In this paper, we present a PPI article filtering method, which combines NLP strategies with ML techniques. Our approach uses a dependency parser [6] and a gene mention detection method [7] to extract additional features along with the word-based feature set. A large margin classifier with Huber loss function [8] is used to learn a selected feature set from training data. When applied to BioCreative corpora, our method outperforms ML approaches using the bag-of-words (BOW) representation. In addition, we explore the automatic induction approach for high-order features [9]. According to the experimental results, we found that the performance of PPI article classification can be improved by utilizing grammar relations between words. Gene mention tagging can be used to decrease the data sparseness problem and also increase classification performance.

Methods

Dataset

The training data used in the present work is based on the examples of all BioCreative PPI article classification tasks (Table 1). BioCreative II (6,172 abstracts), Biocreative II.5 (1,190 abstracts), and BioCreative III training data (2,280 abstracts) were combined for training. The BioCreative III development set was used for testing. As a result, the final training set consists of 5,138 positive and 4,504 negative examples. The development set includes 682 positive and 3,318 negative examples.

PPI Article Filtering Procedure

Figure 1 depicts the overview of the proposed method. Input articles are first evaluated whether there are gene/protein names in text, where a gene mention tagger trained with SemCat [7] and Entrez Gene data is used. After gene mention detection, feature generation is performed in three different ways. One is word-based features including multi-words, strings, and MeSH terms. The second is relation-based features,

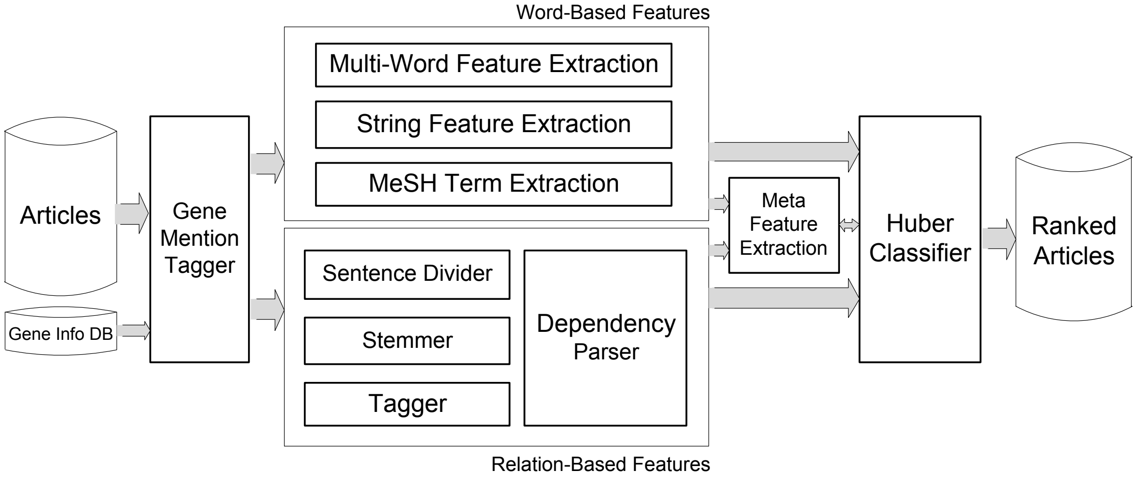


Figure 1: Overview of the proposed PPI article classification method.

which extracts grammar relations between words. The third is to extract meta-information by evaluating a combination of different features.

After all feature extraction procedures, a large margin classifier with Huber loss function [8] is utilized for learning and classifying given examples. The Huber classifier learns examples by minimizing modified Huber loss L, where given p and y, $L(p, y) = \max(0, 1 - py)^2$ when $py \geq -1$ and $-4py$ otherwise. p is the system output and y is the label of a data point $\mathbf{x}$. A gradient descent method is used for optimizing the classifier. Output of the classifier is the score indicating how likely an article would contain PPI information. More detail regarding the features is explained in the following subsection.

Feature Set

Feature generation is the most important part when it comes to machine learning real-world problems. In this paper, we investigate six feature sets added to a simple BOW representation. 'Multi-word features' are bigrams and trigrams of words. 'String features' are strings with n characters. Huang *et al.* [10] suggested this feature set to reduce the difference between distributions on training and test sets. In the experiments, four through seven characters were tested as string features, and 6-consecutive characters produced the best classification performance for the current dataset. 'MeSH terms' are also considered as one candidate feature set. MeSH is a thesaurus for indexing and searching biomedical literature, hence this controlled vocabulary set might be helpful for PPI article detection.

'Grammar relation features' indicate dependency relationships between words. Since we detect gene/protein names beforehand, each gene or protein name can be handled as a word. Our assumption here is that PPI information can be revealed by analyzing word-to-word relationships. The C&C CCG parser [6] is used to obtain dependency relations in the proposed method. 'Gene tagging' is a further step in utilizing grammar relation features. The purpose of PPI article classification is to identify whether an article contains PPI information, not a gene/protein name itself. Therefore, in a dependency relationship, the particular protein named is not important. The gene tagging strategy simply exchanges a detected gene/protein word for a special tag, called 'PTNWORD'. By doing this, the complexity of relationship features is decreased, while

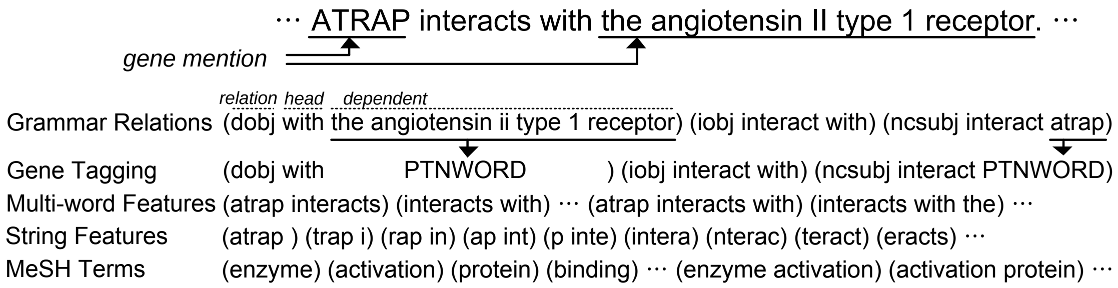

··· ATRAP interacts with the angiotensin II type 1 receptor. ···
gene mention
relation head dependent
Grammar Relations (dobj with the angiotensin ii type 1 receptor) (iobj interact with) (ncsubj interact atrap)
Gene Tagging (dobj with PTNWORD) (iobj interact with) (ncsubj interact PTNWORD)
Multi-word Features (atrap interacts) (interacts with) ··· (atrap interacts with) (interacts with the) ···
String Features (atrap) (trap i) (rap in) (ap int) (p inte) (intera) (nterac) (teract) (eracts) ···
MeSH Terms (enzyme) (activation) (protein) (binding) ··· (enzyme activation) (activation protein) ···

Figure 2: An example of feature generation. Feature values are extracted based on gene mention detection. Grammar relations and gene tagging are relation-based features. Multi-word, string, and MeSH term features are word-based features.

Table 2: Average precision rate for Naïve Bayes, SVM and Huber approaches. The best score is obtained when using both BOW (bag-of-word representation) and GR (grammar relations).

Feature Set	Naïve Bayes	SVM	Huber
BOW	0.616935	0.659952	0.664582
GR	0.628117	0.639090	0.641717
BOW + GR	**0.653827**	**0.672551**	**0.677123**

the relationship information remains the same. Figure 2 shows an example sentence and its word-based and relation-based feature sets used in our approach.

'Meta features' are higher-order features automatically extracted by evaluating a set of feature combinations. When system prediction is incorrect, each feature combination is evaluated by a sum of partial derivatives of loss function terms on data points [9]. For experiments, bigram meta features were evaluated, and the top-scoring 784 bigrams were used as meta-features in the Huber classifier.

Results

Table 2 shows the average precision rates for the BioCreative III (BC3) development set when single words and their dependency relationships are used. 'BOW' means unigram features. 'GR' means word-relationship features. All classifiers were optimized for giving the best scores on both training and development sets. 'SVM' and 'Huber' are the support vector machine classifier with linear kernel and the proposed large margin classifier, respectively. As shown in the table, adding word dependencies to single word features boosts up the performance by 3.7% in naïve Bayes classifiers. However, our Huber approach produces the best average precision overall.

For the BioCreative ACT task, possible feature candidates were tested and analyzed. As a result, five feature sets were further selected for better classification. Table 3 presents the performance changes by adding those five feature types, gene tagging, multi-word features, string features, MeSH terms, and meta features. A row shows the evaluation results when all of its above features and the feature presented in the row are used. Our system is originally designed to give ranked results, rather than labels. However, the

Table 3: Performance changes by adding additional features. All features contribute to the performance in some way. However, using gene tagging and MeSH terms are more effective than using other features. 'Word features' are bigrams and trigrams of words. 'String features' are strings with six characters. 'Meta features' are automatically induced meta bi-gram features.

Used Features	Avg. Prec.	Precision	Recall	F1
Baseline (BOW + GR)	0.677123	0.594005	0.639296	0.615819
+ Gene Tagging	0.689493	0.586842	0.653959	0.618585
+ Multi-word Features	0.694726	0.616379	0.629032	0.622641
+ String Features	0.696727	**0.633577**	0.636364	0.634967
+ MeSH Terms	**0.707015**	0.620739	0.640762	0.630592
+ Meta Features	0.704403	0.632737	**0.674487**	**0.652945**

system output can be binarized by using signs of the Huber classifier output. Precision, recall, and F1 scores in the table were evaluated on these binarized results. While all features contribute to the performance in some way, the biggest improvement is made when gene tagging and MeSH term features are introduced. In particular, 'Meta features' does not improve the average precision, but it increases F1 score to 65.29%.

The proposed approach produced good scores for training and other balanced data sets, whereas it was less successful for imbalanced data sets such as the BC3 development and test sets. Our ACT system on the BC3 test set obtained 89.15% accuracy and 61.32% F1 score, where only some of the feature types in Tables 2 and 3 were used, i.e., grammar relations with gene tagging, bigrams of words, string features, and MeSH terms.

Conclusions

We presented a machine learning approach combined with natural language processing techniques to classify protein-protein interaction articles. A large margin classifier with Huber loss function is used to identify PPI-relevant abstracts. The experimental results show that grammar relations between words help improve the classification performance, and other features such as gene mention tagging, MeSH terms, and automatic meta-feature selection can be effective for the PPI article filtering task. However, the imbalanced nature of PPI vs. non-PPI articles in PubMed makes this problem much harder. Hence, exploring additional feature sets and controlling the number of positive/negative examples for better classification remain as future work.

Acknowledgements

The authors would like to thank Won Kim and Larry Smith for valuable comments on implementing the proposed method. The authors are supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Blaschke C, Leon EA, Krallinger M, Valencia A: **Evaluation of BioCreAtIvE assessment of task 2**. *BMC Bioinformatics* 2005, **6(Suppl 1)**:S16.
2. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I: **Extracting human protein interactions from MEDLINE using a full-sentence parser**. *Bioinformatics* 2004, **20**:604–611.
3. Jang H, Lim J, Lim JH, Park SJ, Lee KC, Park SH: **Finding the evidence for protein-protein interactions from PubMed abstracts**. *Bioinformatics* 2006, **22**:e220–e226.

4. Kim S, Shin SY, Lee IH, Kim SJ, Sriram R, Zhang BT: **PIE: an online prediction system for protein-protein interactions from text**. *Nucleic Acids Research* 2008, **36(Suppl 2)**:W411–W415.

5. Miyao Y, Sagae K, Sætre R, Matsuzaki T, Tsujii J: **Evaluating contributions of natural language parsers to protein-protein interaction extraction**. *Bioinformatics* 2009, **25**:394–400.

6. Curran JR, Clark S, Bos J: **Linguistically motivated large-scale NLP with C&C and Boxer**. In *Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo): 23-30 June 2007; Prague* 2007:33–36.

7. Tanabe L, Wilbur, J W: **A priority model for named entities**. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology: 4-9 June 2006; New York* 2006:33–40.

8. Zhang T: **Solving large scale linear prediction problems using stochastic gradient descent algorithms**. In *Proceedings of the 21st International Conference on Machine Learning: 4-8 July 2004; Banff* 2004:919–926.

9. Ando RK: **BioCreative II gene mention tagging system at IBM Watson**. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop: 23-25 April 2007; Madrid* 2007:101–103.

10. Huang M, Ding S, Wang H, Zhu X: **Mining physical protein-protein interactions from the literature**. *Genome Biology* 2008, **9(Suppl 2)**:S12.

Gene normalization as a problem of information retrieval

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

[1]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[2]School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore
[3]Department of Zoology, The University of Melbourne, Parkville, Victoria, Australia
[4]Information Sciences Institute, University of Southern California, Marina del Rey, California, USA

Email: Cheng-Ju Kuo - clarkkuo@iis.sinica.edu.tw; Maurice HT Ling - mauriceling@acm.org; Chun-Nan Hsu*- chunnan@isi.edu;

*Corresponding author

The goal of gene normalization (GN) task is to link the names of gene or gene products mentioned in the literature to standard database identifiers [1]. In the BioCreative II GN task, we realized that finding out gene mention candidates as many as possible may be a key to implement a GN system with high recall. We propose to use bi-directional parsing models of Conditional Random Fields (CRFs) to tag gene mentions (any possible composition of tokens likely to be a gene mention) in a region of text. It is useful to alleviate the requirement to manually create rule sets which may change with time. The GM module [2], which was ranked second among twenty-one participants in BioCreative II challenge, was built on BioCreative II gene mention tagging (GM) training corpus. In order to finding all candidates of gene mentions, we collected twenty prediction sequences from two parsing models and transform each sequence to one set of gene mentions from a sentence. That is, there are at most fourty sets of gene mentions that might be produced from a single input sentence. We merged them into one set, tested on BioCreative II GM test corpus and achieved a recall of 0.9419 in the internal test. The resulting fourty sets of gene mentions were merged as input to the GN system.

We used BIOADI [3] to identify all pairs of abbreviation and its long form in the article. If the long form of an abbreviation is not tagged as a gene mention in the sentence where the abbreviation pair is extracted, the abbreviation is marked as an invalid gene mention, then has to be removed from gene mentions extracted in GM step. We took advantage of contextual information among sentences to remove mentions which are not referred to gene or gene products. For example, ”NAA” is a candidate gene mention tagged from a text of “...metabolites in 5 patients: N-acetyl-aspartate (NAA), creatine...”. We can link the definition of “NAA” to “N-acetyl-asparate” and know that “N-acetyl-asparate” is not been tagged as a gene mention, we can ignore all ”NAA”s tagged in the sentences of the same article. It is useful to improve

GM performance by reducing false-positives.

The rest of gene mentions will be resolved into its species and assigned a taxonomy identifier (taxid) of NCBI Taxonomy database. We combined the Taxonomy database and LINNAEUS species dictionary [4] to a taxid-name dictionary. We used the dictionary to resolve species word to unique taxonomy identifier of Taxonomy database by applying a heuristic approach described in [5]. There are three rules for species assignment. Firstly, assign the nearset taxid which is precending the gene mention. Secondly, assign the taxid which locates in the same sentence. Lastly, assign the taxid which has the highest occurrence of the literature.

After taxid assignment, we obtain a list of gene mentions with its taxid of each input sentence which will be used to query Entrez Gene database to know whether there is a similar text in Entrez Gene. If no record is found, the gene mention will be removed from the list. This step can remove most of non-gene mention records. As a rule, we choose the longest gene mention of any region where gene mentions were found by the GM system. This can avoid mention overlapping which might reduce the precision of GN module. After the process, we took an expensive matching process to resolve each gene mention for its Entrez Gene identifier candidates. We applied Lucene, with a customized tokenizer and analyzer, to speed up the matching process. Each name of retrieved Entrez Gene record will be compiled to a regular expression pattern for fuzzy match to the query text of gene mention. If the gene mention matches the pattern, the gene mention is assigned to the identifier. If more than one identifier are assigned to a gene mention, we set a heuristic rule to select one as the output.

We propose a system that can evaluate the quality of gene mention. Each GN result is assigned a confidence score. The system is built based on logistic regression and trained on BioCreative II GM training and test data. Each GN result will be pass to the model to evaluate the quality of its gene mention. The logisitic regression model will return a decision value between 0 and 1. We directly used the output value as a significance for the GN result. In the training stage, we achieved a precision of 0.643 with 0.588 recall (F-score = 0.614) on BioCreative III GN 32 full annotated articles. TAP-5, TAP-10 and TAP-15 scores on the training articles are shown in Table 1. Table 2 is the results for three submitted runs on the most difficult 50 articles selected by the task organizers.

References

1. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu Hh, Torres R, Krauthammer M, Lau W, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L: **Overview of**

Table 1: TAP-5, -10 and -20 on 32 training articles in inside test

Run	Unweighted Average TAP		
	TAP-5	TAP-10	TAP-20
1	0.3123	0.4151	0.4151

Table 2: TAP-5, -10 and -20 of three submitted runs on 50 test articles

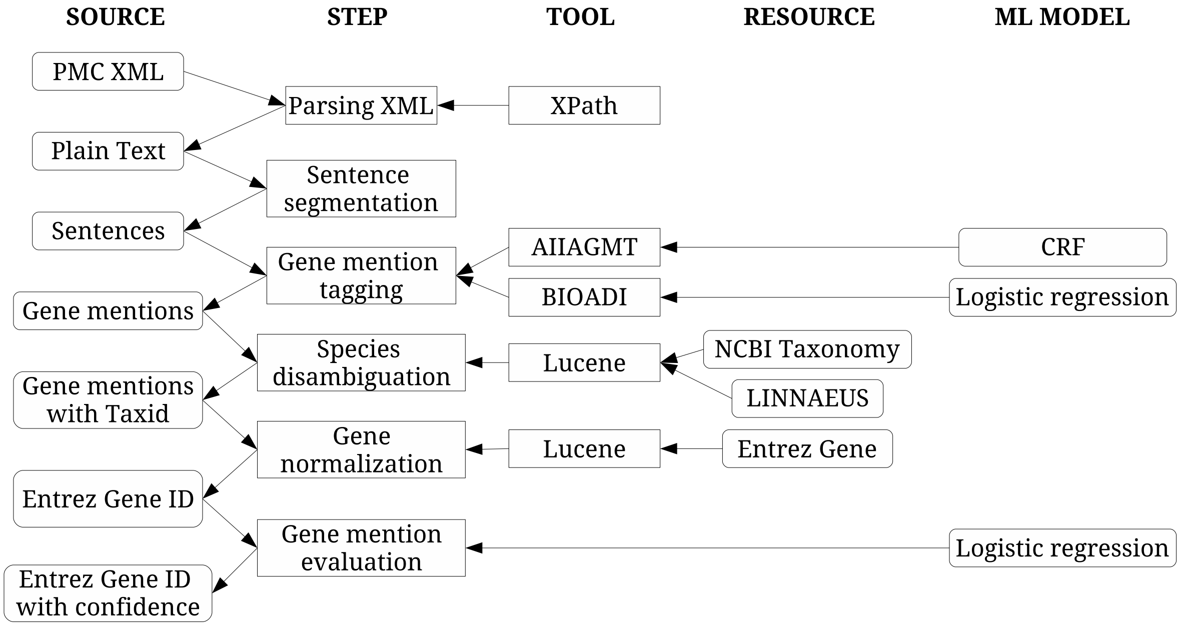
Run	Unweighted Average TAP		
	TAP-5	TAP-10	TAP-20
1	0.2099	0.2447	0.2447
2	0.2048	0.2420	0.2420
3	0.2061	0.2432	0.2432

BioCreative II gene normalization. *Genome Biology* 2008, **9**(Suppl 2):S3, [http://genomebiology.com/2008/9/S2/S3].

2. Hsu CN, Chang YM, Kuo CJ, Lin YS, Huang HS, Chung IF: **Integrating high dimensional bi-directional parsing models for gene mention tagging**. *Bioinformatics* 2008, **24**(13):i286–i294.
3. Kuo CJ, Ling M, Lin KT, Hsu CN: **BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature**. *BMC Bioinformatics* 2009, **10**(Suppl 15):S7, [http://www.biomedcentral.com/1471-2105/10/S15/S7].
4. Gerner M, Nenadic G, Bergman C: **LINNAEUS: A species name identification system for biomedical literature**. *BMC Bioinformatics* 2010, **11**:85, [http://www.biomedcentral.com/1471-2105/11/85].
5. Wang X, Tsujii J, Ananiadou S: **Disambiguating the species of biomedical named entities using natural language parsers**. *Bioinformatics* 2010, **26**(5):661–667, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/5/661].

Figures

Figure 1 - System flowchart



IISR Gene Normalization System for BioCreAtIvE III

Po-Ting Lai [1], Hong-Jie Dai [2,3], Chi-Hsin Huang [3], Richard Tzong-Han Tsai [1*]

[1] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan

[2] Department of Computer Science, National Tsing-Hua University, Hsinchu

[3] Institute of Information Science, Academia Sinica, Taipei

[*]Corresponding author

Email addresses:

PTL: s951416@mail.yzu.edu.tw

HJD: hongjie@iis.sinica.edu.tw

CHH: sinyuhgs@iis.sinica.edu.tw

RTHT: thtsai@saturn.yzu.edu.tw

Background

The goal of Gene Normalization (GN) is to link gene or gene products mentioned in the literature to standard database identifiers. Three main subtasks are involved in the GN task: gene mention recognition (GMR), dictionary matching, and disambiguation processing. The Intelligent Information Services Research (IISR) team (team identifier: #101) employed several machine learning (ML) techniques and natural language processing (NLP) techniques to deal with the three subtasks.

Methods

GMR is handled by two taggers. The first is a conditional random field (CRF)-based gene mention tagger, NERBio [1], which is trained on the BioCreAtIvE II gene mention dataset [2] with a set of features selected by a sequential forward search algorithm. The GMR problem is formulated as a word-by-word sequence labeling task, where the assigned tags delimit the boundaries of any gene names. The second gene mention tagger is a rule-based gene mention tagger that is developed to recognize genes mentioned in regular formats. For example, locus_tags are identifiers that are systematically applied to every gene in a genome. These tags have become surrogate gene names by the biological community. The prefix of a locus_tag should start with a letter and can contain only alpha-numeric characters. It must be at least three characters long. Numerals can be in the second position or later in the string (e.g. A1C.) The rule-based tagger use regular expressions to recognize these terms. After GMR, we employ several post-processing rules to identify more gene mentions [3]. For instance, if a parenthesized phrase follows an identified gene mention, we also regard the contents of the parentheses as a gene mention. The recognized gene names are finally examined against a blacklist to filter out false positives.

