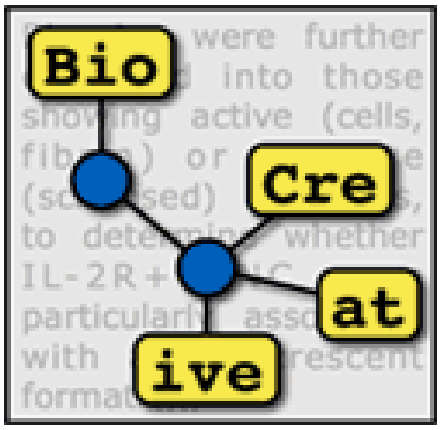
Proceedings of
2012 BioCreative Workshop



April 4 -5, 2012
Washington, DC USA

Editors:

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

UNIVERSITY OF DELAWARE · 1743 ·

2012 BioCreative Workshop Proceedings
Table of Contents

Track 3

Preface

Welcome to the BioCreative 2012 workshop being held in Washington DC, USA on April 4-5, 2012. 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 continued 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 introduced a new interactive task (IAT), ran as a demonstration task. The goal of IAT was 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 included ranking genes by importance based preferably on the amount of described experimental information regarding genes.

The BioCreative-2012 Workshop on Interactive Text Mining in the Biocuration Workflow aims to bring together the biocuration and text mining communities towards the development and evaluation of interactive text mining tools and systems to improve utility and usability in the biocuration workflow. To achieve this goal, the workshop consists of three Tracks: I-Triage a collaborative biocuration-text mining development task for document prioritization for curation; II-Workflow a biocuration workflow survey and analysis task; and III-Interactive TM an interactive text mining and user evaluation task. The workshop includes a demo/testing session where curators will be able to test system presented in Track I and III.

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

The BioCreative 2012 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
Ben Carterette, 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
Carolyn Mattingly, Mount Desert Island Biological Laboratory, MDIBL, USA
Alfonso Valencia, Spanish National Cancer Centre, CNIO, Spain
Thomas Wiegers, Mount Desert Island Biological Laboratory, MDIBL, USA
John Wilbur, National Center for Biotechnology Information, NCBI, NIH, USA
Cathy Wu, University of Delaware and

Local Organizing Committee
Cecilia Arighi, University of Delaware, USA
Sun Kim, National Center for Biotechnology Information (NCBI), NIH, USA
Peter McGarvey, Georgetown University, USA
Zhiyong Lu, National Center for Biotechnology Information (NCBI), NIH, USA
Susan Phipps, University of Delaware, USA
Baris Suzek, Georgetown University, USA
John Wilbur, National Center for Biotechnology Information (NCBI), NIH, USA
Cathy Wu, University of Delaware and Georgetown University, USA
Mehershrutisrin Yerramalla, Georgetown University, USA

Proceedings Committee
Cecilia Arighi, University of Delaware, USA
Katie Lakofsky, University of Delaware, USA

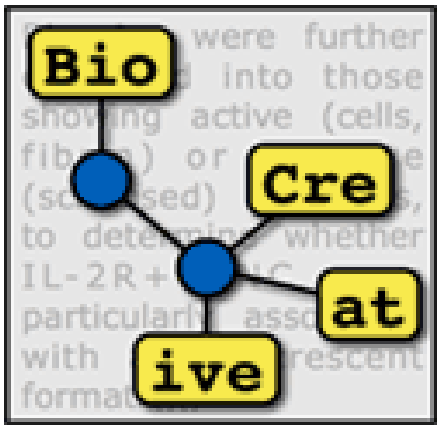
2012 BioCreative Workshop Agenda

April 4-5, 2012
Georgetown University Hotel and Conference Center
Washington, DC USA

Wednesday, April 4, 2012	
8:30 AM – NOON	**Registration: West Lobby**
7:30 AM – 9:00 AM	Breakfast: South Gallery
10:30 – 12:30 PM	**BioCuration 2012 Joint Session: Conference Room 4** • Session 6: Integrating text mining into biocuration workflows
12:30 PM – 1:30 PM	**Lunch** (Salons ABG)
1:30 PM – 1:40 PM	**Workshop Opening:** Lynette Hirschman, Salon DE
1:40 PM – 2:15 PM	**Overview on Track I (Triage) results**: Thomas Wiegers, MDI Biological Laboratory Salon DE
2:15 PM – 3:40 PM	**Participant Track I: Selected Team Participants,** Salon DE • 2:15 – 2:30 pm: Team 121 – System Description for BioCreative 2012 Triage Task • 2:30 – 2:45 pm: Team 116 – Ranking of CTD Articles and Interactions Using the OntoGene Pipeline • 2:45 – 3:00 pm: Team 120 – Selection of Relevant Articles for Curation for the Comparative Toxicogenomic Database • 3:00 – 3:15 pm: Team 130 – CoIN: a Network Exploration for Document Triage • 3:15 – 3:40 pm: Discussion (Moderated by Thomas Wiegers)
3:40 PM – 4:00 PM	Break: South Gallery
4:00 PM – 5:00 PM	**BioCuration 2012 Joint Session: Conference Room 4** • Plenary session 3: Rich Roberts
5:00 PM – 5:30 PM	Break
5:30 PM – 7:30 PM	**BioCreative Workshop Reception and Poster Session: Salon CH**

Thursday, April 5, 2012	
7:30 AM – 12:00 PM	**Registration**
7:30 AM – 8:30 AM	Breakfast South Gallery
8:00 AM – 8:20 AM	**Overview on Track II (Workflow): Zhiyong Lu,** Salon DE
8:20 AM – 10:00 AM	**Participant Track II: Selected team participants,** Salon DE • 8:20 – 8:35 am: Team 142 – WormBase Literature Curation Workflow • 8:35 – 8:50 am: Team 50 – Literature curation workflow at The Arabidopsis Information Resource (TAIR) • 8:50 – 9:05 am: Team 151 – Summary of Curation Process for one component of the Mouse Genome Informatics Database Resource • 9:05 – 9:20 am: Team 156 – The Xenbase Literature Curation Process • 9:20 – 9:35 am: Team 159 – Summary of the FlyBase-Cambridge Literature Curation Workflow • 9:35 – 9:50 am: Team 162 – Incorporating text-mining into the biocuration workflow at the AgBase database • 9:50 – 10:00 am: Discussion (Moderated by Lynette Hirschman)
10:00 AM – 10:30 AM	Break
10:30 AM – 10:50 AM	**Overview on Track III (Interactive TM): Cecilia Arighi,** Salon DE
10:50 AM – 12:30 PM	**Participant Track III: Selected team participants,** Salon DE • 10:50 – 11:00 am: Team 132 – T-HOOD: Text-mined Hypertension, Obesity, Diabetes Candidate Gene Database • 11:00 – 11:15 am: Team 142 – Textpresso Text mining: Semi-automated Curation of Protein Subcellular Localization Using the Gene Ontology's Cellular Component Ontology • 11:15 – 11:30 am: Team 143 – PCS for Phylogenetic Systematic Literature Curation • 11:30 – 11:45 am: Team 153 – PubTator: A PubMed-like interactive curation system for document triage and literature curation • 11:45 – 12:00 pm: Team 158: PPInterFinder – A Web Server for Mining Human Protein–Protein Interactions • 12:00 – 12:15 pm: Team 160 – Mining Protein Interactions of Phosphorylated Proteins from the Literature using eFIP • 12:15 – 12:30 pm: Discussion (Moderated by Ben Carterette, Martin Krallinger, Kevin Cohen and John Wilbur)
12:30 PM – 1:30 PM	Lunch: Faculty Club Restaurants
1:30 PM – 4:00 PM	**Participant Tracks I & III: Demos and system testing,** Salon BG
4:00 PM – 4:30 PM	**Retrospective & future: BioCreative IV – Organizers,** Salon DE
4:30 PM	**Workshop Closing**

Track 1



Collaborative Biocuration-Text Mining Development Task for Document Prioritization for Curation

Thomas C. Wiegers*, Allan Peter Davis, and Carolyn J. Mattingly

Department of Biology, North Carolina State University, Raleigh, NC, USA

*Corresponding author: Tel: 207-288-9880, E-mail: twiegers@ncsu.edu

Abstract

The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation is a community-wide effort for evaluating text mining and information extraction systems for the biological domain. The *BioCreative Workshop 2012* subcommittee identified three areas, or tracks, that comprised independent, but complementary aspects of data curation in which they sought community input: literature triage (Track I); curation workflow (Track II); and text mining/Natural Language Processing (NLP) systems (Track III). Track I participants were invited to develop tools or systems that would effectively triage and prioritize articles for curation and present results in a prototype web interface. Training and test data sets were derived from the Comparative Toxicogenomics Database (CTD; http://ctdbase.org) and consisted of manuscripts from which chemical-gene-disease data were manually curated. In this paper we present a detailed description of the challenge and a summary of the results.

Introduction

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org) is a publicly available resource that aims to promote understanding about the mechanisms by which drugs and environmental chemicals influence the function of biological processes and human health (1). CTD data are manually curated by a team of PhD-level biocurators. Articles are typically prioritized by chemicals of interest and distributed to biocurators who then capture relevant data using our first-generation web-based curation application (2). Curated data include chemical-gene/protein interactions, chemical-disease relationships, and gene-disease relationships. These data are integrated with select external data sets to facilitate development of novel hypotheses about chemical-gene-disease networks (3).

All manually curated data are captured using freely available controlled vocabularies. Chemicals are represented using terms from the Chemicals and Drugs subset of the National Library of Medicine's Medical Subject Headings (MeSH) vocabulary (4); genes and proteins are represented using the Entrez Gene vocabulary (5); diseases are represented using CTD's novel disease vocabulary MEDIC (6) that merges OMIM and the Disease subset of the MeSH vocabulary (4,7); and chemical-gene/protein interactions

are captured using CTD's action vocabulary (1). The implementation of a web-based curation application has had many positive effects on the CTD curation process including, increasing the efficiency of curation, enhancing the flexibility of biocurator location, introducing real-time quality control, and easing data management and storage (2). Research has demonstrated that further enhancement of the curation process for CTD, as well as for many manually curated biomedical resources, would be achieved by improving a) the triage and prioritization of data-rich relevant articles and b) the identification of curatable content within these articles (8). The *BioCreative Workshop 2012* subcommittee dedicated a focus area, or track (Track I), to development of systems that would address these important yet unmet needs of the biocuration community.