Dictionary-matching is able to assign candidate identifiers to each recognized gene mention. We use a lexicon compiled from collected gene names from Entrezgene and their orthographical variants. We collect gene names from the following fields: “official symbol”, “official full name”, “all also known as”, “names of general protein information”, and “prefer names of general protein information.” In addition, we have observed that, the identifiers of the 22 common organisms recorded in the NCBI Entrez Taxonomy account for 91.44% of those in the training test. Therefore, we only compiled gene identifiers and corresponding names from those organisms. Each recognized gene mention is looked up in the lexicon. If a gene mention is assigned two or more gene identifiers, we must determine which is more appropriate through disambiguation processing.

We have constructed several rule-based classifiers which use context information, such as chromosome location, sequence length and so on, to determine the given identifier’s label. Each classification rule consists of a conjunction of attribute tests and a weighted score. The final disambiguation process is based on the linear combination of the weighted scores of the various classifiers’ predictions.

The dataset of BioCreAtIvE III GN is assembled by full-length articles. Several studies have shown that scientific authors do, in the majority of cases, follow the basic principles of the research article structure and assign information accurately to each section. Each section has different characteristics which we can use to guide GN. We employed a multi-stage processing framework developed by the IASL-IISR interactor normalization task (INT) system [4] to process the articles. The following sections are separately processed by our GN system: title, abstract, introduction/background, methods, materials, result, discussion, conclusion, figure/table captions, abbreviation definitions, and gene names described in the tables.

The final step is ranking all normalized identifiers in a paper. We formulate the ranking problem as a classification problem, and incorporate the confidence of the normalized identifiers and context information as features. For an article, the normalized identifiers that match/miss the gold standard identifiers are treated as true positive (TP) and false negative (TN) instances respectively. However, training set 1 (32 articles), which contains full annotated identifiers, contains only limited numbers of TP and TN instances. To increase the generality of our classification model, we use the training set 2 (493 articles), released by BioCreAtIvE III GN challenge, as another resource to bootstrap our model. We employ our GN system to process training set 2. The identifiers that match the gold answers are extracted as TP instances. Ambiguous partners of each extracted identifier are treated as TN instances.

Table 1 summarizes all resources used by our system.

References

1. Tsai RT-H, Sung C-L, Dai H-J, Hung H-C, Sung T-Y, Hsu W-L: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition**. *BMC Bioinformatics* 2006, **7**(Suppl 5):S11.
2. Smith L, Tanabe LK, Ando RJn, Kuo C-J, Chung I-F, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K *et al*: **Overview of BioCreative II gene mention recognition**. *Genome Biology* 2008, **9**(Suppl 2):S2.
3. Lai P-T, Bow Y-Y, Huang C-H, Dai H-J, Tsai RT-H, Hsu W-L: **Using Contextual Information to Clarify Gene Normalization Ambiguity**. In: *The IEEE International Conference on Information Reuse and Integration (IEEE IRI 2009).* Las Vegas, USA; 2009: 1-5.

4. Dai H-J, Lai P-T, Tsai RT-H: **Multi-stage gene normalization and SVM-based ranking for protein interactor extraction in full-text articles**. *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010, **7**(3):412-420.

Tables

Table 1 - Resources used by Team 101.

Team Identifier	101
Machine Learning Technique	Maximum entropy Conditional random field Support vector machine
NLP Components	GeniaTagger NERBio IASL-IISR Multi-stage Processing Framework
External Lexical Resources	NCBI Entrezgene set NCBI Taxonomy UniProt Cell Bank HyperCLDB Invitrogen HPRD
Relevant Resources	BioCreAtIvE II Gene Mention corpus

A top-down approach for finding interaction detection methods

Robert Leaman*[1], Ryan Sullivan[2] and Graciela Gonzalez[2]

[1]School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, Arizona, USA
[2]Department of Biomedical Informatics, Arizona State University, Tempe, Arizona, USA

Email: Robert Leaman*- robert.leaman@asu.edu; Ryan Sullivan - rpsulli@asu.edu; Graciela Gonzalez - graciela.gonzalez@asu.edu;

*Corresponding author

Abstract

Background: The experimental methods used to detect protein interactions are an essential part of the evidence for the interaction. The BioCreative 3 PPI-IMT task consists of determining which interaction detection methods were used to detect protein interactions in a given full text article.

Results: We created a machine-learning based system where each interaction detection method was modeled by one classifier. The classifiers were trained and applied at the document level to provide the overall determination of whether a given method was used in the document. However, we also apply the same classifiers at the sentence level to determine which sentence from the document best supports the interaction detection method. The highest macro-averaged f-measure we achieved in our cross-validation experiments was 0.6278, and the highest AUC iP/R we achieved was 0.6217.

Background

The experimental methods used to verify protein interactions are an important factor in weighing their reliability [1]. Determining the interaction detection methods used in an article is a term recognition task that is therefore of strong interest to database curators. The BioCreative 3 PPI-IMT task asks the biomedical text mining community to address this need. The task organizers provided a training set consisting of 2,035 full-text articles and a development set consisting of 587 full-text articles, all annotated at the document level with interaction detection method concepts from the PSI-MI ontology [2]. In this paper we describe our machine learning-based system for determining the interaction detection methods used within a document.

Methods

Our initial analysis of the IMT data suggested that a lexical approach would be difficult for several reasons. First, interaction detection methods are often implied rather than mentioned directly. Second, even when the method is mentioned, the mention is usually not sufficiently specific to determine the exact method used. However, we noted some correlations between interaction detection methods and words that are unrelated to any of the names for the method. For example, “ultracentrifuge” is a strong indicator for MI:0028, “cosedimentation in solution.” We also found that the average number of times each interaction detection method appears in the training data is about 50, implying that if an interaction detection method appeared at all, there were typically several examples of it.

We therefore hypothesized that this task might be productively modeled as a document-level classification problem. Given the availability of sufficient document-level training data, we decided to implement and evaluate a machine learning approach for finding interaction detection methods. As our method is based on document-level classification, this approach is inspired more by techniques for problems such as topic classification than by standard term and named entity recognition.

Document preprocessing

For our initial input, we utilized the text format distributed by the organizers, rather than extracting text from the PDF files ourselves. We broke each document into sentences using the Java sentence breaker, and each sentence was then tokenized by splitting at whitespace and punctuation. Each token was converted to lower case and Unicode characters such as ligatures were normalized. Stop words were then removed [3] and the remaining tokens were stemmed using the Snowball implementation of the Porter2 stemmer [4].

Machine learning setup

Our system creates one classification model for each interaction detection method. Each document annotated with the given method is considered a positive instance of that method, and all other documents are considered negative instances of the method. Each interaction detection method is modeled on its own, without regard for any subtypes or supertypes of the method.

We use the same feature set for all classifiers, which consists of two feature types. One feature type is binary-valued and indicates the presence or absence of a single stemmed token. The other feature type indicates the presence or absence of a name from a single interaction detection method concept. This type comes in two variants: strict and fuzzy. Strict lexicon membership features are binary-valued and will be true if there is a sentence in the document that contains all of the tokens in any of the names of the detection method being located. Fuzzy lexicon membership features are similar to strict lexicon membership except that they are real-valued and represent the proportion of the tokens of the interaction detection method name that the sentence contains.

The lexicon used for the lexicon membership features consisted primarily of the name, synonyms and unique identifiers (e.g. “MI:0006”) from the PSI-MI ontology. We added additional synonyms by locating concepts in the UMLS Metathesaurus [5] of the semantic types listed in table 1 which share a name with a concept in the PSI-MI ontology. Once a synonymous concept was found, we added all of the names for the UMLS concept as synonyms for the PSI-MI concept. Each name in the lexicon was preprocessed in the same manner as the document text prior to use (i.e. stop words removed from each name, tokens converted to lower case and stemmed).

Classification and training

We chose logistic regression as the classification model since it tends to give reasonable probability estimates. The classifiers were trained using L_1 regularization, which has been the subject of considerable interest in recent years, largely due to the tendency it has of setting the weight of most parameters to 0. This is in contrast to L_2 regularization, which usually learns many weights that approach 0 asymptotically. Setting most feature weights to 0 has the same effect as feature selection, and results in models that are more compact, more interpretable, and execute faster. They should also be more robust to irrelevant features, since the amount of training data needed rises only logarithmically in the number of irrelevant features provided [6]. The standard L-BFGS optimization cannot handle weights with a zero value, however, and thus cannot be used to train an L_1 regularized model. Instead, we use the orthant-wise limited memory

Table 1: UMLS Semantic Types searched for synonyms to concepts in the PSI-MI ontology, along with their unique identifier (TUI) and the number of terms of that type in the final lexicon.

TUI	Semantic Type Name	Count
T059	Laboratory Procedure	23
T070	Natural Phenomenon or Process	5
T116	Amino Acid, Peptide, or Protein	4
T060	Diagnostic Procedure	4
T063	Molecular Biology Research Technique	2
T067	Phenomenon or Process	1
T121	Pharmacologic Substance	1
T074	Medical Device	1
T169	Functional Concept	1

quasi-Newton algorithm (OWL-QT) [7]. Our system uses the MALLET implementation of both logistic regression and L_1 regularized training with OWL-QT [8].

System output

We use the probability output by each classifier as the confidence for the associated interaction detection method in the system output. This allows for easy tuning towards higher recall or higher precision by thresholding the results. In our submitted runs, we chose the thresholds to approximate the maximum possible value for the following measurements:

- Area under the interpolated precision / recall curve (AUC iP/R)
- F-measure, balanced
- F-measure $\beta = 0.5$, meaning that precision is weighted twice as heavily as recall

To find support statements, we used the trained classifiers by applying them to each sentence in the document. The sentence from the document with the highest probability output by the classifier is used as support for the corresponding interaction detection method. We found that while these classifiers were trained at the document level, applying them at the sentence level still resulted in inferences of reasonable quality. Objectively quantifying their accuracy was not possible, however, due to a lack of data.

System Variants

In addition to the system described to this point, which we shall call version 1, we evaluated two additional versions. Version 2 uses L_2 regularization rather than L_1 regularization. Version 3 replaces the binary features representing the presence or absence of individual tokens with the TF-IDF values for the token, normalized so that the TF-IDF values for the document sum to one.

Evaluation

We used a variation of 10-fold cross validation for evaluation. Since the organizers stated that the test set would more closely match the development set than the training set, we did not use any part of the training set for evaluation. Instead, we split the development set into 10 folds, and then repeatedly trained on the entire training set plus 9 folds of the development set, and evaluated on the remaining fold from the development set. The results were then averaged across all 10 folds.

Figure 1: Precision vs. Recall curves for system versions 1, 2, and 3

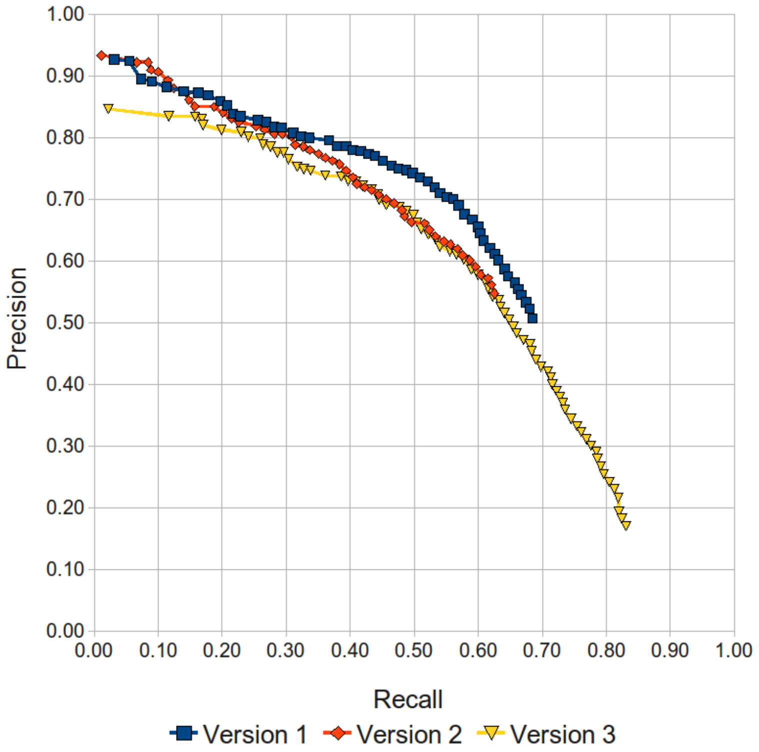


Table 2: Results of the cross-validation experiments. Each value is macro-averaged and represents the maximum value achievable by selecting a cutoff threshold.

Version	**AUC iP/R**	**F-measure (balanced)**	**F-measure $\beta = 0.5$**
1	0.6217	0.6278	0.6783
2	0.5730	0.5946	0.6349
3	0.6684	0.4548	0.3666

Results and Discussion

The results of the cross validation experiments can be seen in figure 1 and in table 2. Figure 1 plots the points for the macro-averaged and interpolated precision and recall of all three system versions. Table 2 lists the maximum performance each system version could achieve by choosing and applying a threshold to the output.

We see that system version 1 outperforms version 2 on all measurements and at nearly all points on the precision/recall curve. Since these versions are the same except for the choice of L_1 regularization or L_2 regularization, this result is empirical evidence that L_1 regularization returns a better classifier. We also see that version 1 outperforms version 3 at all points on the precision/recall curve except that the highest recall version 1 achieves is just under 0.69, while the highest recall for version 3 is over 0.83. We conclude that the binary representation has better overall performance, but that TF-IDF representation is better at emphasizing high recall.

Conclusions

The highest macro-averaged f-measure we achieved in our cross-validation experiments was 0.6278, and the highest AUC iP/R we achieved was 0.6217. While the results are not directly comparable since different datasets were used, we note that the best f-measure reported at the BioCreative 2 PPI-IMS task was 0.4836 [1]. We believe that document-level classification is a promising approach for finding interaction detection methods and that it may be applicable to other term recognition problems where sufficient document-level training data is available.

In future work we intend to explore the improvements to be gained by feature combinations, for example, using the presence or absence of a pairs of tokens as a feature. We also intend to make better use of the interaction method hierarchy by creating classifiers trained to determine the presence of an interaction detection method or any of its children, rather than looking for a single method as we have done here.

Authors contributions

RL analyzed the data, designed the system, performed implementation and drafted the paper. RS performed literature search and assisted with system implementation. GG supervised the project and assisted with system design.

Acknowledgements

The authors would like to thank Annie Skariah for her assistance with data analysis.

References

1. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II**. *Genome Biology* 2008, **9**(Suppl 2):S4.

2. **Proteomics Standards Initiative - Molecular Interaction ontology** [http://psidev.sourceforge.net/mi/rel25/data/psi-mi25.obo].

3. The Information Retrieval Group University of Glasgow: **Stop Words** [http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words].

4. **Snowball** [http://snowball.tartarus.org].

5. National Library of Medicine: *Unified Medical Language System (UMLS) Knowledge Sources, revision 2009AA*. Bethesda, Maryland, USA 2009.

6. Ng AY: **Feature selection, L1 vs. L2 regularization, and rotational invariance**. In *Proceedings of the 21st International Conference on Machine Learning* 2004:78–85.

7. Andrew G, Gao J: **Scalable training of L1-regularized log-linear models**. In *Proceedings of the 24th International Conference on Machine Learning* 2007:40–47.

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

Species-specific Gene Normalization – is it feasible?

Hongfang Liu[1]*, Manabu Torii[2], and Zhangzhi Hu[3]
Lab of Text Intelligence in Biomedicine
[1] Department of Biochemistry, Cell and Molecular Biology,
[2]ISIS Center, [3] Department of Oncology
Georgetown University Medical Center

Abstract

Identification of gene and protein names in literature and their mapping to corresponding gene/protein database records are critical for biomedical literature mining applications. In this paper, we report our gene/protein name normalization system in BioCreAtive III. We become aware of inherent challenge in the current normalization task and propose to define a sense inventory for the normalization task in the future.

Background

One crucial requirement for biomedical literature mining applications is the ability to identify gene/protein entities discussed in the text. [1-4] In general, this task can be divided into several steps: 1) identifying gene/protein mentions in the text, ii) associating the mentions to one or more potential gene/protein database records, iii) selecting the correct record in case of ambiguity, and iv) assembling the final list of genes/proteins in the document[3, 5]. In the first and second BioCreAtive workshops, the gene/protein normalization task was defined and evaluated for single species (i.e., yeast, mouse, and fly for the first workshop and human for the second workshop). Generally, the first step of identifying gene/protein mentions in text can be classified into two groups: i) matching against a gene/protein terminology resource [6, 7], and ii) using a rule-based or machine learning name tagger. After associating gene/protein mentions to potential gene/protein database records, the methods to select the correct record fell into two categories: i) pruning the gene/protein terminology resource by removing ambiguous name entries, and ii) performing word sense disambiguation. Most systems assign confidence scores in this step and set a threshold to derive the final list of genes/proteins. Machine learning classifiers such as Support Vector Machine or Maximum Entropy methods can be used to remove likely false positives.

For the first BioCreAtive workshop, we used a flexible dictionary-lookup method where the dictionary consists of synonyms obtained from online resources of genes/proteins. The system achieved the best recalls among the participating systems for yeast and mouse but the precisions were very low. We found that using an extensive list of synonyms could improve recall, while word sense disambiguation would be critical to improve the precision as also indicated by Hirschman [5].

For the second BioCreAtive workshop, we assembled a dictionary of gene/protein synonyms from online resources, such as BioThesaurus and HUGO, conducted flexible dictionary lookup, and obtained a list of mapping pairs (Term, GeneID), where Term is a term in text and GeneID is the gene identifier. We then applied machine learning to classify each mapping pair as positive or negative, where positives were those considered as appropriate mappings.

The gene normalization task in BioCreAtive III is different from the previous workshops in that full-length articles, instead abstracts, are used and no species information is provided. This setting is rather practical for model organism database curation.

Implementation

Our gene/protein name normalization system is based on the existing systems and resources developed previously in this domain. The following describes the normalization process we employed.

Preprocessing – We extracted sentences in full-length articles that were dynamically retrieved from PubMed Central (PMC) according to article identifiers. For each sentence, gene/protein mentions were detected using BioTagger-GM[8a], BANNER[9], and ABNER[10].

Taxonomy assignment –We sampled a collection of sentences from GeneRIF using the following criteria: i) at most one sentence per gene, ii) at most 5000 sentences per taxonomy identifier for those among the top 30 species ranked according to the number of genes with GeneRIF records. An SVM classifier was then constructed and used to assign taxonomy identifier for each sentence.

Gene normalization – We applied the same dictionary lookup method as the one we used in the previous BioCreAtive workshops. The latest release of gene-centric BioThesaurus[b] was used as the dictionary. For each pair (Term, GeneID), we derived a descriptive feature vector to represent i) ambiguity and systematic ambiguity features of Term based on onto-BioThesaurus and GeneRIF, ii) document-level taxonomy assignment counts, iii) counts of GeneID in the document, iv) number of synonyms representing GeneID, and v) whether Term detected by gene mention systems or not. Failing to obtain acceptable performance using the similar machine learning approach as the one used for gene normalization in BioCreAtive II, we assigned weight to each feature type and ranked each pair (Term, GeneID) according to the sum of the weight of the present features. Heuristic rules were used to avoid the return of too many homologous genes/proteins in the result.

Discussion and Conclusions

The current normalization task is very challenging due to the high systematic ambiguity associated with gene/protein names. From a practical point of view, however, we feel that mapping gene/protein names in text to the species-specific gene/protein database records may not be a well-defined task to suit the needs of biologists, who often describe or glean biological knowledge about genes/proteins in text without much emphasis on the species information. Thus, a sense inventory defining the meanings of gene/protein names in text is urgently needed.

[a] http://biomine.dbb.georgetown.edu/BioTagger

[b] http://biomine.dbb.georgetown.edu/ontoBioT

Acknowledgement

The project was supported by ABI: 0845523 from National Science Foundation and R01LM009959 from National Institute of Health.

Reference

1. Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform.* Dec 2004;37(6):512-526.

2. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics.* 2005;6 Suppl 1:S2.

3. Morgan AA, Lu Z, Wang X, et al. Overview of BioCreative II gene normalization. *Genome Biol.* 2008;9 Suppl 2:S3.

4. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 2008;9 Suppl 2:S8.

5. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics.* 2005;6 Suppl 1:S11.

6. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics.* 2005;6 Suppl 1:S13.

7. Liu H, Wu C, Friedman C. BioTagger: a biological entity tagging system. Paper presented at: BioCreAtive Workshop, 2004; Spain.

8. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: An gene/protein entity tagging system. *Journal of American Medical Informatics Association.* 2008;In press.

9. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput.* 2008:652-663.

10. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics.* Jul 15 2005;21(14):3191-3192.

A Novel Ranking-based Gene Normalization System

Jingchen Liu[1], Minlie Huang[1§], Xiaoyan Zhu[1]

[1]Department of Computer Science and Technology, Tsinghua University, China
[§]Corresponding author

Email addresses:
JLiu: liu-jc04@mails.tsinghua.edu.cn
MHuang: aihuang@tsinghua.edu.cn
XZhu: zxy-dcs@tsinghua.edu.cn

Abstract

Gene normalization is one of the most challenging tasks in bio-literature mining. In this paper, we present a novel ranking based gene normalization system. Our system has four modules: (1) a gene mention recognition module which combines results from a CRF-based NER procedure, a dictionary-based NER procedure, and the ABNER system; (2) a candidate gene ID generation module using Lucene indexing and search APIs; (3) a disambiguation module based on a learning-to-rank algorithm; and (4) a confidence score generation module which outputs a list of final gene IDs and corresponding scores for each gene ID. Evaluation results on the most difficult 50 articles provided by BioCreAtIvE III show that our system is very competitive.

Introduction

Gene Normalization (GN), which maps a gene mention in the literature to a unique database identifier, is a quite challenging task. The task has been addressed by BioCreAtIvE challenge I, II and III. The following issues, which will be addressed in our system, have made the task very difficult:

(1) Gene mentions, recognized from upstream NER components, are not always necessary or suitable for gene normalization. Particular examples are gene families and protein complexes, which should be recognized as gene mentions but should not be normalized. However, most GN systems used Gene Mention Recognition components as a black box. The NER results were used as input of the GN systems directly. In our system, as inspired by reference 1, we filter those gene mentions that are not suitable for normalization, such as some gene families and complexes.