The CTD project was chosen as a source for the project data because it possesses a large and high quality set of manually curated information that contains elements that are of broad interest and relevance to the biomedical research community, specifically chemicals, genes/proteins, and diseases. Track I invited text-mining teams to develop a system to assist biocurators in the selection and prioritization of relevant articles for curation for CTD (http://www.biocreative.org/events/bc-workshop-2012/CFP/#track1). Participants were asked to provide two major deliverables that included a) benchmarking results on the prioritization of relevant articles and b) a prototype web interface that would present a biocurator with these articles and the relevant information highlighted using integrated text-mining recognition tools.

Methods

Training Phase

In order for participants to effectively rank articles and identify relevant data, it was critical for them to gain an understanding of the CTD curation process. To facilitate this understanding, a detailed document entitled, *Summary Of Curation Details For The Comparative Toxicogenomics Database,* was distributed to participants (http://www.biocreative.org/tasks/bc-workshop-2012/Triage/). In addition, a training data set was made available to participants that consisted of 1,725 articles that had been previously triaged and curated by CTD biocurators. The data were presented in a series of input files that included all associated curated data for eight target chemicals (raloxifene, aniline, amasacrine, doxorubicin, aspartame, quercetin, 2-acetylaminofluorene, and indomethacin).

In January 2012, the *BioCreative Track I File Upload Facility* web site was released (Figure 1). This web site enabled participants to upload their benchmarking files. The web site in turn produced a report containing detailed information regarding their benchmarking performance and aggregate statistics. Specifically, a report was generated that calculated the aggregate *Mean Average Precision* (MAP) score, as well as the recognition scores for each data type curated (chemicals, genes, diseases, and action terms). Additional details were provided that enabled participants to understand how these scores where calculated (Figure 1).

Recognition scores were provided for each data category (chemicals, genes, diseases, and

action term) within each article. Three fields were provided for each data category, *on a per article basis*, and included:

- *Curated Terms* - This field listed the terms, if any, that a CTD biocurator previously curated for each data category.
- *Text Mined Terms* - This field listed the text-mined terms, if any, that a participant provided for each data category.
- *Match Explanation* - This field provided an explanation of how matches between the curated and text-mined terms were determined. Because providing synonyms to curated terms are counted as matches, the notation of *CYP1-->CYP1A1*, for example, indicated that the term *CYP1* was text mined, which is a valid synonym for the actual underlying curated term *CYP1A1*; alternatively, *FZR1-->FZR1* indicated that the text mined term of *FZR1* exactly matched the curated term.

In all, the following information was provided for each article submitted in the form of a post-submission report:

- PubMed ID
- Curated (Y or N)?
- Intermediate *MAP* Score
- Curated Gene Hit Rate
- Curated Chemical Hit Rate
- Curated Disease Hit Rate
- Curated Action Hit Rate
- Text Mined Genes
- Curated Genes
- Gene Match Explanation
- Text Mined Chemicals
- Curated Chemicals
- Chemical Match Explanation
- Text Mined Diseases
- Curated Diseases
- Disease Match Explanation
- Text Mined Action Terms
- Curated Action Terms
- Action Term Match Explanation
- Curated Ixns

The final line of the report provided the aggregate *MAP* and recognition scores in each category. The reports were provided in both HTML and text formats; summary versions were also provided at the participant's discretion that contained solely the aggregate statistics.

Test Phase

On February 6, 2012, a Track I Test Dataset was released to participants. The purpose of this data set was to evaluate the performance of the participants' text-mining pipeline without their prior knowledge of the curated results. The Track I Test Dataset comprised 444 articles that were previously manually curated by CTD biocurators and contained information about three additional target chemicals (urethane, phenacetin, cyclophosphamide). Unlike the comprehensive curated data provided in the Training data set, the Test Dataset contained only the basic identification information for each article (PubMed ID, Title, Abstract, Journal Name, Date). Each participant was asked to process the Test Dataset using their text-mining pipeline, and provide the following information for each article/target chemical combination:

- PubMed ID
- Title
- Abstract
- Journal
- Cited Gene Actors
- Cited Chemical Actors
- Cited Disease Actors
- Marked-up HTML of abstract with tagged links back to CTD for all corresponding terms (see note below)
- Document Relevancy Score
- optional: Marked-up HTML of relevant sentences/phrases extracted with tagged links back to CTD for all actors and terms
- optional: Cited Action Terms
- optional: Cited Interactions

The benchmarking results and associated documentation were due on February 20, 2012. Upon receipt of the benchmarking data from the participants, CTD staff evaluated the results by calculating the following metrics for each participant:

- *MAP* score
- Curated Gene Term Recognition Rate
- Curated Chemical Term Recognition Rate
- Curated Disease Term Recognition Rate

Results

A total of seven groups participated:

- BiTeM Group; Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva; Information Science Department, University of Applied Science; Geneva, Switzerland
- Department of Computer Science and Information Engineering, National Cheng Kung University; Department of Information Engineering, Kun Shan University, Tainan, Taiwan

- Institute of Computational Linguistics, University of Zurich
- Two groups from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
- Dept. Of Computer Science East China Normal University
- National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

To maintain anonymity, each group was randomly assigned a coded identification number.

Mean Average Precision. For MAP score calculations, an article was counted as relevant if it had one or more associated curated interactions. Across the groups, MAP scores were fairly high and consistent, ranging from 70% to 80% (Figure 2).

Curated Term Recognition. The results for recognition scores were significantly more mixed than the MAP scores. The recognition rate for each gene, chemical, and disease term was calculated by comparing the list of text-mined terms to the list of curated terms for each article and in each respective data category. As indicated above, if the curated term, or a synonym for the curated term (as defined by the corresponding CTD controlled vocabulary), was found in the text mined list, it was counted as a match.

Gene recognition ranged from 1% to 49% (Figure 3). Chemical recognition ranged from 4% to 82% (Figure 4). Disease recognition ranged from 1% to 64% (Figure 5). The results of MAP scores, and chemical, gene, disease, and action term recognition scores were also aggregated onto a single bar graph for each participating group (Figure 6).

Aggregate Benchmarking Results Summary. Two of the groups clearly distinguished themselves with respect to aggregate benchmarking results. Group 121 held the highest *MAP* score (80%) while also delivering strong recognition scores in the three major recognition categories (chemicals, genes and diseases). Group 116 delivered the highest recognition scores in two of the three major data categories (*i.e.*, gene and disease recognition). Three other groups (120, 139 and 130) had respectable recognition scores in most, if not all, of the major data categories.

The groups were also asked to provide a system description, all of which were reasonably clear and well-written.

Prototype Web Interface. The participants were asked to deliver a prototype web interface to Track I organizers by March 1, 2012. All seven groups that participated in the benchmarking portion of the challenge also submitted a prototype web interface. Each interface was then evaluated based on functionality and ease-of-use by CTD's biocuration project manager.

Of the seven entries, six provided very sophisticated functionality. (Please note that the web interfaces described below that tag gene, chemical, and disease terms were only as effective at doing so as their benchmarking results suggest; the same is true for those web

interfaces that provided a ranked list of PubMeds: their ranking effectiveness is reflected in their benchmarking *MAP* scores).

Group 121. The web interface developed by Group 121 was outstanding. The biocurator accesses the system by clicking the *Login* link. Once login is complete, the user is presented with a list of chemicals for curation. Clicking on one of the chemicals takes the biocurator to a ranked list of articles associated with the chemical with the following information:

- Title,
- Author(s),
- Journal name, date, and page numbers,
- PubMed ID,
- Related citations hyperlink,
- Abstract hyperlink

The biocurator may remove an article from the list by simply clicking on a single *delete* hyperlink (e.g., Delete from [BC2012-test-urethane]). Clicking on the *Abstract* hyperlink causes an expansion of the screen to include the complete abstract text. All genes, chemicals, and disease actors contained in the title or abstract and identified by the text-mining tool are color-coded and hyperlinked back to the CTD web interface.

Clicking on the title causes a detail page to be displayed. The detail page contains most of the same information as the main page, but also includes a list of text-mined chemical, gene and disease actors, each of which is hyperlinked back to CTD. The interface enables the user to save new annotations, as well as confirm and/or reset existing entries. Although this particular feature as currently implemented does not appear to be of direct application to CTD, it certainly has interesting long-term implications.

The interface includes several additional and very convenient options. A list of text-mined target chemicals is displayed on the main target chemical screen, enabling the biocurator to easily jump from one chemical list to another for curation. The ranked list of articles can be re-sorted based on date or relevancy score. Clicking a chemical, gene or disease checkbox on the main target chemical screen or on the detail page causes these actors to either be highlighted and hyperlinked, or made simply plain text. Because there is sometimes overlap in chemical, gene and disease names, there is a feature that enables the user to correct and save a text-mined actor designation to another category. Finally, there is a *Display Management* screen that enables biocurators to select their highlighting preferences in the interface. For example, a user can specify whether or not to display chemicals, genes, and diseases by default, as well as to set the colors of the display.

Group 116. The web interface developed by Group 116 was also outstanding. The biocurator is presented with a list of target chemicals to curate. After clicking on a target chemical hyperlink, the user is presented with a ranked list of articles by their PubMed ID, relevancy score, and title. Clicking a PubMed ID presents detailed information in a new tab. The new tab displays a split screen; on the left hand side is the *Document* panel

that displays the title and abstract text, along with all of the MeSH terms associated with the paper; the right side of the screen is the *Annotation* panel.

The *Annotation* panel initially consists of two tabs: *Concepts* and *Interactions*; a *Terms* tab may also be displayed if the user selects it from the toolbar. The *Concepts* tab lists the chemical, disease, and gene terms identified during the text mining process, including the accession, term name, frequency of appearance in the abstract, and type of term (*i.e.*, chemical, disease, or gene); the *Concepts* tab contains an entry for each concept identified in at least one term in the document. Each of the concepts is also scored using an algorithm developed by the team. If concept rows are expanded by clicking the plus button, a hyperlink to the relevant Web page of the CTD site appears. The *Interactions* tab displays potential interactions contained within the abstract; these interactions are also derived using a scoring algorithm developed by the team. For each potential interaction, a confidence score is displayed, along with the type and name of each chemical, disease, and gene actor. In the I*nteractions* tab, clicking on the name of a participating concept opens the relevant CTD web page. The *Terms* tab contains an entry for each stretch of text considered as a technical term. However, no concept disambiguation is made, *i.e.,* a term can contain references to more than one concept, even of different types (*e.g.,* genes, chemicals, *etc.*). One of the excellent features of the *Annotation* panel is that check boxes are displayed next to each interaction and concept; clicking these check boxes will cause the associated text mined-data to be highlighted and hyperlinked within the abstract text, or alternatively simply plain text. All of this is cleverly done without a screen refresh, so it is extremely fast.