(2) Genes are often mentioned in the text with non-canonical names, rather than referred by the names defined in the database [2]. For example: '*p65 subunit of NF-kappaB*' and '*light chain-3 of microtubule-associated proteins*'. Exact matching for such cases is hard so that it's difficult to normalize such genes.

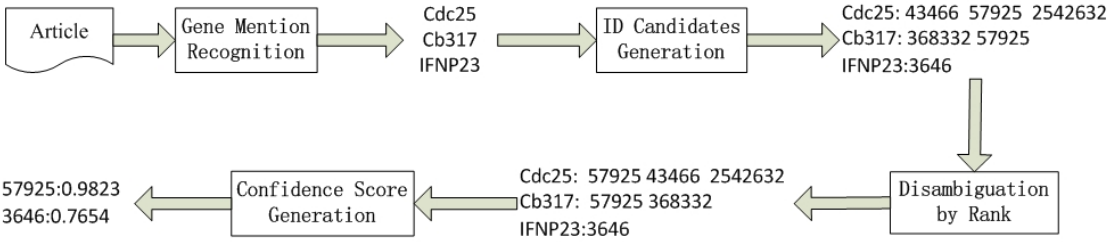
(3) Some gene names are highly ambiguous [2] and it's hard to decide to which species a gene belongs. Homologous genes share same or similar name. Even in the same species, different genes can share one same name.

Most systems participating in the BioCreAtIvE II GN task and some follow-up studies used rule-based methods to deal with these problems [1][2]. Some systems obtained good performance by making great efforts to compile lexicon and to refine rules. The rule based method requires specific domain knowledge and makes the system inflexible. In this work, we aim to leverage statistical machine learning technologies to reduce the system's dependence on specific domain lexicons and rules.

System Description

Our system has a cascade framework consisting four major modules, as shown in Figure 1. The details of each module are described in the following subsections.

Figure 1 - The cascade framework of our system



Gene Mention Recognition

The first module of our system recognized gene mentions (GM) in the text by combining four independent NER components. We have four NER components, including (1) a component extracting the text labeled by the <*itac*> tag from full-texts, with species names removed; (2) a gene mention recognition module based on a CRF that was trained on the corpus from the BioCreAtIvE II Gene Mention Recognition task; (3) a dictionary-based gene mention recognition module, where the lexicon was compiled from Etrenz Gene; (4) the ABNER system[3], which is an open source NER system for biomedical text. The recognized GMs from different components were merged by retaining those GMs that come from at least two components out of the last three components, while the results from <*itac*> tag was always kept due to its high precision. The overlapping part of mentions from different components was taken as the final mention if individually recognized boundaries are different.

Gene ID Candidates Generation

The second module generated gene ID candidates for each recognized GM. Lucene was used to index the gene entries in Entrez Gene. Each GM was searched by Lucene to get the top 50 gene IDs as its ID candidates. During indexing and search, the gene names from Entrez Gene and the GMs from text were processed by a set of rules: (1) removing all special characters such as dashes and underscores; (2) removing stop words; (3) inserting a white space at where lowercase changes to uppercase or vice versa, such as *'hBCL'* into *'h BCL'*; (4) separating digits, Greek letters (*alpha*, *beta* etc.), Roman numbers from other alphabet letters; (5) changing to lowercases.

Disambiguation using Learning-to-Rank

The third module ranked the ID candidates for each GM using a learning-to -rank algorithm: ListNet [4]. The training data was built from the 32 training articles with full annotation. For each gene mention recognized in the training article, the gene ID candidates were compared with the gold standard. If a gene ID is in the gold standard, it is marked as a positive example; otherwise it is a negative example. For each gene ID, its detailed information was found in Entrez Gene. A feature vector was then built for each gene ID. Features are described in Table 1. We tried to rank the correct gene ID to the top one position of the candidate list.

Table 1 - Features for disambiguation using learning-to-rank

Feature Name	Description
Species In GM	Whether the species of the gene ID is implied by GM (gene mention) like *hBCL*.
Species In Document	Whether the species of the gene ID appears in the document.
Species In Title	Whether the species of the gene ID appears in the title.
Nearest species	Whether the species of the gene is the nearest species from the GM in the context.
Type of gene	Whether the value of *type_of_gene* attribute (in the EntrezGene database) appears in the nearby context. For example, if the type of the gene ID is *protein encoding*, the appearance of the word *protein* nearby might be useful evidence.
Symbol Edit Distance	The Edit Distance between GM and the *symbol* of the Gene ID.
Synonyms Edit Distance	The minimum Edit Distance between the GM and all the synonyms of the gene ID.
Number similarity	The number of same 'numbers' between GM and the symbol of the gene ID. The 'numbers' here includes the digitals, the Roman numbers, the Greek letters and single English letters.
GM Lucene Score	The score returned by Lucene using the GM as the query.
Extended GM Lucene Score	Extend the GM by 3 words before and after it, and then get the Lucene score using the extended GM as the query.
Full Name or Abbreviate	If the GM has full or abbreviated name in the context, compute the Edit Distance between the full or abbreviated name and the synonyms of the gene ID and then use the minimum value as the feature.
Synonyms In Sentence	Find the words in the synonyms indicating the gene's function (*death*, *binding*, *interacting* etc) then check whether such signature words appear in the context.

Confidence Score Generation

The last module was used to output the final gene ID set and to associate a confidence score with each gene ID. Two strategies were attempted to generating the output gene ID set. The first run (denoted as TOP1) was only keeping the top 1 gene ID candidate of each gene mention, and the second run (TOP10) was obtained as follows: firstly find the top 10 gene IDs from each gene mention, and then only maintain one gene ID with the highest rank for every species. Then we used a supervised classification method to decide the confidence score for each gene ID. The training data was built similarly as mentioned before: the system was run on the 32 fully annotated articles to get the gene ID set, and then positive and negative examples were labeled according to the gold standard. A feature vector was built for each gene ID, as described in Table 2. We experimented with two classification models: Logistic Regression and Support Vector Machine. The probability of a gene ID being positive, given by the classification model, was used as the gene ID's confidence score in the final output.

Table 2 - Features for confidence score generation

Feature Name	Description
Best values of the Ranking Features	Reuse the features in the **Disambiguation using Learning-to-Rank** module. As one gene ID may be mapped from several gene mentions and each mention has a feature vector, only the 'best' value of each feature from different mentions is used here. For instance, the maximal value for *GM Lucene Score* feature and the minimal value for *Symbol Edit Distance* feature.
GM amount	The number of GMs containing the gene ID
GM sources	Whether the gene ID has GM recognized by a specific NER component described in **Gene Mention Recognition** section. For example, if the gene ID has a GM recognized by the CRF component, then the value is *1* for the *GM source* feature of CRF. This feature was built for each NER component.
Highest Rank	The highest rank of the gene ID among all the GMs containing the ID.
Min Word Number	The minimum word number of the GMs containing this ID.
Uppercase or digital	Whether one of the GMs containing the ID has Uppercase letters or digits.

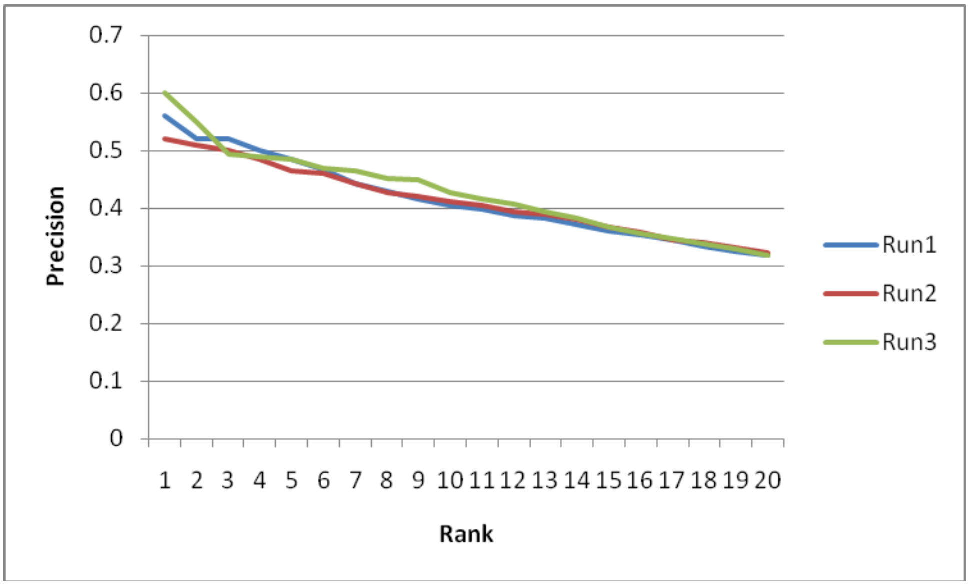
Results

We tested our system on the most difficult 50 articles provided by BioCreAtIvE III. The results were evaluated by TAP-5, TAP-10 and TAP-20[6]. The test results of our 3 submissions are shown in Table 3. The 3 Runs differ in the strategies to generating the output gene ID set: TOP1 and TOP10 described in the previous section, and in the classification model to generate the confidence score: Logistic Regression (LR) and Support Vector Machine (SVM). In addition to the official evaluation of BioCreAtIvE III, we also evaluated our submissions by Precision At different ranks, as shown in Figure 2. From the evaluation results, we can conclude that the TOP10 strategy and SVM are more suitable for this task.

Table 3 - Test results of 3 submissions

Run	Gene ID set strategy	Classification Model	TAP-5	TAP-10	TAP-20
Run 1	TOP1	LR	0.2805	0.2971	0.3064
Run 2	TOP10	LR	0.2850	0.3033	0.3044
Run 3	TOP10	SVM	0.2973	0.3125	0.3248

Figure 2 - The curve of Precision At Ranks for 3 submissions



References

1. Hakenberg J, Plake C, Leaman R, Schroeder M and Gonzalez G: **Inter-species normalization of gene mentions with GNAT**. *Bioinformatics,* Vol. 24, pages i126-i132. 2008
2. Morgan A, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu H, Torres R, Krauthammer M, Lau WM，Liu H, Hsu C, Schuemie M, Cohen KB and Hirschman L: **Overview of BioCreative II gene normalization.** *Genome Biology* , 9(Suppl 2):S3. 2008
3. Settles B: **ABNER: An Open Source Tool for Automatically Tagging Genes, Proteins, and Other Entity Names in Text**. *Bioinformatics*, 21(14):3191-3192. 2005.
4. Cao Z, Qin T, Liu TY, Tsai MF, Li H: **Learning to Rank: From Pairwise Approach to Listwise Approach.** *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.
5. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA Data Mining Software: An Update**. *SIGKDD Explorations*, Volume 11, Issue 1. 2009
6. Carroll HD, Kann MG, Sheetlin SL and Spouge JL: **Threshold Average Precision (TAP-*k*): a measure of retrieval designed for bioinformatics.** *Bioinformatics,* 26(14):1708-1713. 2010

Testing Extensive Use of NER tools in Article Classification and a Statistical Approach for Method Interaction Extraction in the Protein-Protein Interaction Literature

Anália Lourenço[1] , Michael Conover[2,3] , Andrew Wong[4] , Fengxia Pan[5] , Alaa Abi-Haidar[2,3] , Azadeh Nematzadeh[2,3] , Hagit Shatkay*[6] , Luis M. Rocha*[2,3]

[1]IBB/CEB, University of Minho, Campus Gualtar, Braga, Portugal
[2]School of Informatics and Computing, Indiana University, USA
[3]FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Portugal
[4] School of Computing, Queen's University, Kingston, ON, Canada
[5] Microsoft Corp., Redmond, WA, USA
[6] Dept. of Computer and Information Sciences, University of Delaware, USA

Email: Anália Lourenço - analia@deb.uminho.pt; Michael Conover - midconov@indiana.edu; Andrew Wong - 3aw14@queensu.ca; Fengxia Pan - fepan@microsoft.com; Alaa Abi-Haidar - aabihaid@indiana.edu; Azadeh Nematzadeh - azadnema@indiana.edu; Hagit Shatkay*- shatkay@cis.udel.edu; Luis M. Rocha*- rocha@indiana.edu;

*Corresponding author

Abstract

We participated (as Team 81) in the *Article Classification* (ACT) and *Interaction Method* (IMT) subtasks of the *Protein-Protein Interaction* task of the Biocreative III Challenge. For the ACT we pursued an extensive testing of available Named Entity Recognition (NER) tools, and used the most promising ones to extend our the *Variable Trigonometric Threshold* (VTT) linear classifier we successfully used in BioCreative II and II.5. Our main goal was to exploit the power of available NER tools to aid in the document classification of documents relevant for Protein-Protein Interaction. We also used a Support Vector Machine Classifier on NER features for comparison purposes. For the IMT, we experimented with a primarily statistical approach, as opposed to a deeper natural language processing strategy; in a nutshell, we exploited classifiers, simple pattern matching, and ranking of candidate matches using statistical considerations. We will also report on our efforts to integrate our IMT method sentence classifier into our ACT pipeline.

Article Classification Task

We participated in both the online submission with our own annotation server implementing the VTT algorithm via the BioCreative MetaServer platform, as well as the offline component of the Challenge. We used three distinct classifiers: (1) the lightweight *Variable Trigonometric Threshold* (VTT) linear classifier that employs word-pair textual features and protein counts extracted using the ABNER tool [1], and which we successfully introduced in the abstract classification task of BioCreative II [2] as well as on the full-text scenario of Biocreative II.5 [3], (2) a novel version of VTT that includes various NER features as well as various sources of textual features, and (3) a Suport Vector Machine (using SVM^{light}) that takes as features various entity count features from the NER tools we tested.

In the novel version of VTT that included various NER features, a document d is considered to be relevant if:

$$M.\sum_{f=1}^{F}\frac{P_f(d)}{N_f(d)} \geq \lambda_0 + \sum_{\pi=1}^{EP}\frac{\beta_\pi - n_\pi(d)}{\beta_\pi} - \sum_{\nu=1}^{EN}\frac{\beta_\nu - n_\nu(d)}{\beta_\nu} \tag{1}$$

where λ_0 is a constant threshold for deciding whether a document is positive/relevant or negative/irrelevant. $P_f(d)$ and $N_f(d)$ are occurrence counts of discriminative features (see [3] for details) for feature set f. These features can be textual features (such as bigrams) or features from entity recognition tools. EP is the number of entity count features, π, correlated with relevant documents, and EN is the number of entity count features, ν, correlated with irrelevant documents; $M = EN + EP$.

In addition to testing the power of available NER tools to aid in the document classification of documents relevant for Protein-Protein Interaction, we were interested in answering a few other questions: (1) is there a benefit to using word bigrams as textual features, in comparison to the simpler word-pairs we previously employed [2, 3]? (2) Is it advantageous to use additional PPI classification data from previous Biocreative challenges, or is it best to use only Biocreative III data? (3) how much, if at all, does full-text data help on the classification? Given the time limitations of the challenge, the submitted runs will only allow us to respond to our main question (the utility of existing NER tools) and additional question (1) above. We intend to test questions (2) and (3) post-challenge.

Towards responding to our main question, we utilized the following NER tools and dictionaries: ABNER [1], NLProt, Oscar 3, CHEBI (Chemical names), PSI-MI, MeSH terms, and BRENDA enzyme names. With each one of these tools, we extracted various types of features in abstracts and in figure and table captions. We then computed occurrence counts of the various feature types, for instance: Number of protein mentions in an abstract identified by ABNER, or PSI-MI method mentions in figure captions. Finally, we selected those *entity feature counts* that best discriminated relevant and irrelevant documents in the training and development data. This was done via the analysis of charts such as those described in Figure 1, which depicts a comparison of the counts of ABNER protein mentions in abstracts and BRENDA enzyme names in figure captions on Biocreative III training data (excluding development data). As can be seen, the counts of BRENDA enzyme name mentions in figure captions of documents in the training data does not discriminate well between relevant and irrelevant documents. In contrast, counts of ABNER protein mentions in abstracts are distinct for relevant and irrelevant documents. We used this type of plot to identify which features from NER tools and which document portions behaved differently for relevant and irrelevant documents. For our extended VTT classifier, we used the following five entity feature counts: ABNER protein mentions in abstracts, NLProt protein mentions in abstracts, PSI-MI methods in abstracts, ABNER protein mentions in figure captions, and Oscar compound names in figure captions—which were all positively correlated with relevant documents (therefore $EN = 0$ and $M = EN = 5$ in equation 1) We rejected many other entity feature counts, but provide the community with our feasibility study of the various NER Tools as aids for PPI-relevance article classification. Moreover, we used all entity count features to a SVM classifier to understand the performance of those features alone in classifying PPI-relevant documents.

The Interaction Method Task

We note that the BioCreative training set consisted of full-text articles along with the identifiers of the PPI detection methods that were judged to be discussed in them, *without* any tagging of the sentences that formed the actual evidence for the method. Hence the training corpus could not be used to directly train a classifier to identify PPI method sentences. To make up for this shortcoming, we used a corpus that was developed independently and used in a previous work by Shatkay et al [4]. In that work, Support Vector Machine (SVM) and Maximum Entropy classifiers were trained using a corpus of 10,000 sentences from full-text biomedical articles, which were tagged at the sentence-fragment level, along five dimensions: *focus* (methodological, scientific or generic), *type of evidence* (experimental, reference, and a few other types), *level of confidence* (from 0 - no confidence, to 3 - absolute certainty), *polar-*

ity (affirmative or negative statement), and *direction* (e.g. up-regulation vs. down-regulation). Notably, that corpus had little or nothing to do with protein-protein interaction, but a classifier trained on the *Focus* dimension showed high sensitivity and specificity in identifying *Methods sentences*, and as such we have used it without any retraining. We also used classifiers trained to tag text along the other dimensions, but as almost all sentences were of affirmative polarity and high confidence, we decided to use only the Focus classifier (particularly, whether or not a sentence was classified as a *Methodology* sentence). Using the converted text files provided by BioCreative, we applied a simple strategy for breaking the corpus into sentences based on a modified version of the Lingua-EN-Sentence Perl module [5]), and eliminated any text segment that looked like a bibliographic reference using a simple rule-based strategy. The remaining sentences were converted into a simple binary term-vector representation, for the purpose of classifying each sentence by Focus, utilizing a SVM classifier [4]. This classification step did not identify which method is discussed; rather, it only identifies candidate sentences that may discuss methods.

The specific Method Identifiers (MIs) were then associated with sentences by simple pattern-matching to PSI-MI ontology terms (the primary name and synonyms characterizing each concept)), loaded using the *OBO::Parser::OBOParser* Perl module (part of *ONTO-Per*l package [6]). To allow, to some extent, partial matches, and shuffling of word-order in matches we used two Perl modules: *Text::Ngramize* [7], and *Text::RewriteRules* [8]. The module *Lingua::StopWords* was used to avoid the matching of common English words [9]. As such simple pattern matching can lead to many spurious matches, we scored matches such that exact matches are scored higher than partial ones and longer matches score higher than shorter ones.

Each sentence was thus tentatively associated with all the MIs whose terms hit the sentence. Statistical considerations were used to post-process this many-to-many mapping, selecting one MI among multiple MIS that hit the same sentence, while selecting a single sentence as evidence for each matched MI. Employing several scoring schemes similar in spirit to TF*IDF, we scored each sentence for each candidate MI, based on the length of the match (the higher the better), how rare or frequent the matched terms were in the corpus, in the sentence, and in the methods ontology (rare terms score higher, frequent - lower), and increasing the score for sentences that were classified as *Methodology* in the classification step described earlier. The MIs that scored the highest were reported, and the sentence that gave rise to the score was provided as evidence. The different runs we have submitted varied in the scoring methods used, and in the thresholds placed over the scores to select the MIs that were actually reported.

Integrating the ACT and IMT pipelines

While we were unable to integrate both pipelines for an ACT submission, we are working post-challenge to utilize the output of our IMT pipeline as additional entity features in our ACT pipeline. We will report on this development at the Biocreative III workshop.

Authors contributions

Anália Lourenço was responsible for all NER tool extraction for the ACT and participated in the development and testing of the IMT pipeline. Michael Conover was responsible for the processing of all NER entity features in the ACT, design and implementation of data model for ACT pipeline, analysis of entity features, and design and implementation of SVM classifier. Andrew Wong participated in the development and testing of the IMT pipeline. Fengxia Pan developed, implemented, trained and tested the classifiers which were used in the IMT pipeline, as part of her MSc work at the School of Computing, Queen's University, Kingston, Ontario. Alaa Abi-Haidar produced the code necessary for implementing the VTT method. Azadeh Nematzadeh produced code to extract textual features from documents. Hagit Shatkay developed the methodology and experimental set up for the IMT. Luis M. Rocha developed the methodology and experimental set up for the ACT.

Acknowledgements

Andrew Wong's work was supported by NSERC Discovery Grant, NSERC Discovery Acceleration Supplement, and Ontario's Early Researcher Award, awarded to Hagit Shatkay. Michael Conover, Alaa Abi-Haidar and Azadeh Nematzadeh were supported with a grant from the FLAD Computational Biology Collaboratorium at the Instituto Gulbenkian de Ciencia in Oeiras, Portugal. We also thank support from this grant for travel, hosting and providing facilities used to conduct part of this research. We thank Artemy Kolchinsky for assistance in setting up the online server for the ACT.

References

1. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text**. *Bioinformatics* 2005, **21**(14):3191–3192.
2. Abi-Haidar A, Kaur1 J, Maguitman A, Radivojac P, Retchsteiner A, Verspoor K, Wang Z, Rocha LM: **Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks**. *Genome Biology* 2008, :9(Suppl 2):S11, [http://genomebiology.com/2008/9/s2/S11/abstract/].
3. Kolchinsky A, Abi-Haidar A, Kaur J, Hamed AA, Rocha LM: **Classification of Protein-Protein Interaction Full-Text Documents Using Text and Citation Network Features**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7**:400–411.
4. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ: **Multi-Dimensional Classification of Biomedical Text: Toward Automated, Practical Provision of High-Utility Text to Diverse Users**. *Bioinformatics* 2008, **24**(18):2086–2093.
5. **CPAN module, Lingua-EN-Sentence**[http://search.cpan.org/~shlomoy/Lingua-EN-Sentence-0.25/lib/Lingua/EN/Sentence.pm].
6. **CPAN module, ONTO-PERL**[http://search.cpan.org/~easr/ONTO-PERL-1.23/].
7. **CPAN module, Text-Ngramize**[http://search.cpan.org/~kubina/Text-Ngramize-1.03/lib/Text/Ngramize.pm].
8. **CPAN module, Text-RewriteRules**[http://search.cpan.org/~ambs/Text-RewriteRules-0.23/lib/Text/RewriteRules.pm].
9. **CPAN module, Lingua::StopWords**[http://search.cpan.org/dist/Lingua-StopWords/].