The interface included several additional and very convenient options. The user may remove concepts, interactions or terms from the *Annotation* tab by simply selecting an associated checkbox and clicking the *Remove Selected* button. One may also highlight a term and add it to the concepts list by simply double clicking on it and completing the necessary data, including term, term type, concept values, comments, and search databases (*i.e.,* CTD or Entrez), in the *Inspectors* tab. Mousing over a term/concept causes the term type and associated accession IDs to be displayed. The user may dump all the information associated with an article into XML format by simply selecting the option from the menu. The curation actions taken upon a document are logged into the document itself and/or into a separate database.

Group 133 The user is initially presented with a selection of chemicals to curate. The biocurator selects a chemical, clicks the *Submit* button, and is presented with a split screen. On the left hand side of the screen is an ordered list of ranked articles with associated information including relevancy score, PubMed ID, article title, journal name and abbreviated abstract.

The biocurator may begin curating from the list. Clicking one of the ranked articles causes a *Detail Info* frame containing detailed information to be displayed on the right hand side of the split screen. More specifically, each of the data elements described above is provided, along with the complete abstract text. The title and abstract contain highlighted genes and chemicals within the text, as well as lists of each beneath the

abstract; the lists hyperlink each text-mined actor back to the CTD web interface. A link is also provided to view the PubMed at NCBI on a separate tab. The interface is very clean and easy to use.

Group 120 In order to begin curation, the biocurator enters a chemical as well as a list of associated PubMed IDs separated by tab or new line characters. The list of PubMed IDs is text mined and processed on a real-time basis. Once the text mining is complete, the biocurator is presented with a list of ranked and relevancy score-sorted PubMeds, including the following information:

- PubMed ID
- Article Title
- Journal Name
- Text Mined Genes
- Text Mined Chemicals
- Texted Mined Diseases
- Relevancy Score

To the left of each PubMed ID, a +/- button either expands or contracts display of the PubMed's abstract. The abstract contains highlighted genes, chemicals, and diseases within the text, each of which is hyperlinked back to the CTD web interface. The interface is very clean and easy to use.

Group 139 The web interface provided was very similar to Group 133's prototype; only subtle differences were apparent.

Group 130 Clicking on the *System Demo* link presents the user with the main curation screen. Portions of the main screen are apparently under construction and are not currently functional. However, selecting a chemical from the *Data set* field and clicking the *Submit* button presents the biocurator with a list of ranked PubMeds that are associated with the selected target chemical. For each PubMed, its numeric sequential rank is provided, along with the article's title, abstract, and a list of text-mined chemical, gene and disease actors. Each of the text-mined actors is highlighted within the title and the abstract text. Each of the actors provided in the respective lists beneath the abstract are hyperlinked back to CTD, although the hyperlinks may or may not actually link to a CTD actor, i.e., the actors do not appear to have been mapped to actual CTD terms.

Group 141 The biocurator is presented with 2 options for curation:

- *Single Mode* – Allows a user to enter a single PubMed ID for text mining.
- *Batch Mode* – Allows a user to load a file containing one or more PubMed IDs; the file must contain one PubMed ID on each line without a blank line including the last line.

In *Single* Mode, entering a single PubMed ID and pressing *Submit* resulted in a report being displayed, providing the PubMed ID entered and a relevancy score. There were

columns available for text mined gene, chemical, disease, and action terms, but these fields were blank. The report provided no further functionality.

In *Batch* Mode, uploading an input file will cause a new page to be opened, providing information associated with the upload along with a link to a results file. Clicking on the link causes TAB-delimited records to be displayed, one for each PubMed ID in the input file. Each TAB-delimited record contains basic information about the PubMed, as well as a relevancy score.

In conclusion, six of the seven submissions for the web interface component of the Track I challenge effectively presented the ranked and highlighted data. Of the six submissions, however, the products provided by groups 121 a nd 116 pr ovided exceptional functionality and were deemed very user-friendly with potential for future expansion and application.

Conclusions

The Track I project was a very involved assignment. Development of effective ranking and recognition tools, as well as a prototype web interface that conveyed these results in a user-friendly manner required a high degree of systems development and integration.

Of the seven groups, five performed very well in virtually every category. Overall, the groups far surpassed expectations and are to be congratulated on their efforts and accomplishments in a short period of time. In addition to the successful generation of systems that may have long-term application for either CTD or other curated database groups, the success of the Track I program underscores the enhanced benefits that result from collaborative efforts among otherwise disparate biological and computational groups.

References

1. Davis, A.P., King, B.L., Mockus, S., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., and Mattingly, C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res*. **39**, D1067-72.
2. Davis, A.P., Wiegers, T.C., Murphy, C.G., and Mattingly, C.J. (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database (Oxford)*. **2011,** bar034.
3 Davis, A.P., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wiegers, T., and Mattingly, C.J. (2009) The Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res*. **37**, D786-92.
4. Coletti, M.H. and Bleich, H.L. (2001) Medical Subject Headings used to search the biomedical literature. *J Am Med Inform Assoc.*, **8,** 317-323.
5. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2011) Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Res*. **39,** D52-57.
6. Davis, A.P., Wiegers, T.C., Rosenstein, M.C., and Mattingly, C.J. (2012) MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database (Oxford)*. **2012**, bar057.
7. Amberger, J., Bocchini, C., and Hamosh, A. (2011) A new face and new challenges for online mendelian inheritance in man (omim(r)). *Hum Mutat*. **32,** 564-567.
8. Wiegers, T.C., Davis, A.P., Cohen, K.B., Hirschman, L., and Mattingly, C.J. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*. **10,** 326.

Figure 1. The BioCreative Track I File Upload Facility. A web interface was developed to allow participants to upload their results (back panel). Following successful uploads, a report was generated and returned to each participant that contained summary or detailed information for each data set.

Figure 2. *Mean Average Precision (MAP)* score results for each participating group. For MAP score calculations, an article was counted as relevant if it had one or more associated curated interactions. Across the groups, MAP scores were fairly high and consistent, ranging from 70% to 80%.

Figure 3. Gene recognition results for each participating group. The ability for text-mining tools to recognize curated genes was measured; terms and synonyms to terms were counted as matches. Gene recognition ranged from 1% to 49%

Figure 4. Chemical recognition results for each participating group. The ability for text-mining tools to recognize curated chemicals was measured; terms and synonyms to terms were counted as matches. Chemical recognition ranged from 4% to 82%

Figure 5. Disease recognition results for each participating group. The ability for text-mining tools to recognize curated diseases was measured; terms and synonyms to terms were counted as matches. Disease recognition ranged from 1% to 64%

Figure 6. Aggregate metrics for each participating group. The results of *Mean Average Precision (MAP)* scores, and chemical, gene, disease, and action term recognition scores are aggregated onto a single bar graph for each participating group. Two of the groups clearly distinguished themselves with respect to aggregate benchmarking results. Group 121 he ld the highest *MAP* score (80%) while also delivering strong recognition scores in the three major recognition categories (chemicals, genes and diseases). Group 116 de livered the highest recognition scores in two of the three major data categories (i.e., gene and disease recognition). Three other groups (120, 139 and 130) had respectable recognition scores in most, if not all, of the major data categories.

Figure 1.