Figures

Figure 1 - Entity Counts Analysis

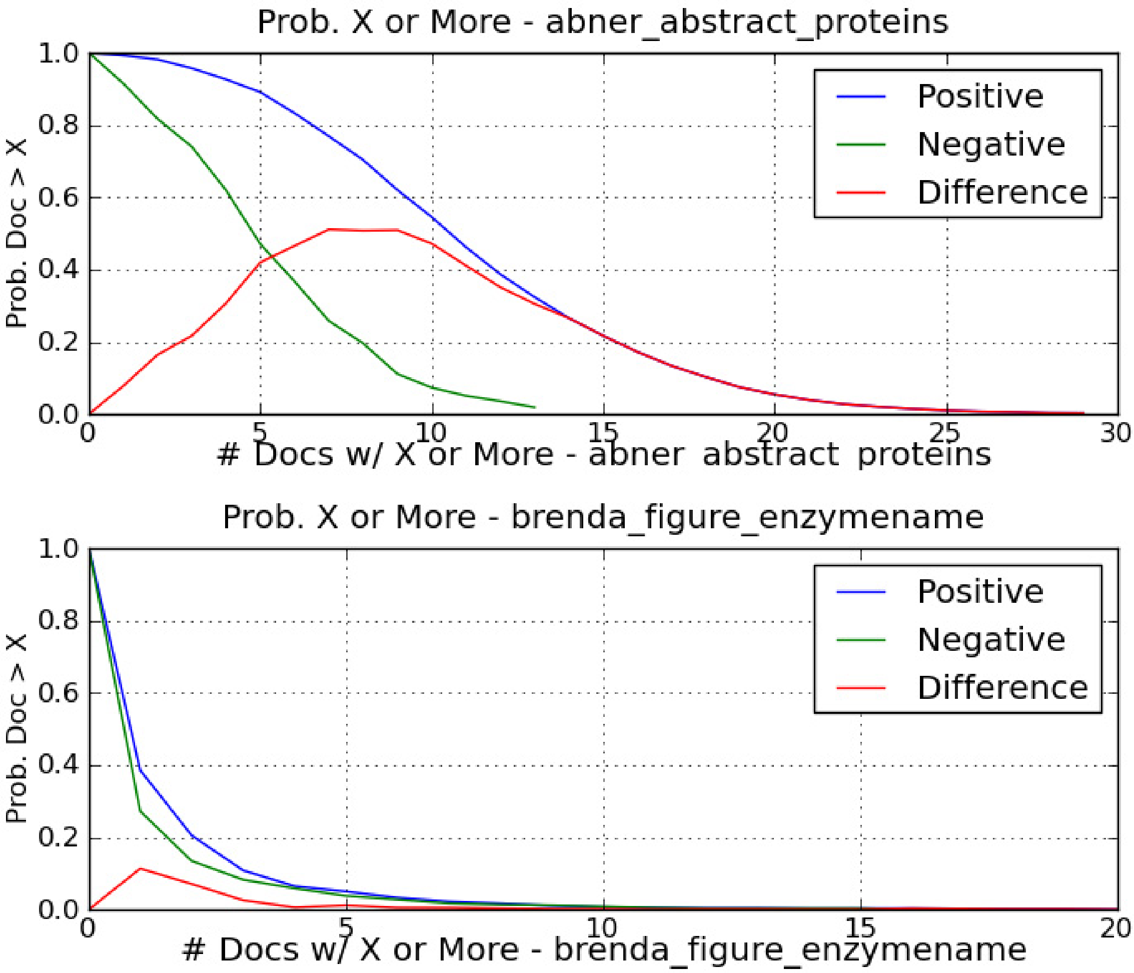


Figure 1: Comparison of the counts of protein mentions as identified by ABNER in abstracts of the articles (top), and BRENDA enzyme names in figure captions (bottom). Results shown for iocreative III training data (excluding development data). The horizontal axis represents the number of mentions x, and the vertical axis the probability $p(x)$ of documents with at least x mentions. The blue lines denote documents labeled relevant, while the green lines denote documents labeled irrelevant; the red lines denote the difference between blue and red lines.

Vector-space models and terminologies in gene normalization and document classification

Sérgio Matos[§], David Campos, José L Oliveira
{aleixomatos, david.campos, jlo}@ua.pt
[1]Institute of Electronics and Telematics Engineering of Aveiro, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[§]Corresponding author

Abstract

Background

This article presents the results of the participation in the BioCreative III tasks: Gene Normalization (GN); Protein-Protein Interaction (PPI) - Article Classification Task (ACT); and Protein-Protein Interaction - Interaction Method Task (IMT).

Results

We obtained a TAP score of 0.2790 on the training set of the GN task. On the PPI-ACT task, we obtained an AUC iP/R of 0.568 on the test set. The highest accuracy obtained was 0.874, and the highest MCC value was 0.466. For the PPI-IMT task, we submitted three runs with AUC iP/R above 0.3, and one run with AUC iP/R above 0.4 (although this run only returned results for 30 documents).

Conclusions

We participated on two tasks of the BioCreative III challenge, with encouraging results on both PPI subtasks. Further work is required in order to improve our approach for gene normalization.

Methods and Results

Gene normalization task

We used a disambiguation approach based on gene-profiles and obtained an average Threshold Average Precision (TAP) score of 0.2790 on the training set. The workflow is composed of five modules, described below:

A. Pre-processing
This module is responsible for parsing the input XML to plain text and for splitting this into sections (title, abstract, etc.) and into sentences.

B. NER

This module is based on a Conditional Random Fields (CRF) model trained in Mallet [1] using the BioCreative II Gene Mention corpus. The set of features used includes: word stemming (Snowball stemmer), part-of-speech tagging (OpenNLP [2]), orthographic and morphological features; roman numbers and greek letters; dictionary-matching of gene/protein names (BioThesaurus [3]), dictionary-matching of relevant verbs (BioLexicon [4]); dictionary-matching of other biological concepts, such as nucleobases, amino acids and nucleic acids. For tokenization we used the OpenNLP tokenizer. In the end, a {-1,1} window of features is used to model local context. Post-processing techniques were not used in this implementation.

C. Dictionary matching

Each mention returned by the NER module is mapped to possible identifiers through dictionary-lookup. We implemented a Lucene [5] index of the BioThesaurus resource for efficient approximate string search. To create the index, each entry of the BioThesaurus dictionary was first pre-processed using various string-editing rules, in order to create lexical variations of the name. These string-editing rules include: removing dashes; replacing dashes by spaces; inserting a dash on a letter-digit sequence; replacing arabic numerals by roman numerals (and vice-versa); replacing greek letter names by their initial (e.g. *alpha* → *a*). All variations were added to the index, together with the corresponding UniProt accession number and Entrez Gene identifiers. Entrez Gene IDs were obtained from the mappings file available from UniProt [6].
We tested the coverage of our dictionary using the gene mentions from the GN training data and obtained a recall rate of 93.7%.

D. Context matching

This module compares the local context of each mention to previously computed biological knowledge profiles. In this implementation, the sentence was used as local context. The gene/protein profiles were implemented as a Lucene index (separate from the dictionary index). To create this index, we parsed the UniProt data file and retrieved Gene Ontology terms and descriptive fields. We also obtained the descriptive fields from mapped entries in Entrez Gene and OMIM. All these text fields were added to the index as free text. To obtain the scores for context matching, the sentence where the mention occurs is used to search this index. The resulting identifiers are then cross-matched to the candidate identifiers returned from the dictionary matching step.

E. Rule-based decision

All possible identifiers obtained after context matching, for the complete article, are assembled and the most likely identifiers are selected, according to some empirically created rules. The rules select identifiers matched with more than five mentions in text, or matched to at least three mentions and with an average context-matching score equal to or higher than a preset threshold (0.8). Other identifiers matched to the same text mentions but with lower scores are rejected.

PPI Article Classification Task

Our approach in the ACT task is based on vector-space similarity. Documents in the training set are represented as vectors of the biologically relevant words occurring in the text. The underlying lexicon includes a list of interaction methods from the Interaction Method Ontology (PSI-MI) [7], distributed by the organizers for the PPI-IMT task, and biological verbs and their nominalizations, obtained from the BioLexicon term repository. Before creating the index, each document is pre-processed in order to identify occurrences of these terms. For each term found, we add to the document vector both the textual occurrence and its corresponding lemma. This index also contains the class of each document (1 for relevant, or 0 for non-relevant).

During the classification of a new document, the same pre-processing stage is performed and each occurrence of a lexicon term is added to the query string that is then used to search the index. From this search we retrieve the top M documents, together with the corresponding scores and classes. The class probability for the new document is then calculated as the sum of the Lucene scores for each class, normalized by the total scores for the M documents. A threshold is then used to select the class for that document. Figure 1 illustrates these classification steps.

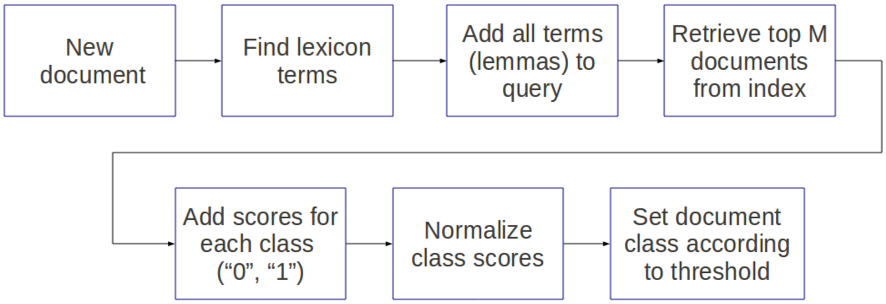

New document
Find lexicon terms
Add all terms (lemmas) to query
Retrieve top M documents from index
Add scores for each class (“0”, “1”)
Normalize class scores
Set document class according to threshold

Figure 1 - Document classification

We compared the use of lemmas to the use of the textual occurrence of the lexicon terms, and observed improvements in AUC iP/R between 3% (for M=50) and 6% (for M=500). Figure 2 shows the iP/R curves for M=500 when using the textual occurrence of BioLexicon terms, the lemmas of these terms, and the lemmas plus PSI-MI terms.

Using training and development data, with 5-fold cross-validation, we obtained an average AUC iP/R of 0.5947, using M=500 documents.

On the test set, we obtained a AUC iP/R of 0.56760. The full test set results are shown in Table 1.

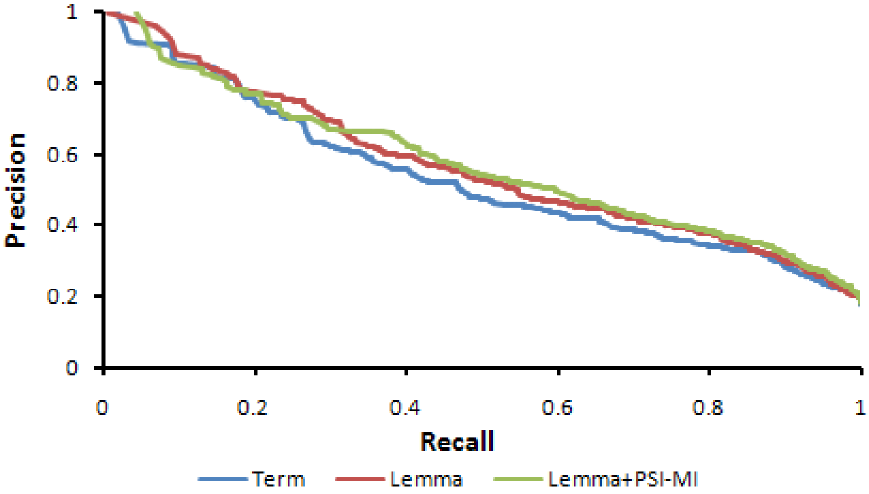


Figure 2 - Use of different vocabularies

Table 1 - PPI-ACT test results

Results on the ACT test data (M=500). For all runs: AUC iP/R=0.56760; P at full R=0.15744. MCC=Matthew's correlation coefficient

Run	Sensitivity	Specificity	Accuracy	MCC	Precision	F-score
1	0.94176	0.49695	0.56445	0.31789	0.25088	0.39621
2	0.38791	0.96108	0.87410	0.43346	0.64065	0.48323
3	0.72527	0.83605	0.81924	0.46563	0.44177	0.54908
4	0.96593	0.39041	0.47774	0.27060	0.22085	0.35951
5	0.20989	0.98624	0.86843	0.34488	0.73180	0.32622

PPI Interaction Method Task

For the IMT task our approach was to find mentions of methods names and synonyms in the texts and apply a very simple heuristic to validate and rank the classifications. We used the list of valid names provided by the organizers.

To facilitate approximate string searches, we also used Lucene in this task. All documents in the test set were added to an index, using Lucene's standard analyzer. We then search this index for each entry in the dictionary of methods names and retrieve the top 100 documents for each search. For synonyms of the same method (same PSI-MI identifier), the document scores are added together. Finally, a method ID is assigned to a document if that document/method score is above a threshold.

This method could possibly be improved by introducing relevant keywords that are specific to particular interaction methods. This could be achieved, at a first stage, through a simple co-occurrence analysis in the training and development sets. However, due to time constraints, this was not performed.

Table 2 - PPI-IMT test results

Results on the IMT test data. N is the number of documents with results for each run. Precision, Recall and F-Score are macro-averaged.

Run	N	Precision	Recall	F-Score	AUC iP/R (macro)	AUC iP/R (micro)
1	143	0.51783	0.35012	0.37838	0.31402	0.18053
2	72	0.71759	0.36806	0.45608	0.36215	0.24091
3	30	0.80000	0.41500	0.51508	0.41500	0.27245
4	205	0.31648	0.38715	0.31747	0.32295	0.16436
5	159	0.36363	0.21258	0.24754	0.18976	0.07689

Discussion and Conclusions

The use of domain terminologies and vector-space models for classification of PPI relevant documents provided positive and encouraging results. The use of other lexical resources (e.g. Gene Ontology terms) may help improve the results obtained. Comparing to the use of classification models such as SVMs, our approach has the advantage that to add more classified documents as new information to the classifier only involves adding those documents, with the corresponding classification, to the index.

Our work on the GN task, based on knowledge profiles, show that this is a valid approach. From our inspection of the results, the main shortcomings of our implementation are in matching the local context of the entity mention in the text with the stored profiles, and in the final decision stage.

Acknowledgements

The research leading to these results has received funding from “Fundação Para a Ciência e a Tecnologia” (FCT) under the project number PTDC/EIA-CCO/100541/2008 (FCOMP-01-0124-FEDER-010029). S. Matos is funded by FCT under the Ciência2007 programme.

References

1. Mallet - MAchine Learning for LanguagE Toolkit [http://mallet.cs.umass.edu/]
2. OpenNLP [http://opennlp.sourceforge.net/]
3. Liu H, Hu ZZ, Zhang J, Wu C: **BioThesaurus: a web-based thesaurus of protein and gene names.** Bioinformatics 2006, **22**(1):103-5.
4. Sasaki Y, Montemagni S, Pezik P, Rebholz-Schuhmann D, McNaught J, Ananiadou S: **BioLexicon: A Lexical Resource for the Biology Domain.** In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine: 1-3 September 2008; Turku, Finland.* Edited by Salakoski T, Rebholz-Schuhmann D, Pyysalo S; 2008:109-116.
5. Apache Lucene [http://lucene.apache.org/]
6. UniProt Knowledge Base [ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/]
7. Molecular Interaction Ontology

[http://psidev.sourceforge.net/mi/rel25/data/psi-mi25.obo]

The gene normalization and interactive systems of the University of Tokyo in the BioCreative III challenge

Naoaki Okazaki*[1], Han-Cheol Cho[2], Rune Sætre[1], Sampo Pyysalo[1], Tomoko Ohta[1]and Jun'ichi Tsujii[1,3]

[1]Graduate School of Interdisciplinary Information Studies, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
[2]Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
[3]National Centre for Text Mining, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street M1 7DN UK

Email: Naoaki Okazaki*- okazaki at is.s.u-tokyo.ac.jp;

*Corresponding author

Abstract

This paper presents our systems of gene normalization (GN) and interactive demonstration (IAT) tasks in the BioCreative III challenge. The GN system is based on various NLP modules including the GENIA sentence splitter, tokenizer and POS tagger, the NERsuite system and a species mention recognizer. For each gene mention in a given article, the system enumerates candidate Gene-IDs that can be assigned to that gene mention, and scores each candidate Gene-ID by using logistic regression. Gene-IDs found in the article are ranked by the sum of the mention-level scores. The results of gene normalization are integrated into the IAT system, together with results from the MEDIE system. The user interface consists of one tab for each source of input to the system (MEDIE, GNsuite and Linnaeus), and a table-tab that summarizes and ranks genes based on the combined input. The gene table can be manipulated both manually and automatic, and can be stored to a local file on the users computer. The whole IAT system is completely web based, but it currently relies on preprocessed input from the underlying systems. As more underlying systems become available as web-services, the IAT system will be able to process the data on-the-fly.

Results and Discussion

Gene normalization

We submitted three runs with the following configurations: 1) Gene-ID candidates without species filtering; 2) Gene-ID candidates with species filtering; and 3) the configuration 2 but Gene-IDs appearing only in the experimental section are removed. Table 1 reports the evaluation results on the test data. The top-x precision (P), recall (R), and F1 scores are micro-averages of those computed for each article; we compute top-x P, R, and F1 scores for each article, where x is the number of gold-standard Gene-IDs for the article.

Interactive Demonstration (IAT)

The Web-page for visualizing all the results from our system shows PMC and PubMed identifiers for the available full text articles. The number of normalized gene mentions in the title/abstract/full-text for each given paper is also shown. The user can click on, or type in, one of the identifiers to show the full text. All the recognized gene names are highlighted in the text, and on the top of each paper's visualization page is a summary of all the genes in the paper. The user can click on any gene symbol to look up the corresponding gene entry in Entrez Gene.

In the overview, each gene symbol and the number of mentions in the current paper is listed (Figure 1 (b)). The user can jump from one occurrence to the next by clicking.

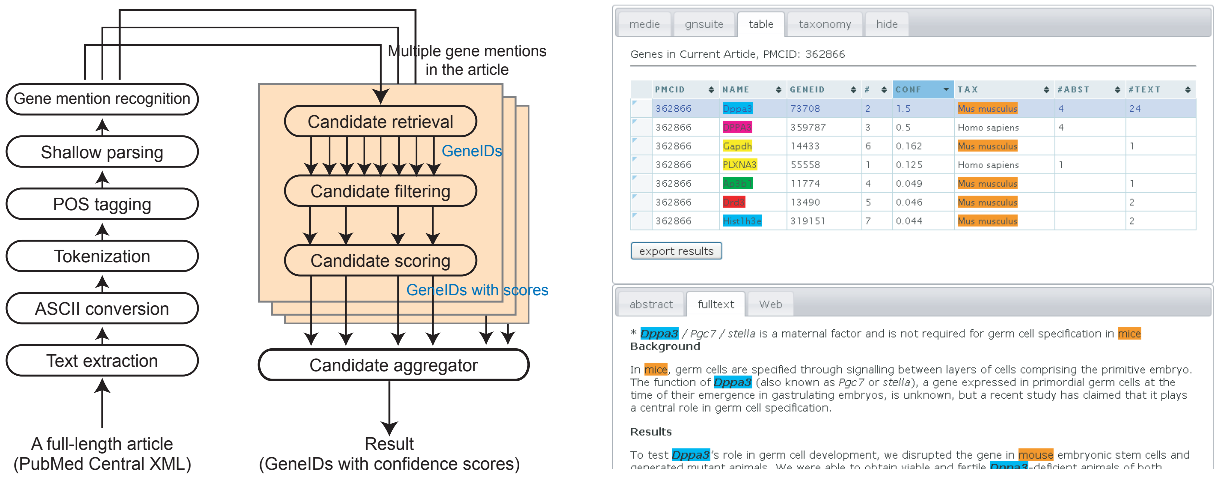
The gene normalization results are integrated with the genes found in the abstract by the MEDIE system (http://www-tsujii.is.s.u-tokyo.ac.jp/medie/). MEDIE uses the GENA dictionary with entries normalized to both Entrez Gene, Swiss-Prot, TrEMBL, Fly-base and several other major DBs. To map the names for a specific gene entry into the text we use a fast web service providing cached information from Entrez Gene (http://entrezajax.appspot.com/). The same web-service is also used to find alternative names for the species for each gene, and to highlight these species names in the text as well. 18 000 full text papers (from BioCreative III) are currently indexed. The list of genes for each paper can be sorted by relevance scores based on frequency, confidence etc. The list can also be edited by authorized users, and the results can be stored in individual user-accounts.

Method

Figure 1 (a) illustrates the outline of the system for gene normalization. The system recognizes gene mentions in source articles. For each gene mention, the system enumerates candidate gene identifiers (Gene-IDs), and computes the confidence score of each Gene-ID. Gene-IDs found in the article are ranked

Run	TAP-5	TAP-10	TAP-20	Top-x P R F1
1 (without species filtering)	0.1599	0.1842	0.2010	0.3337 0.1617 0.2178
2 (with species filtering)	0.1517	0.1804	0.2000	0.2981 0.1520 0.2014
3 (2 exc. experimental section)	0.1611	0.1856	0.2032	0.3400 0.1641 0.2213

Table 1: Results of gene normalization



Genes in Current Article, PMCID: 362866

	PMCID	NAME	GENEID	#	CONF	TAX	#ABST	#TEXT
	362866	Dppa3	73708	2	1.5	Mus musculus	4	24
	362866	DPPA3	359787	3	0.5	Homo sapiens	4	
	362866	Gapdh	14433	6	0.162	Mus musculus		1
	362866	PLXNA3	55558	1	0.125	Homo sapiens	1	
	362866	Ap3b1	11774	4	0.049	Mus musculus		1
	362866	Drd3	13490	5	0.046	Mus musculus		2
	362866	Hist1h3e	319151	7	0.044	Mus musculus		2

* ***Dppa3*** / *Pgc7* / *stella* is a maternal factor and is not required for germ cell specification in mice

Background

In mice, germ cells are specified through signalling between layers of cells comprising the primitive embryo. The function of ***Dppa3*** (also known as *Pgc7* or *stella*), a gene expressed in primordial germ cells at the time of their emergence in gastrulating embryos, is unknown, but a recent study has claimed that it plays a central role in germ cell specification.

Results

To test ***Dppa3***'s role in germ cell development, we disrupted the gene in mouse embryonic stem cells and generated mutant animals. We were able to obtain viable and fertile ***Dppa3***-deficient animals of both

(a) The workflow of GN

(b) Visualization of IAT

Figure 1: The workflow of gene normalization and visualization of interactive system

by the sum of the mention-level scores.

Resource

We converted the ASN.1 version of the Entrez Gene database into XML format by using the gene2xml[1] tool. In order to locate information described in Entrez Gene XML records, we defined shorthands for XPaths from the root node to content nodes. For example, we introduced a shorthand `gene/locus` as `/Entrezgene-Set/Entrezgene/Entrezgene_gene/Gene-ref/Gene-ref_locus` to access a gene locus. We defined 47 shorthands in this task (we omit the exact XPaths in this paper): 7 name fields (`gene/locus`, `gene/desc`, `gene/syn`, `prot/name`, `prot/desc`, `nomenclature/symbol`, `nomenclature/fullname`), 4 organism fields (`org/taxname`, `org/common`, `org/taxid`, `org/linage`), 26 descriptive fields (e.g., `summary`, `generif/text`, `comment/text`), and 10 PMID fields (e.g., `generif/pmid`, `comment/pmid`).

[1] gene2xml: http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_

Preprocessing

We extracted texts from contents of `<article-title>`, `<title>`, and `<p>` elements in the PubMed Central XML files. We replaced non-ASCII characters into ASCII, e.g., α into *alpha*, by using a conversion table because existing NLP tools (e.g., POS tagger and NER tagger) have not been trained with non-ASCII characters. Then, we applied various NLP modules to the texts: GENIA sentence splitter[2], tokenizer, GENIA POS tagger[3] [1], NERsuite[4], and species mention recognizer[5].

Gene mention recognition

We employed NERsuite for gene mention recognition. NERsuite is a toolkit for named entity recognition based on Conditional Random Fields (CRFs). The toolkit also provides functionalities for training a model from an annotated text with gazetteers as external dictionaries. We used the training corpus of the BioCreative II Gene Mention Recognition task and gazetteers extracted from UMLS (135 categories - "Gene or Genome", "Enzyme", "Chemicals", ...) and Entrez Gene (one category). We applied the tokenizer and the GENIA tagger which are bundled in NERsuite to obtain tokens, lemmas, POS tags and chunk tags. To increase the coverage of gazetteers, dictionary entries are normalized as follows: alphabets are lowercased; all consecutive numbers are converted into a single zero; and all consecutive non-alphanumeric characters excluding whitespaces are converted into a single under-bar symbol.

Gene normalization

Gene normalization assigns an Entrez Gene-ID for a gene mention. This is performed by two subtasks: *candidate retrieval* and *candidate scoring*. Candidate retrieval enumerates Gene-IDs that can be assigned to a given gene mention. Candidate scoring assesses a score of each candidate Gene-ID being referred to by the gene mention. Because the Gene Normalization task of BioCreative III requires a list of Gene-IDs and their confidence scores at article level, we introduce an additional component *candidate aggregation*, which computes article-level confidence scores from mention-level scores of Gene-IDs.

Candidate retrieval

Candidate retrieval is a specialization of information retrieval in which queries correspond to gene mentions and documents correspond to Entrez Gene records. We built inverted indices that associate the contents of

[2] GENIA sentence splitter: http://www-tsujii.is.s.u-tokyo.ac.jp/~y-matsu/geniass/

[3] GENIA POS tagger: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[4] NERsuite: http://www-tsujii.is.s.u-tokyo.ac.jp/nersuite/

[5] LINNAEUS 1.5: http://linnaeus.sourceforge.net/

Feature category	Feature value	Text region	Entrez Gene fields
mention-name	exact, approximate, token, token-subset, n-gram cosine similarity, n-gram overlap similarity	mention	`gene/locus, gene/desc, gene/syn, prot/name, prot/desc, nomenclature/symbol, nomenclature/fullname`
context	uni-gram cosine similarity	abstract, title, titles of references, paragraph, sentence, preceding five words, subsequent five words	`summary, generif/text, other/text, other/anchor, other/post-text, phenotype/text, phenotype/anchor, phenotype/post-text, peptide/text, peptide/anchor, peptide/post-text, comment/text, comment/anchor, comment/post-text, function/text, function/anchor, function/post-text, process/anchor, process/post-text, component/anchor, component/post-text, mRNA/peptide/anchor, mRNA/peptide/post-text`
PMID	identity [binary]	PMID of the paper, PMID of the reference papers	`generif/pmid, other/pmid, phenotype/pmid, peptide/pmid, comment/pmid, function/pmid, process/pmid, component/pmid, refseq, mRNA.peptide/pmid`
Organisms	identity [binary]	abstract, title, titles of references, paragraph, sentence, preceding five words, subsequent five words	`org/taxname, org/common, org/taxid, org/lineage`

Table 2: Features for candidate scoring

name fields (e.g., `gene/syn`, `prot/name`, `nomenclature/symbol`) and some descriptive fields (`summary` and `generif/text`) with Gene-IDs. We used the reductive tokenization method [2] and Porter stemmer [3] for both queries (gene mentions) and record contents.

When designing this component, we prioritize recall over precision because the subsequent components (candidate scoring and aggregation) cannot recover from misses (false negatives) of candidate retrieval. At the same time, it might be difficult for candidate scoring to choose a true (positive) Gene-ID from a large number of irrelevant (negative) Gene-IDs. Therefore, we introduced some heuristics to reduce the number of candidate Gene-IDs. We discard Gene-IDs that satisfy none of the following conditions:

1. the Gene record includes the gene mention somewhere in its name fields;
2. the matching score (Okapi's BM25) of the Gene record is within the top 20 for the mention;
3. the species of the Gene record is mentioned somewhere in the source article.

Candidate scoring

We score each candidate Gene-ID by using a logistic regression model. Features for the scorer are categorized into four types: *mention-name features*, *context features*, *PMID features*, and *organism features*. A mention-name feature captures orthographic similarity between a gene mention and the name fields. We prepared a mention-name feature for every combination of fields (e.g., `gene/syn`, `prot/name`, `nomenclature/symbol`) and matching methods (e.g., exact match, approximate match, letter n-gram similarity). Context features compute the cosine similarity between the surrounding expressions (context) of a gene mention and descriptions in a candidate Gene record. We designed a context feature for each window of context (abstract, title, titles of references, paragraph, sentence, preceding five words, and succeeding five words) and for each descriptive field. A PMID feature indicates whether the Gene record includes the PMID(s) related to the target paper (the PMID of the paper and PMIDs of the related work) in the reference fields (e.g., `generif/pmid` and `function/pmid`). An organism feature examines whether the species of the Gene record appear in a context window of the mention. Here, we use the results of the species mention recognizer to link taxonomy identifiers (TaxIDs) and context expressions (e.g., TaxID #9606 and the expression *patients*).

In order to train the logistic regression model, we manually annotated the gold-standard mention(s) for each Gene-ID in the training sets 1 and 2. In Gene-IDs enumerated by the candidate generator for each gold-standard mention, the Gene-ID in the training sets presents a positive instance, and the rest presents negative instances. We used Classias[6] as a tool-kit for training the logistic regression model. The score of a Gene-ID is defined to be the probability estimate when the instance is classified into positive.

Candidate aggregation

A paper usually contains multiple mentions of the same gene. In order to obtain confidence scores of Gene-IDs at article level, we compute the sum of scores of Gene-IDs appearing in the paper. After removing Gene-IDs that have score sums lower than 0.1, we compiled and submitted the list of Gene-IDs for the articles.

Acknowledgments

This work was partially supported by Grants-in-Aid for Scientific Research on Priority Areas (MEXT, Japan) and for Solution-Oriented Research for Science and Technology (JST, Japan).

References

1. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J: **Developing a Robust Part-of-Speech Tagger for Biomedical Text**. In *Proceedings of the 10th Panhellenic Conference on Informatics* 2005:382–392.
2. Wermter J, Tomanek K, Hahn U: **High-performance gene name normalization with GeNo**. *Bioinformatics* 2009, **25**(6):815–821.
3. Porter MF: **An algorithm for suffix stripping**. *Program* 1980, **14**(3):130–137.

[6] Classias: http://www.chokkan.org/software/classias/

OntoGene (Team 65): preliminary analysis of participation in BioCreative III

Fabio Rinaldi*[1], Gerold Schneider[1], Simon Clematide[1], Silvan Jegen[1],
Pierre Parisot[2] , Martin Romacker[2] and Therese Vachon[2]

[1] Institute of Computational Linguistics, University of Zurich, Switzerland
[2] NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland

Email: F. Rinaldi*- rinaldi@ifi.uzh.ch; G. Schneider - gschneid@ifi.uzh.ch; S. Clematide - simon.clematide@cl.uzh.ch; S. Jegen - silvan.jegen@novartis.com; P. Parisot - pierre.parisot@novartis.com; M. Romacker - martin.romacker@novartis.com; T. Vachon - therese.vachon@novartis.com;

*Corresponding author

Abstract

Background: The BioCreative series of competitive evaluations of text mining systems provide a major test bed for novel techniques in biomedical text mining. Results from the previous and current competition are of fundamental importance for further development in the area.

Results: The OntoGene group participated in all tasks of the current edition. Preliminary results seem satisfactory, however a detailed analysis cannot be performed without a comparison with the results of the other participants.

Background

OntoGene is a research project based at the Institute for Computational Linguistics of the University of Zurich, focusing on the usage of advanced natural language processing techniques for the purpose of biomedical text mining. Since the beginning of our activities in this domain (2005), our core focus has been on relation extraction [1], rather than on entity extraction.

We participated in the previous two editions of the BioCreative shared evaluation. In BioCreative II (2006) we had the best reported results in the extraction of experimental methods task (PPI-IMT) and very competitive results in the extraction of protein interactions (PPI-IPT) [2]. In BioCreative II.5 (2009) we obtained the best results (according to the 'raw' AUC metric) in the IPT task (extraction of protein interactions) [3, 4].

Due to very recently obtained additional research funding, we decided to increase our effort in the current competition, and participate in all of the tasks on offer. In the rest of this research report we describe in detail our approach to each of the tasks.

RUN 1	RUN 2
Positives: 15101	Positives: 17973
Relevant: 1670	Relevant: 1670
TP: 451	TP: 467
FN: 1219	FN: 1203
FP: 14650	FP: 17506
Recall: 0.2701	Recall: 0.2796
Precision: 0.0299	Precision: 0.0260
Averaged-TAP-5: 0.0718	Averaged-TAP-5: 0.0891
Averaged-TAP-10: 0.0992	Averaged-TAP-10: 0.1073
Averaged-TAP-20: 0.1077	Averaged-TAP-20: 0.1156

Table 1: GN results on the 50-articles evaluation set

Results and Discussion

GN Task

In the GN task we used a variant of the OntoGene text mining system which was previously developed for the detection of protein-protein interactions. While the full OntoGene system includes modules for syntactic parsing and relation extraction, the version used for the GN task included only part of the complete pipeline. The following processing steps are performed: (1) XML cleanup and transformation into our own basic XML format; (2) preprocessing with Lingpipe [5] (sentence splitting, tokenization, tagging); (3) terminology recognition; (4) detection of ‘focus organisms’; (5) terminology filtering and scoring.

The terminology recognition module is based on an efficient lexical lookup approach, with the contribution of a ‘normalization’ module (rule based) which can take into account the most frequent surface variants of a term. The lookup uses an internal terminological resource built using terms extracted from UniProt, Entrez Gene, NCBI Taxonomy, Cell Line Knowledge Base (CLKB). An additional aim of our participation was to test an extensive gene resource provided by TMS (Text Mining Services, Novartis AG, Basel).

One characteristic of our approach is the usage of a specific module for the detection of the ‘focus organism’, i.e. the core specie(s) discussed in the paper. This information is later used for the disambiguation of gene and protein mentions. This module was originally optimized for disambiguation of protein mentions over the set of IntAct ‘snippets’.[1] No further adaptation for the GN task in BC III was performed

We use a terminology filtering and scoring approach, which is based on the one hand on textual features, on the other hand on the detected organism. It functions as follows: for each term for which a focus organism above a probability threshold filter has been identified, and which is not in a stop word list, a score based on frequency of the term, the zone (title, abstract, main text), and organism-related keywords is calculated. Organism-related keywords express e.g. that the presence of the word ‘murine’ gives increased scores to terms related to mouse. The scores and the organism-related keywords were manually adapted to the training documents. Broadly speaking, for each term candidate $SCORE = f * org$, where

f : frequency of term in text (an occurrence in the title has a weight of 200, an occurrence in the abstract a weight of 8; additionally terms in italics are weighted 3 times higher).

org : organism score from “focus organism” detection module (rebalanced through some specific additional organism-related keywords).

The difference between our two submitted runs is mainly in the terminological resources. RUN 1 does not use EntrezGene or UniProt, but instead used an extensive terminological resources provided by TMS (Text Mining Services, Novartis AG), which however covers only the five most important species (human, mouse, rat, yeast and drosophila). Additionally, we included organism resources extracted from the NCBI taxonomy and terms from the CLKB. The TMS resource contains 670,000 term senses. Our own organism and CLKB resource contains 49,000 term senses. This resulted in 520,000 normalized terms, and 172,000 different gene IDs from 5 different organisms.

RUN 2 additionally used 2,203,000 terms from UniProt (version from June 2010) and 1,021,000 terms from EntrezGene (only 20 topmost organisms from the training data, for efficiency reasons). This resulted in 1,856,000 normalized terms and 833,000 different gene IDs from 2,113 different organisms.

[1] A snippet is a short textual reference provided by the IntAct curators.

ACT	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
TP	351	539	756	648	475
FP	120	353	1823	1317	285
FN	559	371	154	262	435
TN	4970	4737	3267	3773	4805
sensv.	0.38571	0.59231	**0.83077**	0.71209	0.52198
specf.	**0.97642**	0.93065	0.64185	0.74126	0.94401
accur.	**0.88683**	0.87933	0.67050	0.73683	0.88000
Matthew	0.48297	**0.52727**	0.34244	0.34650	0.50255
P at full R	**0.16189**	**0.16189**	0.15182	0.15182	0.15660
AUC iP/R	0.63847	**0.63890**	0.41741	0.41740	0.62394

Table 2: PPI-ACT Performance: specf (specificity), sensv (sensitivity), accur (accuracy).

The results obtained on the 50-articles set released by the organizers after the end of the competition are shown in table 1. Not having seen the results by other teams, the only conclusion which we can draw at present is that the resource used for RUN 1 appears to be sufficiently complete, in comparison with the subset of EntrezGene used for RUN 2. In fact, in RUN 2 we have an increase of only 16 TP (+3.5%), which is small compared with the increase of 2,856 FP (+19.5%). Unexpectedly, the TAP-k measures are definitely better for RUN 2. This would suggest that RUN 2 produced a better ranking than RUN 1. A possible explanation for this difference is that the contribution of the “focus organism” detection module is better in RUN 2 than in RUN 1 (therefore genes belonging to the selected organisms are ranked higher). Our “focus organism” module [6] was initially developed for PPI detection. In order to derive an organism ranking it uses all relevant terminology in the article: in particular terms from NCBI and CLKB, but also proteins mentions. Crucially however, it does not use gene mentions to the same extent as protein mentions (in retrospect, we should have adapted it to the nature of the competition). Therefore the lack of sufficient protein mentions in RUN 1 produced a lower quality ranking of organism, which in turn resulted in a worse ranking for genes.

On the set of the 50 most difficult articles, we reached an unweighted average TAP-20 of 0.07 for RUN 1. On the training data we had reached an unweighted TAP-20 of 0.3453. The low results for the 50 articles set are mostly due to the fact that only 103 gene IDs out of 1,219 false negatives were available in this resource. For RUN 2, we had 1,203 false negatives. However also here, only 335 gene IDs were available in our resource. On the training data we had reached an unweighted TAP-20 of 0.3751.

PPI-ACT Task

Three of the runs were generated applying Maximum Entropy optimization (specifically the software package ‘MEGAM’ [7]). Features considered include lexical items in the document (+Bow),[2] MeSH annotations (+Mesh),[3] and a score delivered by our PPI detection pipeline (+PPIscore)[4] . Two runs (RUN 3 and RUN 4) used only the result of the PPI pipeline. The development set proved to be representative for the testset.

The feature weights used for the test set were drawn from the development set only. Including the balanced (but therefore biased) training set (which was released earlier in the shared task) proved to detoriate the results in a 10-fold cross-validation experiment on the development set. Using the bow and mesh features, we get a huge number of features. In order to keep the training efficient, and to prevent over-training, each

[2] All words of the articles were stemmed. Than all counts of a stem were used as a feature. E.g, if the word ”protein” was found 3 times, we produced the features ”protein_1”, ”protein_2”, ”protein_3”. This produced for instance 70886 different features for the development set.

[3] Every MeSH descriptor, with and also without every qualifier, was used as a feature. E.g., for the MeSH term ”-Signal Transduction (-drug effects; +physiology)” as it appeared in the textual format, we produced the descriptor features ”signal/transduction/drug/effects”, ”signal/transduction/physiology”. For multi word terms, we added also all descriptor terms produced by iteratively removing the first word, for instance ”transduction”. Additionally, all MeSH qualifiers as ”-drug/effects” and ”+physiology” were added.

[4] This feature is computed using the full pipeline for detection of PPI as used in the BioCreative II.5 challenge. The original system is used to detect candidate interactions, and deliver each of them, together with a numerical score. This value was discretized in order to form few large classes and then used as a feature set.

IMT	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
Evaluated Results	5098	21529	4576	666	21600
TP	447	527	431	223	527
FP	4651	21002	4145	443	21073
FN	80	0	96	304	0
Micro P	0.08768	0.02448	0.09419	**0.33483**	0.02440
Micro R	0.84820	**1.00000**	0.81784	0.42315	**1.00000**
Micro F	0.15893	0.04779	0.16892	**0.37385**	0.04763
Micro AUC iP/R	0.27588	0.24484	0.27727	0.14169	**0.29016**
Macro P	0.09346	0.02448	0.09992	**0.33483**	0.02440
Macro R	0.83206	**1.00000**	0.79377	0.42883	**1.00000**
Macro F	0.16322	0.04750	0.17163	**0.35403**	0.04735
Macro AUC iP/R	0.47884	0.44034	0.47650	0.30927	**0.50111**

Table 3: PPI-IMT Performance

feature had to appear at least 3 times in the development set, and additionally, the feature selection limitation of MEGAM was used to allow not more than 20,000 features. The resulting features are distributed as follows: 69% bow, 31% mesh.

RUN 1, as expected was the run with the highest accuracy (see table 2). Specificity was deliberately maximized at the cost of sensitivity because of the class imbalance. The features used were +PPIscore, +Mesh, +Bow with standard class binarization of MEGAM at 0.5 between classes 0 and 1. RUN 2 was aimed at maximizing Matthew's correlation coefficient. It is also the run with the highest AUC. The features used were +PPIscore, +Mesh, +Bow with lowered binarization threshold of MEGAM at 0.2 between classes 0 and 1, in order to boost the positive class (the threshold was determined heuristically on the basis of the development set). RUN 3 was aimed at maximizing recall (without using maximum entropy optimization). The 'raw PPIscore' was discretized as follows: if $PPIscore > 0.2$ then class=1 else class=0. RUN 4 was aimed at a balanced specificity / sensitivity result. It did not use the maximum entropy approach, but only the raw PPIscore with the following decision rule: if $PPIscore > 1.1$ then class=1 else class=0. RUN 5 used only the +Bow and +Mesh features, with lowered binarization of MEGAM at 0.25 between classes 0 and 1, in order to obtain the best Matthew's coefficient (threshold determined by experimentation on the development set). The comparison with RUN 2 is particularly interesting because it shows the impact of the +PPIscore feature: we gain 64 TP, but also get 68 more FP.

We have made the following observations. First, the class imbalance negatively affects the recall of the smaller class (1), because the classifier optimizes for overall accuracy. One way to improve the high recall results might be to use the several subscores that make up PPIscore (for example syntactic path, word at the top of the path, protein pair salience, zoning information, etc.) as fine-grained individual features, whose weights can also be optimized individually.

PPI-IMT Task

For the PPI-IMT detection task, we have developed two statistical systems (called system A and system B in this document). Both are based on a naive Bayes approach but use different optimizations and heuristics. The submitted runs correspond to the following:

RUN 1: full output of system A
RUN 2: full output of system B
RUN 3: optimized output of system A
RUN 4: optimized output of system B
RUN 5: combined output (average scores of RUN 1 and RUN 2)

The full outputs were aimed at maximizing R and AUC, the optimized outputs at maximizing F-score. We have avoided sending runs which optimize precision, because these can always be obtained by picking for each article only the best prediction (i.e. the method which is ranked first). [8] reports that the curators preferred a high recall setting to a high precision setting, because it is much easier and less time-consuming to reject suggestions (false positives, low precision) than to add new information from scratch (false negatives,

$p(method\|word)$		
Probability	Word	Method
0.490056	L1	MI:0006
0.470270	LT	MI:0019
0.447269	ERK1/2	MI:0006
0.443877	hydrogen-bonding	MI:0114
0.441441	omit	MI:0114
0.438765	synapses	MI:0006
0.436363	tumours	MI:0006
0.435114	REFMAC	MI:0114

Table 4: Statistical association of methods with specific words (examples)

low recall). A good ranking, coupled with good recall, allows the user to decide where to stop examining the results, rather than leaving the decision to the system.

RUN 5 was a blind experiment - due to lack of time we did not try this combination on development and training sets. It is interesting to notice that this RUN achieves the best AUC (50%), while maintaining full recall (like RUN 2). The preliminary conclusion appears to be that system A produces a better ranking, which, when combined with the more complete output of system B, results in a better AUC. While system A has been specifically optimized for the IMT task with task-specific heuristics, system B provides a fairly generic implementation of a naive Bayes multiclass classifier, which therefore does not need a very detailed description. In the rest of this section we provide more information about System A.