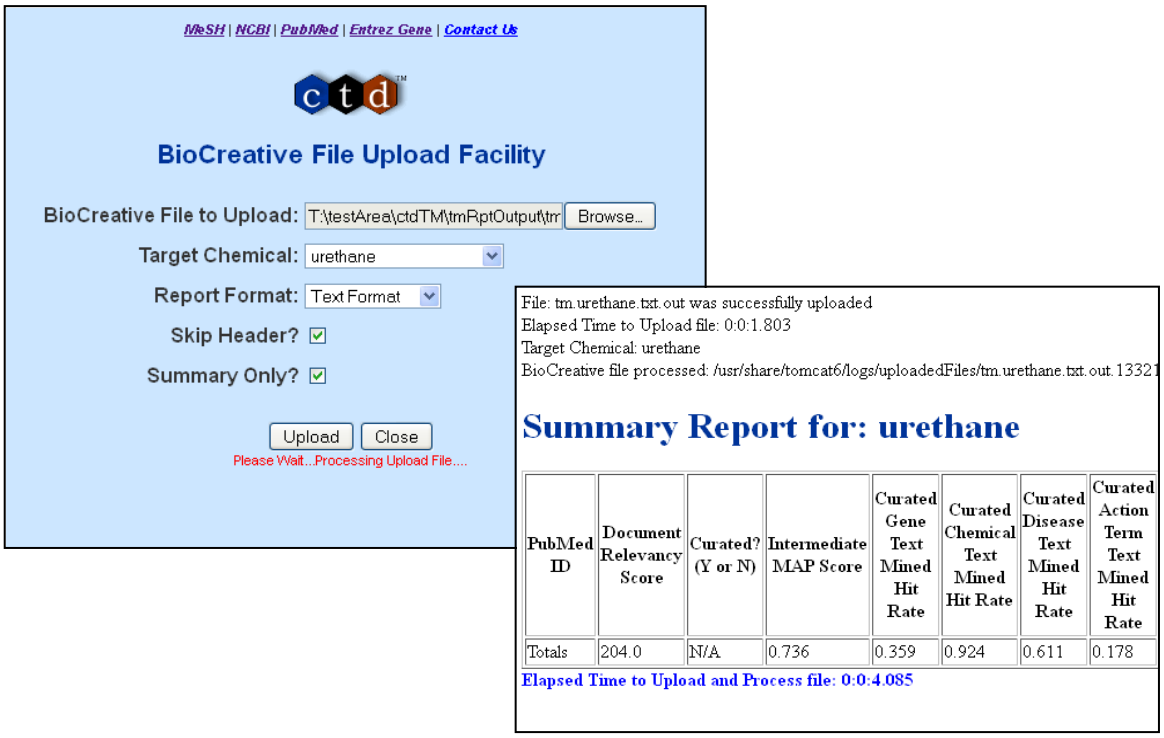

MeSH | NCBI | PubMed | Entrez Gene | Contact Us
ctd
BioCreative File Upload Facility
BioCreative File to Upload:
T:\testArea\ctdTM\tmRptOutput\tm
Browse...
Target Chemical:
urethane
Report Format:
Text Format
Skip Header?
Summary Only?
Upload
Close
Please Wait...Processing Upload File....
File: tm.urethane.txt.out was successfully uploaded
Elapsed Time to Upload file: 0:0:1.803
Target Chemical: urethane
BioCreative file processed: /usr/share/tomcat6/logs/uploadedFiles/tm.urethane.txt.out.13321
Summary Report for: urethane
PubMed ID
Document Relevancy Score
Curated? (Y or N)
Intermediate MAP Score
Curated Gene Text Mined Hit Rate
Curated Chemical Text Mined Hit Rate
Curated Disease Text Mined Hit Rate
Curated Action Term Text Mined Hit Rate
Totals
204.0
N/A
0.736
0.359
0.924
0.611
0.178
Elapsed Time to Upload and Process file: 0:0:4.085

Figure 2.

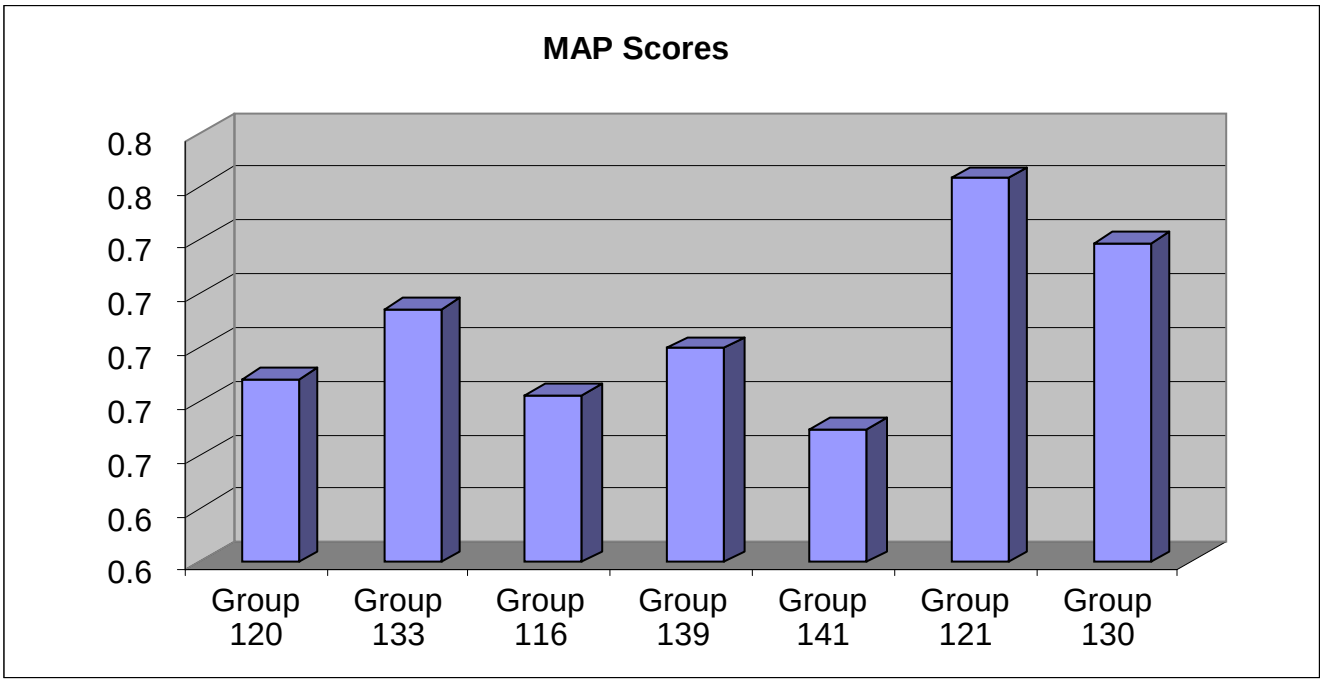

MAP Scores
0.8
0.8
0.7
0.7
0.7
0.7
0.6
0.6
Group 120
Group 133
Group 116
Group 139
Group 141
Group 121
Group 130

Figure 3.