As a first approach, we used a pattern matcher giving high scores to every occurrence of an exact match, and lower scores to every occurrence of a word-submatch, using the PSI-MI dictionary of experimental methods [9] as our standard. No 'stop word' list was used, except for removing the prepositions *of* and *in* which occur in many terms and synonyms. The inclusion of submatches led to overgeneration (increased recall but low precision). Using only full matches led to very low recall. As an intermediate level between full match and word-based submatch, we also used a subset approach: if more than three words of a term or a synonym from the PSI-OBO dictionary appear in a ten word observation window, a mid-range score is given for each occurrence. We observed that some submatch words are contained in many different experimental methods (they do not discriminate well) and at the same time many submatch words very often do not indicate a method mention. For example, method 0231 has the term name *mammalian protein interaction trap*, which means that every occurrence of the word *protein* assigns a score to this method.

To respond to these observations, a statistical method can be used. We use, on the one hand conditional probabilities for the method given a word $p(method \mid term\ word)$ and, on the other hand the conditional probability that a given submatch word occurs in a document where the corresponding term identifier has been effectively assigned by the annotator: $p(term\ word\ =\ yes \mid word,\ document)$. We informally refer to the latter probability as termness. We first use the statistical model, $p(method \mid word)\ *\ termness(word)$ for all words that are matches or submatches of the terms given in the PSI-MI dictionary, and further for all words, irrespective of whether they appear in the PSI dictionary, whenever $p(method|word)$ and $termness(word)$ are above 10%, and whenever the word is used in at least 5 training documents. We have obtained considerably better results when using the statistical model also on all words, including non-term words. The lists containing words which have a high probabilities to be associated with a given method are not obviously interpretable by the non-expert, although some of the inherent knowledge they contain are clear hints. An excerpt of frequent words indicating experimental methods at high probability is given in table 4.

IAT Task

The ODIN system is being developed within the scope of the OntoGene project, as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group (Text Mining Services) of Novartis Pharma AG. The purpose of the system is to allow a human annotator/curator to leverage upon the result of a text mining system in order to enhance the speed and effectiveness of the annotation process.

The OntoGene system takes as input a document in plain text or a number of supported xml-based formats (including PubMed Central) and processes it with a custom NLP pipeline, which includes Named Entity recognition and relation extraction. Entities which are currently supported include proteins, genes, experimental methods, cell lines, species. Entities detected in the input document are disambiguated with respect to a reference database (UniProt, EntrezGene, NCBI taxonomy, PSI-MI ontology).

The annotated documents are handed back to the ODIN interface (as pure XML documents), which allows multiple display modalities, plus various selection and modification options. The curator/annotator can view the whole document with in-line annotations highlighted, or can browse the extracted entities and be pointed back to the mentions of the entities within the original document. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change its extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences) containing desired entities (terms or gene identifiers) is supported. Rapid disambiguation can be achieved through manual organism selection. Additionally, extensive logging functionalities are provided. The curation interface is mainly developed as a JavaScript-based web application using the extjs framework. This allows rapid prototyping of views (tables, highlighting, creation of hyperlinks). Visualization is very flexible through CSS and DOM manipulation.

Acknowledgements

The OntoGene group is partially supported by the Swiss National Science Foundation (grants 100014 − 118396/1 and 105315 − 130558/1) and by NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland.

References

1. Rinaldi F, Schneider G, Kaljurand K, Hess M, Romacker M: **An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA**. *BMC Bioinformatics* 2006, **7**(Suppl 3):S3, [http://www.biomedcentral.com/1471-2105/7/S3/S3].
2. Rinaldi F, Kappeler T, Kaljurand K, Schneider G, Klenner M, Clematide S, Hess M, von Allmen JM, Parisot P, Romacker M, Vachon T: **OntoGene in BioCreative II**. *Genome Biology* 2008, **9**(Suppl 2):S13, [http://genomebiology.com/2008/9/S2/S13].
3. Schneider G, Kaljurand K, Kappeler T, Rinaldi F: **Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources**. In *Proceedings of CICLING 2009* 2009.
4. Rinaldi F, Schneider G, Kaljurand K, Clematide S, Vachon T, Romacker M: **OntoGene in BioCreative II.5**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7**(3):472–480.
5. Alias-i: **LingPipe**. [http://alias-i.com/lingpipe].
6. Kappeler T, Kaljurand K, Rinaldi F: **TX Task: Automatic Detection of Focus Organisms in Biomedical Publications**. In *Proceedings of the BioNLP workshop, Boulder, Colorado* 2009.
7. Daumé III H: **Notes on CG and LM-BFGS Optimization of Logistic Regression** 2004. [Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/].
8. Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, Roebuck S, Tobin R, Wang X: **Assisted Curation: Does Text Mining Really Help**. In *BIOCOMPUTING 2008. Proceedings of the Pacific Symposium on Biocomputing.* Edited by Altman RB, Dunker AK, Hunter L, Murray T, Klein TE, Kohala Coast, Hawaii, USA 2008[http://psb.stanford.edu/psb-online/proceedings/psb08/alex.pdf].
9. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG C Sander, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, R A: **The HUPO PSI's molecular interaction format - a community standard for the representation of protein interaction data**. *Nat. Biotechnol* 2004, **22**:177–183.

GeneView – Gene-Centric Ranking of Biomedical Text

Philippe E. Thomas*[1], Johannes M. Starlinger[1] , Christoph Jacob[1] , Illés Solt[1,2], Jörg Hakenberg[3] and Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany,[2]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary,[3]Computer Science Department, Arizona State University, Tempe, AZ 85287, USA

Email: Philippe E. Thomas*- thomas@informatik.hu-berlin.de; Johannes M. Starlinger - starling@informatik.hu-berlin.de; Christoph Jacob - christoph.jacob@informatik.hu-berlin.de; Illés Solt - solt@tmit.bme.hu; Jörg Hakenberg - joerg.hakenberg@asu.edu; Ulf Leser - leser@informatik.hu-berlin.de;

*Corresponding author

Abstract

Background: Life scientists spend a great amount of time searching for gene-specific information. It is widely acknowledged that research results are primarily published in scientific literature and current curation efforts can not keep up with the fast increase of such literature. It can therefore be estimated that the plethora of gene-specific knowledge is still hidden in large text repositories like MEDLINE. Searching text data sources is difficult, as user queries are usually ambiguous and lead to hundreds of results. Faced with such a number of relevant publications, an appropriate article ranking is important. PubMed, for example, ranks articles per default by indexing date, making it difficult to find seminal papers about a specific topic. In this paper, we introduce GeneView, a gene-centric text mining application capable of searching, ranking, and visualizing biomedical publications.

Results: Our ranking algorithm relies on the assumption that the relevance of a gene for a specific article depends on the frequency with which it is mentioned and on the sections it appears in. For ranking we introduce a simple evaluation strategy by using the NCBI Gene2Pubmed mapping as gold-standard. This strategy is used to evaluate different section specific rankers, where the best one achieves on average a precision of 75.5 %.The evaluation further confirms our expectations, that sections like *title*, *abstract* and *result* are more relevant for gene specific ranking than others. Surprisingly, incorporation of figure- and table-captions decreased the quality of ranking results.

Background

Life scientists spend a great amount of time searching for gene-specific information. As current approaches for gene function annotations are clearly not keeping up with the double-exponential increase of biomedical literature [1, 2], such a search cannot be restricted to structured databases, but must also include scientific publications, either using abstract repositories such as MEDLINE or directly accessing full text articles from publishers web sites.

Searching for biological concepts like genes in full text is difficult for various reasons. First, genes usually have many synonyms. Thus, a simple query with a single gene name will return only a subset of all relevant articles. Second, homologous genes often share the same name across different species.

Also, even non-homologous genes might share names among each other, and genes might also have names that have additional, totally unrelated meanings. For example, the human *EGFR* gene has several aliases, like *ERBB* or *HER1*. But the gene *EGFR* also exists, under the same name, in other species like *rat*, *mouse* or *fruit fly*. Finally, EGFR also is an abbreviation for "estimated glomerular filtration rate". Thus, a simple PubMed query for "EGFR" leads to rather unspecific results. Even if a researcher manages to define an appropriate query, e.g. by incorporating synonyms and excluding non-gene related keywords, often hundreds of papers are reported. Faced with such result sizes, an appropriate article ranking is important. PubMed, for example, ranks articles per default by indexing date, making it difficult to find seminal papers about a specific topic. Such ranking is especially important for database curators to decide which articles to read first to find authoratative and comprehensive information about a given concept.

In this paper, we introduce GeneView, a gene-centric text mining application capable of searching, ranking, and visualizing biomedical publications. Some of its salient features include the possibility to personalize the behavior of the ranking, the pre-tagging of gene mentions using a state-of-the-art gene name tagger, and the application of species-dependent disambiguation. GeneView takes part in the "Interactive Demonstration Task for Gene Indexing and Retrieval" (IAT) of BioCreative III (BC3).

Methods

GeneView consists of several inter-operating components: 1.) named entity recognition modules for genes, mutations, diseases, and drugs 2.) an inverted index for efficient searching 3.) a customized ranking algorithm taking gene-centric information into account, and 4.) a web frontend for querying and visualization. These components are further described in this section.

Preprocessing of Data

GeneView in its application for BC3 searches a full text corpus of 17,780 PubMed-Central (PMC) articles provided as XML files. These files contain, among other things, information about authors, publishing journal, full text, and figure and table captions. Parsed full texts are stored in an inverted index using the open source system Lucene[1], serving as storage, query and ranking engine. Performing text queries is no requirement for the IAT task, but it proved to be useful in executing more detailed queries than just gene names. Further, it is required for our ranking algorithm (see below).

After extracting plain text from the PMC XML files, gene names are identified and grounded to Entrez Gene-Id's using the latest version of GNAT [3]. This version of GNAT has been improved to more efficiently deal with full texts and to allow for a more general species-specific disambiguation of gene names. We further tag single nucleotide polymorphisms (SNP), using a slightly modified version of MutationFinder [4]; improvements encompass additional rules for mutation recognition and an expanded amino acid dictionary, which allows detection of ambiguous amino acid descriptions. For recognizing diseases and drug names, we generate dictionaries from UMLS [5], DrugBank [6] and PharmGKB [7] and perform tagging with a simple dictionary lookup.

All recognized entities are associated with the respective section they were found in, allowing us to base the ranking on different weights for occurrences depending on the section. Such a feature was proven useful in several previous works [8,9]. As most journals have their own guidelines about section naming, we normalize sections to common section names. For example, the section name *background* is normalized to *introduction*. Sections like *author contributions* or *funding* are normalized to the "catch all" section type *others*. For section names not in the dictionary we employ a fuzzy matching, by searching the dictionary for the entry with smallest edit distance. If the ratio between distance and length of the term is below a certain threshold, the term is normalized to the corresponding section. Section names failing this procedure again are normalized to the section *others*.

Entities (genes, SNPs, etc.) are added to the Lucene index, together with their common section name and their entity type, allowing for a very quick search. The index is also used by the section specific ranker for the *retrieval* task and for full text visualization in the *indexing* task.

[1] http://lucene.apache.org/; accessed 09/01/2010

Indexing

The *indexing* task of BC3 addresses the visualization of one selected full text article and the subsequent ranking of all genes it contains. As the importance of a gene for ranking depends on a specific user's needs, GeneView allows users to personalize the ranking to some degree. Per default, genes are ranked by total occurrence in the article, but users have the possibility to exclude sections from this calculation. Figure 1 shows a GeneView screen-shot for a full text article with default gene ranking on the left hand side. Rankings can be modified using the "adjust ranking" function. Found entities are highlighted using different colors and entity-specific information integrated from a number of public data sources is provided on mouse-click. In future work, we also plan to open further ranking information for personalization (see Conclusions).

Retrieval and Ranking

Goal of the *retrieval* task is the proper ranking of articles providing information for one specific gene. It has been previously supposed that the relevance of a gene for a specific article depends, among other things, on the frequency with which it is mentioned and on the sections it appears in [10]. For example, articles mentioning a gene in *title* or *abstract* are often considered to be more relevant than articles describing the gene only in *methods* or *discussion*. To emphasize relevant sections we employ a section specific boosting. The so-ranked list of articles is presented to the user in return for their query. Each entry can then be selected for full text viewing.

GeneView also computes and displays associated genes. To this end, the system identifies all genes co-occurring with a given query gene in any of the articles in the corpus. Each such gene is tested for positive association using a single sided χ^2-test. The five most significantly associated entities are than displayed in GeneView at the top of the search results page. Although several tools like LitMiner [11], EBIMed [12], SciMiner [13], or FACTA [14] are capable of determining query associated entities, only few are able to perform such a query in near real time [14]. Our current implementation processes such a query in approximately 1 second. Application on the whole of PubMed, with more than 20 million citations, takes approximately 7 seconds.

Results and Discussion

The total number of entities found in the BC3 corpus are depicted in Table 1. Almost all articles contain at least one gene, and, on average, an article contains 72 gene mentions. The results of section mapping in Table 2 show that our approach is able to normalize 111,695/114,204 = 97.80 % of all section names. Note that about 20 % of all provided XML files contained only title and abstract. These publications are only available as PDF or image and the full text has not been transcribed into the XML files.

Entity Type	Articles	Entities
Genes	16,013	1,294,875
Drugs	15,102	535,707
Diseases	12,166	940,225
SNPs	4,846	91,410

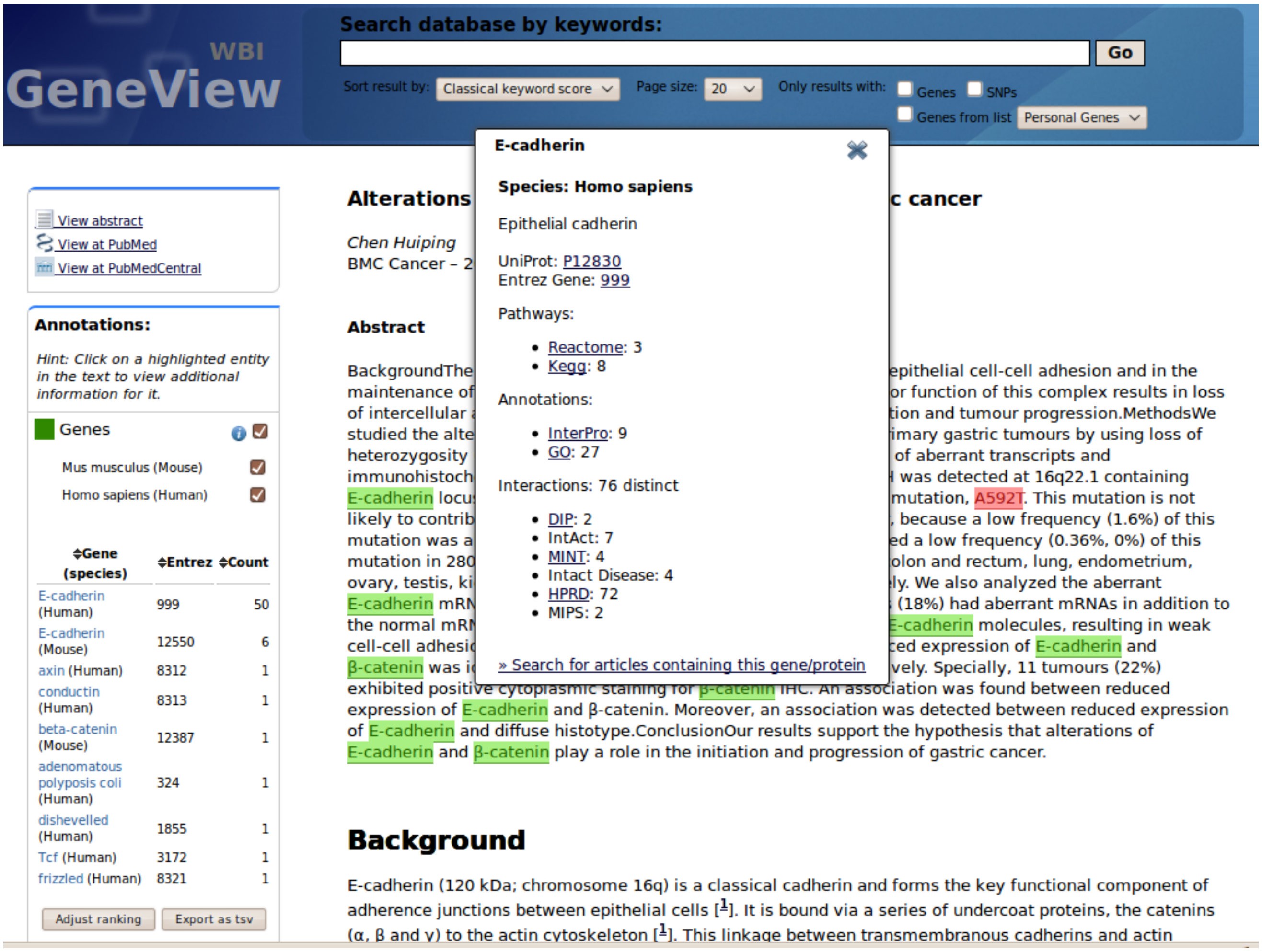
Table 1: Number of articles containing a specific type of entity and the number of entity occurrence in the IAT corpus (17,780 articles).

Normalized Section	Number
Title	17,780
Abstract	17,780
Results	13,879
Methods	13,650
Introduction	13,140
Discussion	12,255
Other	12,228
Conclusion	5,063
Supplement	3,836
Not normalized	2,503
abbreviations	2,088

Table 2: Frequency of normalized section headers.

Both IAT subtasks are intended as groundwork for the future evaluation of interactive systems in BioCreative IV. No gold-standard data was provided for any of the tasks. However, we were curious about the impact of our ranking algorithm and therefore defined our own simple (yet quite fuzzy) evaluation strategy by using the NCBI Gene2Pubmed [2] mapping as gold-standard. This mapping provides links between PubMed articles and Entrez-Gene Id's for

[2] ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz; accessed 09/01/2010



Search database by keywords:

Sort result by: Classical keyword score | Page size: 20 | Only results with: □ Genes □ SNPs □ Genes from list Personal Genes

View abstract
View at PubMed
View at PubMedCentral

Annotations:

Hint: Click on a highlighted entity in the text to view additional information for it.

🟩 Genes ☑

Mus musculus (Mouse) ☑
Homo sapiens (Human) ☑

Gene (species)	Entrez	Count
E-cadherin (Human)	999	50
E-cadherin (Mouse)	12550	6
axin (Human)	8312	1
conductin (Human)	8313	1
beta-catenin (Mouse)	12387	1
adenomatous polyposis coli (Human)	324	1
dishevelled (Human)	1855	1
Tcf (Human)	3172	1
frizzled (Human)	8321	1

E-cadherin

Species: Homo sapiens

Epithelial cadherin

UniProt: P12830
Entrez Gene: 999

Pathways:

- Reactome: 3
- Kegg: 8

Annotations:

- InterPro: 9
- GO: 27

Interactions: 76 distinct

- DIP: 2
- IntAct: 7
- MINT: 4
- Intact Disease: 4
- HPRD: 72
- MIPS: 2

» Search for articles containing this gene/protein

Alterations **c cancer**

Chen Huiping
BMC Cancer – 2

Abstract

BackgroundThe epithelial cell-cell adhesion and in the
maintenance of or function of this complex results in loss
of intercellular tion and tumour progression.MethodsWe
studied the alte imary gastric tumours by using loss of
heterozygosity of aberrant transcripts and
immunohistoch was detected at 16q22.1 containing
E-cadherin locu mutation, A592T. This mutation is not
likely to contrib , because a low frequency (1.6%) of this
mutation was a ed a low frequency (0.36%, 0%) of this
mutation in 280 olon and rectum, lung, endometrium,
ovary, testis, ki ly. We also analyzed the aberrant
E-cadherin mRN (18%) had aberrant mRNAs in addition to
the normal mRN E-cadherin molecules, resulting in weak
cell-cell adhesi ced expression of E-cadherin and
β-catenin was i vely. Specially, 11 tumours (22%)
exhibited positive cytoplasmic staining for β-catenin IHC. An association was found between reduced expression of E-cadherin and β-catenin. Moreover, an association was detected between reduced expression of E-cadherin and diffuse histotype.ConclusionOur results support the hypothesis that alterations of E-cadherin and β-catenin play a role in the initiation and progression of gastric cancer.

Background

E-cadherin (120 kDa; chromosome 16q) is a classical cadherin and forms the key functional component of adherence junctions between epithelial cells [1]. It is bound via a series of undercoat proteins, the catenins (α, β and γ) to the actin cytoskeleton [1]. This linkage between transmembranous cadherins and actin

Figure 1: Screen-shot of GeneView presenting PubMed-Central article 60969. A ranked list of all recognized genes is shown in the left-hand panel. On the right, the full text is displayed with colored markup for recognized entities. The tool-tip shows details on a specific gene and provides links to sources of further information on this gene.

articles being considered as relevant by describing the gene. We performed an evaluation on a subset of 10 randomly selected genes, where each gene on average had 49 associated articles in Gene2Pubmed. For each gene the top 20-results were computed using our system and compared to the complete set of associated articles. Retrieved articles among these top-20 with a corresponding entry in Gene2Pubmed are counted as true positives, all others as false positives. As we investigate only the top 20 results, calculation of false negatives is not applicable. We also used this evaluation to search for optimal section-specific boosts.

Finally, we selected the best combination from approximately 2,000 different settings. This setting has, on average, a precision of 75.5 %, whereas queries without section boosts achieves 72.0 %. The best setting, as shown in Table 3, reflects our expectations, that sections like *title*, *abstract* and *result* are more relevant for gene specific ranking than others. Surprisingly, incorporation of captions decreased ranking results.

Section	Boost
Title	3.0
Abstract	2.0
Introduction	1.0
Result	2.0

Table 3: Section boosts yielding the best results in our evaluation. Sections not mentioned were excluded from the query.

We also tested the quality of associated genes using the full text query "colorectal cancer". For four of the five reported genes, the association to the query is well known in the literature. The fifth gene (Scrib/Entrez-Id: 105782, $p = 5 \cdot 10^{-20}$) reveals a mistake of GNAT, which commonly identifies "CRC" as a synonym for this gene. But in fact, CRC is also often used as an abbreviation for "colorectal cancer"; thus, the token is positively associated with the query, but not the gene reported.

Conclusions

We described and evaluated GeneView, a tool for gene-centric searching, ranking, and visualizing scientific full text articles. We also reported on a very preliminary evaluation which seem to confirm our expectation that section-specific boosting is beneficial for relevance ranking. We are currently also exploring other ways of improving the ranking by concept-based query expansion with associated terms [15]. This query expansion currently focuses on Gene Ontology terms [16], as such terms are associated with the function of genes/proteins. However, a first and naive implementation, using all associated GO terms as query phrases, lead to a minimal decline of 1 % in article ranking. We therefore plan to deeper investigate this expansion approach; furthermore, we think about including centrality of a gene in the document-specific gene network into the ranking. All these features should be turnable by each user.

Acknowledgements

GeneView is developed as part of the ColoNet project, supported by the German Federal Ministry of Education and Research, grant no 0315417B.

References

1. Hunter L, Cohen KB: **Biomedical language processing: what's beyond PubMed?** *Mol Cell* 2006, **21**(5):589–594.
2. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L: **Manual curation is not sufficient for annotation of genomic databases.** *Bioinformatics* 2007, **23**(13):i41–i48.
3. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics* 2008, **24**(16):i126–i132.
4. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics* 2007, **23**(14):1862–1865.
5. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32**(Database issue):D267–D270.
6. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic Acids Res* 2008, **36**(Database issue):D901–D906.
7. Thorn CF, Klein TE, Altman RB: **Pharmacogenomics and bioinformatics: PharmGKB.** *Pharmacogenomics* 2010, **11**(4):501–505.
8. Schuemie MJ, Weeber M, Schijvenaars BJA, van Mulligen EM, van der Eijk CC, Jelier R, Mons B, Kors JA: **Distribution of information in biomedical abstracts and full-text publications.** *Bioinformatics* 2004, **20**(16):2597–2604.

9. Hakenberg J, Rutsch J, Leser U: **Tuning Text Classification for Hereditary Diseases with Section Weighting**. In *Proc International Symposium on Semantic Mining in Biomedicine, SMBM*, Hinxton, UK 2005:34–37.

10. Hakenberg J, Leaman R, Vo NH, Jonnalagadda S, Sullivan R, Miller C, Tari L, Baral C, Gonzalez G: **Efficient extraction of protein-protein interactions from full-text articles.** *IEEE/ACM Trans Comput Biol Bioinform* 2010, **7**(3):481–494.

11. Maier H, Döhr S, Grote K, O'Keeffe S, Werner T, de Angelis MH, Schneider R: **LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W779–W782.

12. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed–text crunching to gather facts for proteins from Medline.** *Bioinformatics* 2007, **23**(2):e237–e244.

13. Hur J, Schuyler AD, States DJ, Feldman EL: **SciMiner: web-based literature mining tool for target identification and functional enrichment analysis.** *Bioinformatics* 2009, **25**(6):838–840.

14. Tsuruoka Y, Tsujii J, Ananiadou S: **FACTA: a text search engine for finding associated biomedical concepts.** *Bioinformatics* 2008, **24**(21):2559–2560.

15. Matos S, Arrais JP, Maia-Rodrigues J, Oliveira JL: **Concept-based query expansion for retrieving gene related publications from MEDLINE.** *BMC Bioinformatics* 2010, **11**:212.

16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.

Gene mention normalization in full texts using GNAT and LINNAEUS

Illés Solt[1,2], Martin Gerner[3], Philippe Thomas[2], Goran Nenadic[4], Casey M. Bergman[3], Ulf Leser[2], Jörg Hakenberg[5§]

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary
[2] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[3] Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK
[4] School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
[5] Computer Science Department, Arizona State University, Tempe, AZ 85287, USA
[§] Corresponding author

Email addresses:
IS: solt@tmit.bme.hu
MG: martin.gerner@postgrad.manchester.ac.uk
PT: thomas@informatik.hu-berlin.de
GN: g.nenadic@manchester.ac.uk
CB: casey.bergman@manchester.ac.uk
UL: leser@informatik.hu-berlin.de
JH: joerg.hakenberg@asu.edu

Abstract

Gene mention normalization (GN) refers to the automated mapping of gene names to a unique identifier, such as an NCBI Entrez Gene ID. Such knowledge helps in indexing and retrieval, linkage to additional information (such as sequences), database curation, and data integration. We present here an ensemble system encompassing LINNAEUS for recognizing organism names and GNAT for recognition and normalization of gene mentions, taking into account the species information provided by LINNAEUS. Candidate identifiers are filtered through a series of steps that take the local context of a given mention into account. On the BioCreative III high-quality training data, our system achieves TAP-5 and TAP-20 scores of 0.36 and 0.41, respectively. On the evaluation set of 50 documents that were provided to participants, we achieve scores of 0.16 and 0.20 for TAP-5 and TAP-20, respectively. Our analysis of the evaluation results suggests that the lower scores primarily are due to significant differences in species composition, and partly due to the method for selecting the evaluation data.

Background

BioCreative is a repeated community challenge addressing various tasks in biomedical text mining, such as named entity recognition (NER) of gene and protein names, extraction of protein-protein interactions, or protein interaction detection methods. In the fourth installment in 2010, one of the tasks addressed the recognition and normalization of gene and protein names in full text publications. Participants of this task had to provide a system capable of finding all mentions of genes or proteins in a full text article and of mapping these mentions to their respective Entrez Gene identifiers. Challenges arise from both synonymity and homonymity. Genes frequently have multiple synonyms, usage of which differs not only between authors and journals [1], but also over time. Names often also are used for several different genes (including orthologs, paralogs or unrelated genes) or even for concepts belonging to completely different semantic classes. Developing systems that overcome these challenges is critical for advancing the application of gene mention normalization in biomedical text mining.

Methods

System overview

Our processing pipeline begins by loading the collection of texts that should be annotated, after which we perform NER of species, Gene Ontology (GO) [http://www.geneontology.org/] and MeSH [http://www.nlm.nih.gov/mesh/] terms. We then use species-specific GNAT [2] gene NER modules to find gene name matches in the texts. These modules consist of combined Entrez Gene and UniProt gene name dictionaries, expanded with typical patterns of gene name variations [5]. The recognized gene mentions are assigned candidate identifiers according to the dictionary. The gene mentions are processed by a set of rule-based methods designed to filter out and score candidate identifiers, based on their syntactic and semantic context [2]. Species disambiguation of gene mentions is done by considering the local findings of species NER. Finally, gene mentions with confidence scores above a threshold are reported.

Using LINNAEUS for species NER

In order to identify the species that are discussed in a paper (which in turn determines what genes to search for), we utilize LINNAEUS [3]. LINNAEUS uses a dictionary of expanded species terms from the NCBI taxonomy, together with a variety of rule-based methods and distributional statistics to disambiguate ambiguous species mentions and reduce the number of false positives and negatives. Compared against a corpus of 100 full-text articles manually annotated for species names, LINNAEUS achieves 94% precision and 97% recall [3]. It has previously been shown that for articles linked to genes in Entrez Gene, LINNAEUS can find the species of the referenced gene in 94% (9,662/10,290) of cases where full-text was available [3].

In order to further increase the utility of LINNAEUS for detecting focus organisms of articles, even if they are not mentioned directly, we have included additional “proxy” dictionaries that link cell-lines and genera to corresponding species. The cell-line dictionary was created from the database of [4]. Genera are also tagged and linked to the member species that is most commonly mentioned in MEDLINE (for example, “Drosophila” is linked to *Drosophila melanogaster*).

Some technical re-linking of species identifiers was also necessary due to recent changes in species associations in Entrez Gene. For example, all genes that previously were linked to *Saccharomyces cerevisiae* (NCBI Taxonomy ID 4932) were instead linked to a specific strain, *S. cer. S288c* (ID 850287). This was performed for all species where we could determine that such changes had occurred in Entrez Gene.

Filtering gene names and candidate identifiers

Dictionary-based matching allows direct assignment of candidate identifiers to recognized gene mentions, based on what dictionary entries a mention matches. In a series of filtering steps, the set of mention candidate identifiers is narrowed down successively by removing false positive gene IDs and species IDs (see Table 1 for the full list, and [5] for further details). Filtering includes:

1. Use of the sentence and paragraph context surrounding the mention. The context is matched against pre-computed gene profiles and scanned for clues indicating the presence of false positives.
2. Use of species name mentions located close to the gene mention, that are used to perform cross-species disambiguation.
3. String similarity searches of the located term against the original (not expanded) terms for the candidate identifiers, which are used to determine the closest (and most distant) matches.

Table 1. List of processing filters. Filtering steps are used to expand and reduce candidate ID lists for each gene mention. Also see Figure 1.

Filter	Filtering method
MDRER	Species-dependent gene NER
REU	Joins overlapping or adjacent gene names
LRCF	Match the text surrounding the mention against context models of FPs
ICF	Filter false positives by immediate context
loadGR	Load the gene profile for each candidate gene
UNF	Filter names that refer to gene families and other un-specific mentions
NVF	Restore names removed during UNF where a synonym is used elsewhere
AF	Score mentions by string similarity against unexpanded gene synonyms
SVF	Verify ambiguous species names ("cancer")
UMF	Mark genes that are unambiguous throughout the text as identified
MSDF	Gene mention disambiguation by context profile
ITF	Adjust mention scores based on whether the terms have been found italicized in other PubMed Central articles
SCSA	Assign relative scores to candidates per text
SCSF	Adjust scores to fit the TAP scoring scheme

Scoring candidate identifiers using context profiles for disambiguation

Gene mention disambiguation in our system is handled by an adaptation of GNAT [2]. Adjustments include: (i) more localized reliability scoring of candidate identifiers using paragraph contexts; (ii) keeping annotations consistent across paragraphs; and (iii) text-wide search for the best evidence to map a gene mention to a species.

Selecting the set of species-specific dictionaries

Due to memory constraints, the gene name dictionaries used by GNAT are restricted to a set of model organisms. The selection of what species to include is critical since it determines the species for which GNAT can recognize gene names. The species were chosen based on mention frequencies in MEDLINE and PubMed Central, to cover the majority of articles discussing particular species. In total, we used gene name dictionaries with genes from 32 species (see Table 2), covering 69% of all species mentions in MEDLINE and PubMed Central.

Table 2. List of species-specific dictionaries.

List of species for which we built and used gene name dictionaries. Column two and three give occurrence statistics in the training and test sets (species with no associated genes in both the training or test sets were omitted due to space constraints). The frequencies represent the number of genes associated to each species.

Species	Training frequency	Test frequency
Homo sapiens	121 (19.9%)	181 (10.8%)
Mus musculus	75 (12.3%)	235 (14%)
Rattus norvegicus	14 (2.3%)	41 (2.4%)
Gallus gallus	10 (1.6%)	4 (0.2%)
Saccharomyces cerevisiae S288c	166 (27.3%)	36 (2.1%)
Escherichia coli str. K-12 substr. MG1655	4 (0.6%)	1 (0%)
Arabidopsis thaliana	30 (4.9%)	9 (0.5%)
Drosophila melanogaster	58 (9.5%)	59 (3.5%)
Bos taurus	9 (1.4%)	3 (0.1%)
Caenorhabditis elegans	19 (3.1%)	9 (0.5%)
Xenopus laevis	17 (2.8%)	3 (0.1%)
Danio rerio	42 (6.9%)	7 (0.4%)
Hepatitis C virus	0	1 (0%)
Magnaporthe oryzae 70-15	0	68 (4%)
Neurospora crassa OR74A	0	2 (0.1%)
Schizosaccharomyces pombe	3 (0.4%)	5 (0.2%)
Zea mays	2 (0.3%)	0
Human immunodeficiency virus 1	0	1 (0%)
Sus scrofa	7 (1.1%)	76 (4.5%)
Triticum aestivum	2 (0.3%)	2 (0.1%)
Xenopus (Silurana) tropicalis	1 (0.1%)	0
Macaca mulatta	2 (0.3%)	0
Total	582 (95.1%)	743 (43.5%)

Results and discussion

TAP, F1-score, recall and precision on the training and test corpora

The TAP-5 scores [6] of our system on the training and test data are 0.363 and 0.157, respectively; the corresponding TAP-20 scores are 0.408 and 0.199. Per the construction of the test set, these data were considered more difficult than the training data (see overview paper). On the high-quality part of the training set, we achieved precision, recall, and F1-score of 0.536, 0.474, and 0.503, respectively.

Species recognition results

By applying LINNAEUS to the training and test corpora and comparing the identified species against the manually annotated gene identifiers, we evaluated to what extent LINNAEUS was able to find mentions belonging to the species that are associated with genes in particular papers. For the fully annotated subset of the training corpus (32 documents), the original version of LINNAEUS could find species mentions for 87% (528/607) of the annotated gene entries. When also incorporating the additional dictionaries produced as part of this work (using cell-lines as proxies for species and linking genus names to commonly mentioned species), this rate increased to 94% (571/607). The species identifier "re-linking" was performed in both cases. Performance was lower for the manually annotated subset of the test set (50 documents), where the software was able to locate only 74% (1,242/1,670) and 80% (1,341/1,670) of gene-associated species using the original and extended dictionaries, respectively.

For both the training and test set, a preliminary inspection of a subset of false negatives suggests that the main reason for false negatives is that articles simply do not contain the appropriate species name. While it may be possible to reduce this problem by adding additional "proxy" dictionaries, it is probably not possible to completely solve it.

Analysis of filtering steps

To assess the impact of the individual components used by GNAT, we performed accuracy evaluations of the predicted gene mentions throughout the GNAT pipeline. We evaluated each filtering step, from initial species-dependent gene NER to the final disambiguation and scoring, on the high-quality portions of the BC III training set (see Table 1 and Figure 1). This analysis show that the pipeline methods that contributed the most to the increase in accuracy were the context-based filters (LRCF and ICF), the string similarity search filter (AF), the species disambiguation filter (SVF) and the gene re-classification filter (UMF).

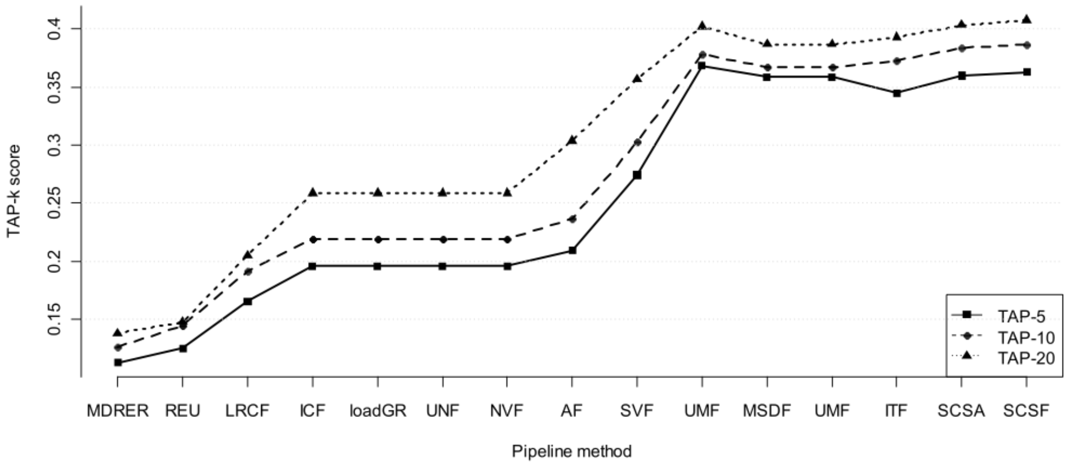


Figure 1. TAP scores after individual filtering steps on the training data. TAP-5, TAP-10, and TAP-20 scores as observed after each individual step of our processing pipeline. Table 1 describes each filtering step.

Differences between the training and test set

Our analysis of the results on the test set suggests that the primary reason for the difference in accuracy seen between the test and training set is the difference in species composition. Our species-specific gene dictionaries covered the species associated to 95% of the annotated gene entries in the training set, but only 43% of the genes in the test set (see Table 2), causing a large number of false negatives. Model organisms were much more common in the training set than in the test set, where species discussed less frequently have a more important role. For instance, 22% of the gene entries in the test data are from *Enterobacter sp. 638*, a species mentioned extremely rarely in MEDLINE. It is clear that while the common model species are heavily over-represented in research [3], the species-specific gene dictionaries used by GNAT represent a limitation for articles that discuss less-frequently mentioned species.

Conclusions and availability

Here, we presented a system for recognizing and normalizing gene mentions in full texts used in the BioCreative III challenge's Gene Normalization (GN) task. We demonstrate the utility for species NER for guiding gene name dictionary recognition, and for gene context profiles used when performing gene normalization. Our training and test set performances differ widely, with TAP-20 scores ranging from 0.4 to 0.2. This difference can primarily be attributed to differences in species composition that could not be handled using the species-restricted approach used by our system, and to some extent the method used for the selection of test data used for evaluation (see overview article). Future work will concentrate on making the initial dictionary NER method less dependent on species-specific dictionaries in order to overcome this problem.

GNAT will be made available at http://gnat.sourceforge.net shortly after the BioCreative III workshop. LINNAEUS and the additional genus and cell-line dictionaries are available at http://linnaeus.sourceforge.net.

Author contributions

IS and MG implemented the adaptations of GNAT and LINNAEUS for BC III. MG and JH wrote the manuscript, with help from the other authors. PT tested GNAT for high-throughput applications. GN, CMB, and UL supervised the work. All authors have read and approved the final manuscript.

Acknowledgements

IS was supported by the Alexander-von-Humboldt Stiftung. MG was supported by the University of Manchester and a BBSRC CASE studentship. PT was supported by the BMBF, grant 0315417B. GN and CMB were supported in part by BBSRC, grant BB/G000093/1. JH was supported by Arizona State University.

References

1. Tamames J, Valencia A: **The success (or not) of HUGO nomenclature.** *Genome Biology* 2006, **7:**402.
2. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzales G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics* 2008, **24:**i126-i132.
3. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature.** *BMC Bioinformatics* 2010, **11:**85.
4. Sarntivijai S, Ade AS, Athey BD, States DJ: **A bioinformatics analysis of the cell line nomenclature.** *Bioinformatics* 2008, **24:**2760-2766.
5. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M: **Gene mention normalization and interaction extraction with context models and sentence motifs.** *Genome Biology* 2008, **9:**S14.
6. Carroll HD, Kann MG, Sheetlin SL, Spouge JL: **Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics.** *Bioinformatics* 2010, **26:**1708-1713.

The Colorado BioCreative III Gene Normalization task submission

Karin Verspoor*, Kevin M. Livingston, Christophe Roeder, Tom Christiansen, Helen L. Johnson, Kevin Bretonnel Cohen, William A. Baumgartner Jr. and Lawrence E. Hunter

Center for Computational Pharmacology, University of Colorado Denver School of Medicine, PO Box 6511, MS 8303, Aurora, CO 80045 USA; Email: karin.verspoor@ucdenver.edu;

1 Introduction

Team 63 from the Colorado Center for Computational Pharmacology prepared a system for tackling the gene normalization task using a novel approach to gene normalization. Our strategy is based on state of the art research in knowledge-based word sense disambiguation from the natural language processing community, and represents the first system to tackle gene normalization through the primary use of *relational* background knowledge. We call this approach **KNoGM**, for Knowledge-based Normalization of Gene Mentions.

Our system includes the basic stages of (1) gene mention tagging, (2) mapping of gene mentions to candidate gene identifiers, an (3) selection of a candidate gene identifier for each mention. We describe each in turn.

2 Gene Mention Tagging

For the gene mention tagging step of the system, we used mentions detected by the **AIIA-GMT** system [1], above a threshold of 0.4.

3 Mapping to Gene Identifiers

3.1 Dictionary matching

Each mention detected by the gene mention tagging system was matched against the names in the full **BioThesaurus** [2]. We utilized version 6.0. For matching, we regularized the gene mention names in the original text and matched against regularized versions of the thesaurus names. The regularization performed involved removing punctuation and whitespace, and transforming greek and roman characters. All Uniprot identifiers matched by the regularized string were retrieved as candidate mappings.

3.2 Abbreviation detection

We perform abbreviation detection in order to reduce the ambiguity in the gene mention candidate sets. We apply the Schwartz and Hearst algorithm [3] to recognize explicit short form-long form pairs. Any occurrence of a detected short form in the document is associated with the candidate set for the long form, rather than the (generally more ambiguous) candidate set for the short form.

4 Gene Identifier disambiguation

The strategy that we pursue for disambiguating a gene/protein mention to the appropriate identifier is to make use of known relationships among proteins. We build on research on word sense disambiguation (WSD) of general English nouns and verbs that demonstrates good performance by taking advantage of the graph structure of a semantic graph connecting word senses [4]. The analogy between the general English case and the gene normalization case is that gene names (words) correspond to multiple gene identifiers (senses).

Graph-based WSD is performed over a graph composed by senses (nodes) and relations between pairs of senses (edges). There may be several types of relations in a single graph and these may have some weight attached to them. The disambiguation is typically performed by applying a ranking algorithm over the graph, and then assigning the concepts with highest rank to the corresponding words.

4.1 Knowledge graph

To support a knowledge-graph-based methodology for gene normalization, we must first construct a knowledge graph. We built a graph combining knowledge from the following sources:

- **iRefWeb:** protein-protein interactions
- **UniProt GOA:** gene ontology annotations for proteins
- **NCBI Taxonomy**: the association between a protein and an organism
- **Homologene**: relations among genes based on homology

In this graph, nodes represent biological entities and concepts – proteins (represented as UniProt identifiers), genes (represented by Entrez Gene identifiers), organisms (represented as NCBI Taxonomy identifiers), and the three types of concepts in the Gene Ontology (cellular components, molecular functions, and biological processes; represented as GO identifiers). Edges among them represent a connection between them. For instance, an edge directly connecting two proteins indicates that there is a known interaction among them based on iRefWeb, and a link between a protein and a gene ontology identifier indicates that the protein has been annotated to that gene ontology term.

4.2 Word sense disambiguation system

To perform the graph-based gene mention disambiguation, we employed the UKB word sense disambiguation system of Agirre and Soroa [4]. Their system is available on-line at http://ixa2.si.ehu.es/ukb/.

The UKB system employs the well-known PageRank algorithm, as well as providing several variations on this core algorithm that introduce use of local document context for disambiguation. The main idea of PageRank is that whenever a link from v_i to v_j exists in a graph, a vote from node i to node j is produced, and hence the rank of node j increases. For the submitted results, we only made use of the static PageRank algorithm but we continue to experiment with this approach.

5 Acknowledgements

We are grateful for the help of Chun-Nan Hsu from the Information Sciences Institute at the University of Southern California and Clark Kuo from Academica Sinica, Taiwan with the application of the AIIA-GMT system to the BioCreative III Gene Normalization data.

References

1. Hsu CN, Chang YM, Kuo CJ, Lin YS, Huang HS, Chung IF: **Integrating high dimensional bi-directional parsing models for gene mention tagging**. *Bioinformatics* 2008, **24**(13):i286–294, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/13/i286].

2. Liu H, Hu ZZ, Zhang J, Wu C: **BioThesaurus: a web-based thesaurus of protein and gene names**. *Bioinformatics* 2006, **22**:103–105, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/1/103].

3. Schwartz A, Hearst M: **A simple algorithm for identifying abbreviation definitions in biomedical text**. In *Pacific Symposium on Biocomputing, Volume 8* 2003:451–462.

4. Agirre E, Soroa A: **Personalizing PageRank for Word Sense Disambiguation**. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)* 2009.

NaCTeM Systems for BioCreative III PPI Tasks

Xinglong Wang*, Rafal Rak, Angelo Restificar, Chikashi Nobata, C.J. Rupp, Riza Theresa B. Batista-Navarro, Raheel Nawaz and Sophia Ananiadou

National Centre for Text Mining and Department of Computer Science, University of Manchester, Manchester, UK

Email: {xinglong.wang,rafal.rak,angelo.restificar,chikashi.nobata}@manchester.ac.uk; {c.j.rupp,batistar,nawazr}@cs.man.ac.uk; sophia.ananiadou@manchester.ac.uk;

*Corresponding author

Abstract

This paper proposes and compares several classification based methods for the BioCreative III ACT and IMT tasks, which respectively address users' requirements for automatically selecting documents relevant to the curation of protein interactions, and for detecting the experimental techniques used in discovering the interactions. For both tasks, we experimented with various classification methods, employing rich contextual and dictionary features. Evaluation on the IMT development data shows that, a new method that classifies pair-wise relations between text phrases and candidate interaction names achieved promising results.

Background

The BioCreative III Interaction Method Task (IMT) concerns automatically detecting experimental techniques used in research articles that support given protein protein interactions (PPIs). For each document, terms describing interaction methods should be recognised and also grounded to unique concept identifiers (MI IDs) as defined in the PSI-MI ontology.[1] The allowed subset of PSI-MI ontology contains 115 interaction methods and each document may be associate with zero or more methods. Therefore, the task can also be cast as a multi-label document classification problem.[2] In addition, full-text documents are provided because descriptions of experimental techniques are not usually found in abstracts.

The Article Classification Task (ACT), on the other hand, requires categorising each document as being relevant or irrelevant to PPI curation. According to the task definition, only documents reporting PPIs are deemed relevant, while those describing interactions between genes or other non-protein biological entities are considered irrelevant. We cast ACT as a binary document classification task.

[1] http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI

[2] Strictly speaking IMT is a multi-class, multi-label classification task but throughout this paper we'll use multi-label for simplicity.

Table 1: Macro-averaged results on IMT development dataset with 10 best models selected by cross-validation on training data (%)

System	Precision	Recall	F1
m-LR	41.36	53.81	46.37
m-SVM	72.12	51.31	59.96
b-SVM	68.35	61.05	**64.49**
union(m-SVM,b-SVM)	65.62	**63.11**	64.33
intersect(m-LR,b-SVM)	75.24	54.96	63.52
intersect(m-LR,m-SVM,b-SVM)	**78.22**	50.17	61.13

Table 2: Macro-averaged 10-fold cross-validation results on ACT training and development datasets (%)

System	Specificity	Sensitivity	Accuracy	MatthewsCoef	AUCiP/R
SVM (C-value=0.25)	70.75	94.19	87.39	68.38	75.96
SVM (C-value=0.50)	69.65	93.99	86.93	67.17	74.99
LR	72.66	93.54	87.48	68.77	86.41

Results and Discussion

This paper describes 3 IMT systems: two follow the multi-label document classification framework and the other uses a binary classifier classifying pairs of synonyms in the PSI-MI ontology and text phrases in target documents. Cross-validation was performed on the training dataset, based on which the best models were selected and tested on the development set. Results are shown in Table 1, where *m-LR* and *m-SVM* are multi-label document classification systems, with the former using Logistic Regression (LR), and the latter support vector machines (SVM), and *b-SVM* is the binary classification system. The features used in b-SVM are different from those in m-SVM and m-LR. For each document, we also took union and intersection of sets of MI IDs obtained from the aforementioned approaches. The three highest performing ensemble systems are also shown in Table 1: as expected, the union of results improved recall and the intersection improved precision.

For ACT, we extracted a rich set of features and then used two machine learning paradigms: LR and SVM. Table 2 shows 10-fold cross validation results as tested on a combined ACT training and development datasets.

Conclusions

We compared several approaches to the BioCreative III IMT and ACT tasks. For IMT, we proposed a new method that first searches candidate interaction method text strings in documents, and then classifies pair-wise relations between the candidates and their matching interaction method names, as defined in PSI-MI. This method utilises a rich set of features extracted from the candidates' surrounding context, together with the definitions and synonyms in PSI-MI. Evaluation results on the development dataset show that, overall, this method is promising and outperformed the more conventional multi-label document classification using the "one-vs-all" strategy. We also tested simple ensemble systems using heuristic rules of union and intersection, which achieved good overall performance, and are especially competitive in recall or precision.

For ACT we tested LR and SVM classifiers exploiting features that are commonly used in PPI classification tasks. It appears that protein entities identified by our named entity recogniser, together with MeSH headings, provided much distinguishing power.

Methods

Document Pre-processing

The IMT documents were provided in various formats and we used PDF-converted plain text. The quality of the text was not satisfactory but we did not find a quick solution to address it. Nevertheless, we normalised typographic ligatures (e.g., ﬃ → ffi), some Unicode punctuation, such as different white space, dashes, single and double quotes, and also removed control characters. By contrast, the ACT documents were of good quality and therefore we did not apply the above text-cleaning steps.

The IMT and ACT documents were then pre-processed using a number of linguistic processors [1], including tokenisation, lemmatisation, part-of-speech tagging and chunking, and processed with a named entity recogniser (NER). The NER is the same as used in our semantic search engine KLEIO[3], which is based on the method described in [2]. It consists of two components: the first finds entity candidates by searching a dictionary; and if a manually annotated corpus is available, the second component trains a conditional random fields model, which is then used to tag unseen text. We applied both components for annotating genes and proteins, and only the first for annotating the following types of entities: metabolite, organ, drug, bacteria, diseases, symptoms, diagnostic/therapeutic procedure and phenomenon. Please refer to [3] for more details regarding the NER.

In addition, for each IMT or ACT document, we retrieved its MeSH headings and information associated with the headings from MeSH ontology. Information of our interest includes descriptor names and identifiers, in both their atomic and hierarchically ordered form (i.e., tree ID), with the latter more closely representing the underlying structure of MeSH ontology. This information was used as a feature in training machine learning models. For IMT, we also manually constructed a mapping from the 10 most frequent MI IDs (as found in the training data), to their corresponding MeSH descriptors.

IMT

Classifying Pairs of Text Chunks and Method Names

This approach first searches every text chunk in a full-text document and collects the chunks that are approximately similar to an interaction method name in the PSI-MI ontology, where the strength of similarity was determined by a string similarity measure. In our work, the text chunks were noun phrases (NP) or verb phrases (VP), and a pair was deemed similar if its text chunk and MI name had a SoftTFIDF [4] similarity score above 0.50. The second-level similarity measure used in SoftTFIDF was JaroWinkler [5], with threshold 0.85. After such pairs were gathered, we classified each pair, where a positive label indicates the text chunk in question entails its pairing interaction method. All interaction method IDs appearing in the *positive* pairs were then assigned to the document.

In more detail, suppose document D contains n NP or VP chunks; we compare each chunk to every name in PSI-MI, and gather all pairs whose similarity scores are above 0.50. For example, if pair p consists of an NP chunk, "anti-His tag antibodies", and a method name "anti tag coimmunoprecipitation" (MI:0007), and their similarity is 0.834, then the model would classify p to determine whether the chunk bears the ID of MI:0007. If p is positive, MI:0007 would be assigned to document D.

This way, a multi-label document classification problem is converted to a binary one, simplifying the machine learning task; and if we carefully choose features that depict the relation between a chunk and an MI ID, by, for example, looking at how much the chunk's surrounding context overlaps the description of the ID in PSI-MI, the actual content of the chunk and the ID become less important. In other words, performance of classifying such pairs is less dependent on the amount of training data available for the MI ID in question. Hence, the approach addressed the problem faced by multi-label document classification where many MI IDs do not have sufficient data to train a good model.

We used SVM^{perf} classifier (with the linear kernel) [6][4] and a rich feature set. Features used are listed below. For the sake of explanation, we define a candidate pair to be classified as $p = \{c, n\}$, where c is a

[3] http://www.nactem.ac.uk/software/kleio/

[4] http://www.cs.cornell.edu/People/tj/svm_light/svm_perf.html

text chunk and n is an MI name that matches c. Let's also define id as n's corresponding MI identifier and the document containing c as D.

Local context includes contextual words within a defined window surrounding c. We chose two window sizes: 10 and 50 with the former additionally accompanied by position information.

Local NER context is the named entities adjacent to c. We took 5 on each side and both the type (e.g., protein) and text of the entities were used.

MI synonym match We searched the local context (window size 20) of c and the global context of D for the synonyms linked to id in PSI-MI. The number of matches in both cases were used as features.

MI definition match In addition to synonyms, key words in the definition associated with id may be useful. We ranked the tokens in each definition according to their TFIDF scores so that the tokens at the top of the rank were more likely to relate to id. Given this rank, we searched the local context (window size 20) of c and also the global context of D for tokens in id's definition, and then used the TFIDF rank linked to each definition token as a binary feature.

Section title Based on the assumption that method names are more likely to be mentioned in some sections (e.g., "Materials and Methods") than others, we searched the commonly used section names in biomedical articles, such as "introduction" and "results and discussion", and tagged them as section titles.

MeSH headings A feature indicating whether D is annotated with a MeSH heading that matches an MI ID, using the mapping previously described.

Other features include the text strings of c and id, and string similarity score between c and n.

Note that all contextual words were lemmatised and "stop words" (e.g., functional words and words consisting of only digits) were removed. We tuned the C-value of SVM with cross-validation on the training set and achieved the best F1-scores with the values 16 or 32. The final model was trained on training and development datasets. This method is referred to as b-SVM in Table 1.

Multi-label Document Classification

We also approached IMT as multi-label document classification, using an ensemble of binary classifiers produced for each class (i.e., MI ID). We trained the binary classifiers using the *one-vs-all* strategy, where each model was trained on all instances, and positive instances were those that belong to a class for which the model was being built. To classify a document D, it would be scored by all the models, and if the score of D when given as input to a particular model was greater than the corresponding threshold, then the label associated with that model would be assigned to D.

The features used were different from those in b-SVM. Two types of features were used: (1) type and text of named entities, words surrounding the entities (window size 10 with position) and the title of the section in which the entities occur; (2) word unigrams and character n-grams ($n = \{2, 3, 4\}$) from the MI definition and synonyms. All features were binary. Based on this set of features, we tested two machine learning algorithms: LR and SVM (referred to as m-LR and m-SVM respectively in Table 1).

Logistic Regression We used models trained via L_2-regularized LR [7, 8] from instance vectors constructed using the features described above. In total, 85 LR models were constructed, one for each interaction method for which at least one instance in the training set was found. In order to assess how the LR models would generalize on unseen data, both the training data and development data provided by the organizers were used in a 10-fold cross-validation experiment. In this experiment, we have set aside the development data as a test set and decided to use the training data to build our models and also determine the threshold value.

We first randomly divided the data into 10-folds, and then performed 10 runs using the training data to build models and separately used the development data for testing. For each run, we used 9 folds to train the LR model LR_j, $j \in [1, 85]$, and the remaining fold to determine the threshold for LR_j. Thus for each run, we trained 85 LR models. During training, a document D corresponds to one training instance. For a specific interaction method MI_j, $j \in [1, 85]$, we associate a corresponding LR model LR_j. D is a positive example for LR_j if it has been assigned a label MI_j in the training data, otherwise D is a negative example

Table 3: Influence of thresholding to SVM as tested by 10 fold cross-validation on IMT training data. All scores are macro-average (%).

System	Precision	Recall	F1
no thresholding	74.35	44.08	55.29
SCut	78.86	54.90	64.61

for LR_j. We performed 10 experiment runs, training a total of 850 LR models and averaged the results of the evaluation on the development dataset, as shown in Table 1. To construct the final model for the official test data, we used the training data to train the 85 LR models and the development set to determine the thresholds for each model, subject to the constraint that each threshold has a minimum value of 0.10.

SVM The implementation of SVM used was SVM^{perf} with the linear kernel. The parameter C-value was tuned using 10-fold cross-validation on the training set in the fashion similar to the LR classification. Since the applied range of parameter values produced the same micro-average F1-scores, we arbitrarily chose $C = 1$ for the final model.

Thresholding Strategies It has been argued that simple thresholding on scores obtained from classifiers (0.5 in the case of LR or zero in the case of SVM) does not always yield the best performance in multi-label classification. Several thresholding strategies have been proposed, which can be categorised into rank-based thresholding, class proportion-based assignment, and score-based local optimisation [9]. Due to the time constraints, we were able to test only the score-based optimisation strategy (SCut), which assigns a class to a document based purely on the score between the two and a class-specific threshold. We compared the performance classifying documents using local (class-fitted) thresholds and global thresholds. The thresholds were tuned using 10-fold cross-validation on the training set whereas the evaluation was performed on the development set. In the case of SVM, the classification with local thresholds resulted in an inferior (albeit marginally) micro-average F1 to the classification with the global nominal (zero) threshold. That could indicate that the local thresholds over-fitted the training set, which was less likely with the global thresholding.

It is important to note that local thresholding driven to optimise individual class's F1-score outperformed the global thresholding in terms of per-class macro-average F1. However, a relatively small improvement of F1-scores for very small classes was obtained with the expense of a large amount of false positives, which significantly lowered the overall micro-average F1-score. Ideally, the subject of optimisation would be the micro-average F1-score, which would require tuning a range of per-class threshold combinations; however, due to the time constraints we opted for simpler and faster per-class accuracy, which, although not ideal, proved to be better balanced than the per-class F1-score. A comparison of systems with and without the SCut thresholding strategy is shown in Table 3.

In contrast to SVM, the choice of the subject of optimisation (F1-score vs. accuracy) when choosing the thresholds for the LR models did not affect the performance. Similarly to SVM, LR with a global threshold outperformed the local thresholding strategy. Interestingly enough, the best micro-average F1 was obtained with threshold 0.1, which is substantially lower than the nominal 0.5. Further analysis revealed that for the majority of classes the range of prediction probabilities occupied the lower part of the $[0, 1]$ range. Lowering the threshold from the nominal 0.5 to 0.1 boosted recall with an acceptable decrease in precision.

ACT

To get a better understanding of the task at hand, we analysed a few randomly chosen positive and negative sample abstracts from the training dataset, in terms of whether the presence of the following attributes in an abstract correlate to its class (i.e., positive or negative): protein names (A1), verbs or nominalised verbs around protein names that signify protein interaction (involving more than one participant) (A2), verbs or nominalised verbs near protein names signifying protein modification (one participant) (A3), protein name

Table 4: Analysis of 10 sample abstracts

Positive Samples						Negative Samples					
PMID	A1	A2	A3	A4	A5	PMID	A1	A2	A3	A4	A5
17517622	yes	yes	no	yes	yes	19413980	no	no	no	no	no
17586502	yes	no	yes	no	yes	19416831	yes	no	no	yes	no
17666011	yes	yes	no	no	yes	19421224	yes	no	no	no	no
17762861	yes	yes	no	yes	yes	19429605	yes	no	yes	no	no
17942705	yes	yes	yes	yes	no	19435285	yes	no	yes	no	no

MeSH headings (A4) and protein-related biochemical process MeSH headings (A5).

Table 4 shows the results of the analysis of the randomly chosen documents. From this analysis, we were able to determine that verbs around protein names (A2) and MeSH terms pertaining to biochemical processes (A5) can be used as indicative features for distinguishing between positive and negative examples.

To classify a document D, we used the following features: bag of words in D, named entities in D, words in the sentences that contain at least one protein (with position) and MeSH headings associated with D. Again, LR and SVM classifiers were adopted. For LR, we trained an L_2-regularized model [7,8] using above features. To assess how LR models perform on ACT data, we ran three sets of experiments: (1) 10-fold cross-validation on training data (2) separately train on the entire training data and test the trained model on the development set (3) 10-fold cross-validation on the combined training and development data. LR achieved accuracies scores of 86.79%, 84.83% and 87.75%, respectively. We did similar experiments using SVM^{perf} and the final models were trained on the combined training and development data.

Acknowledgements

The work is supported by the Biotechnology and Biological Sciences Research Council (BBSRC), UK.

References

1. Alex B, Grover C, Haddow B, Kabadjov M, Klein E, Matthews M, Roebuck S, Tobin R, Wang X: **Assisted curation: does text mining really help?** In *Proceedings of the Pacific Symposium on Biocomputing* 2008.

2. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S: **How to make the most of NE dictionaries in statistical NER**. *BMC Bioinformatics* 2008, **9**(Suppl 11:S5).

3. Nobata C, Sasaki Y, Okazaki N, Rupp C, Tsujii J, Ananiadou S: **Semantic Search on Digital Document Repositories based on Text Mining Results**. In *Proceedings of International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)* 2009:34–48.

4. Cohen WW, Ravikumar P, Fienberg SE: **A Comparison of String Distance Metrics for Name-Matching Tasks.** In *Proceedings of IIWeb-03 Workshop* 2003.

5. Winkler WE: **The state of record linkage and current research problems**. Tech. rep., Statistics of Income Division, Internal Revenue Service Publication R99/04 1999.

6. Joachims T: **A support vector method for multivariate performance measures**. In *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, NY, USA: ACM 2005:377–384.

7. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ: **LIBLINEAR: A library for large linear classification**. *Journal of Machine Learning Research* 2008, **9**:1871–1874.

8. Lin CJ, Weng RC, Keerthi SS: **Trust region Newton method for large-scale logistic regression**. *Journal of Machine Learning Research (JLMR)* 2008, **9**:627–650.

9. Yang Y: **A study of thresholding strategies for text categorization**. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, NY, USA: ACM Press 2001:137–145.

MyMiner system description

David Salgado[1§], Martin Krallinger [2], Elodie Drula[1], Ashish Tendulkar[3], Alfonso Valencia[2], Christophe Marcelle[1]

[1]Australian Regenerative Medicine Institute, Monash University, Australia

[2]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

[3]Department of Computer Science and Engineering, IIT Madras, Chennai-600 036, India

[§]Corresponding author

Email addresses:

DS: david.salgado@monash.edu

MK: mkrallinger@cnio.es

ED: Elodie.Drula@afmb.univ-mrs.fr

AT: ashishvt@cse.iitm.ac.in

AV: avalencia@cnio.es

CM: christophe.marcelle@monash.edu

General Description

A range of biomedical text mining systems are currently available online, but these are generally not flexible enough to be easily adapted to particular information demands posed by the biocuration community, such as (1) assisting the semi-automatic construction of links for a given document of interest to gene/protein identifiers or (2) the ranking a desired article collection according to a curation topic of interest.

The MyMiner system provides an easy to use online interface for constructing a labeled training collection of relevant and non-relevant articles. Labeling efficiently relevant articles is crucial for the development of training data to be used by supervised or semi-supervised classifier systems. MyMiner facilitates recording of the time spend per abstract, the highlighting of positive and negative keywords or gene names, as well the analysis of overlapping data curated by multiple annotators. The labeled data collection derived from MyMiner can be easily used as training set for any existing retrieval system, such as MedlineRanker [1] or data mining packages like SVMLight [2] or Weka [3].

Another module of MyMiner (Entity tagging) allows the automatic tagging of desired bio-entity mentions in a given document of interest. It integrates the facility to use automatic detection of mentions of proteins, genes, cell lines and cell types as well the tagging of species names [4]. Furthermore using a color code system it also highlights these mentions in the text, and allows quick manual correction and editing. The user can add also new bio-entity types, as well as specify relations between co-mentioned bio-entities using a co-mention matrix check box. Such relation types might be useful to the extraction of annotations, e.g. protein-protein interactions or protein – functional term association. Results can be downloaded in an XML-tagged output file.

Finally with the Entity Linking module, the end user (which also can upload a entity tagged file from the previous step) is able to provide database links to proteins mentioned in a document input. The list of automatically recognized protein mentions can be manually edited, and also the end user can add restrictions in terms of the organism source. MyMiner integrates a species tagger system to pre-index organism mentions as constraints for the gene/protein normalization [5]. One advantage of MyMiner is that it allows direct searches against the UniProt database for the protein normalization step, using the efficient search ranking algorithm implemented by the UniProt search query interface. Potential protein identifiers can be easily selected with a simple to use check box. Annotations can be exported as a tagged annotation XML file.

Status

The main facilities of MyMiner have been already implemented. Right now it has some restrictions in terms of the input format (3 tab separated column file), which could be adapted to also handle other formats such as PubMed Central (PMC) input text. Additional aspects which have been already analyzed but not integrated to the current system are the gene normalization confidence scoring (record to text similarity based, as well as qualitative and quantitative protein properties disambiguation approach). URL: http://myminer.armi.monash.edu.au

Acknowledgements

The MyMiner project was supported by, MyoRes European Network of Excellence dedicated to study normal and aberrant muscle development function and repair and the French Association against Myopathies (AFM). The work of MK and AV was funded by the projects BIO2007 (BIO2007-666855), CONSOLIDER (CSD2007-00050) and ENFIN (LSGH-CT-2005-518254).

References

1. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acid Res.* 2009, **37**(S2):W141
2. Joachims T: **Text Categorization with Support Vector Machines: Learning with Many Relevant Features.** *Proceedings of the European Conference on Machine Learning* 1998, 137-142
3. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update.** In: *ACM SIGKDD Explorations Newsletter* 2009, **11**(1):10-18.
4. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.** *Bioinformatics* 2005, **21**(14): 3191-3192.
5. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: A species name identification system for biomedical literature.** *BMC bioinformatics* 2010, **11**(1):85.