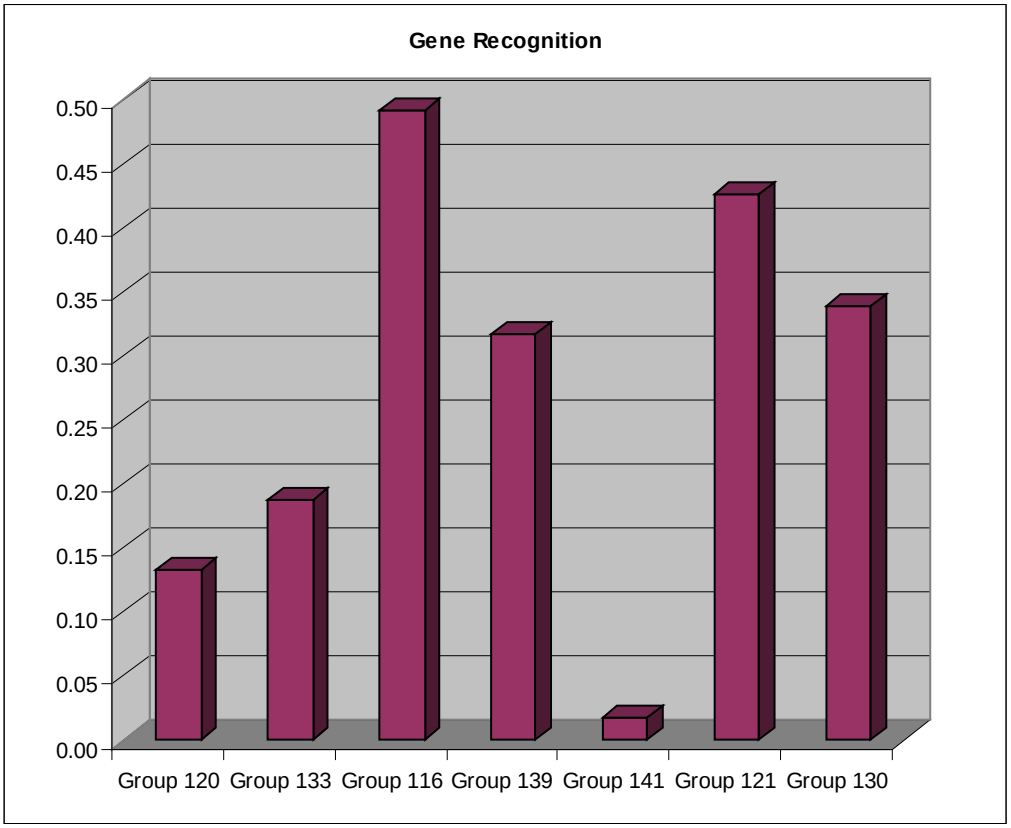

Gene Recognition
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Group 120
Group 133
Group 116
Group 139
Group 141
Group 121
Group 130

Figure 4.

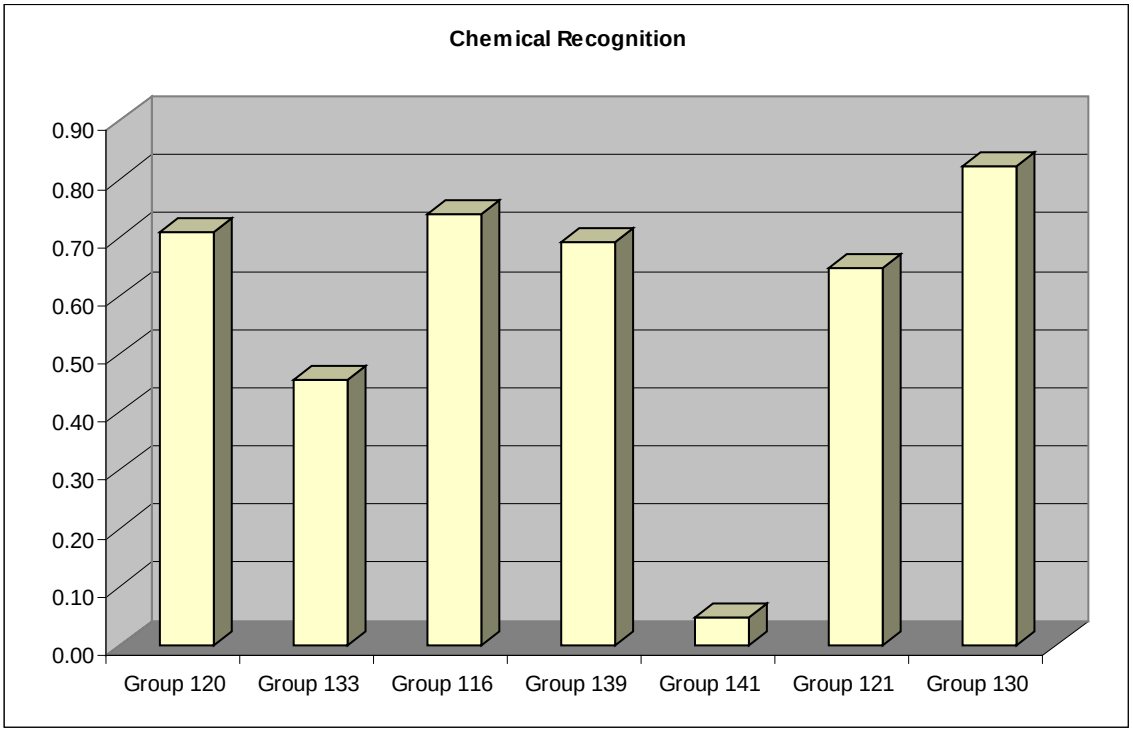

Chemical Recognition
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Group 120
Group 133
Group 116
Group 139
Group 141
Group 121
Group 130

Figure 5.

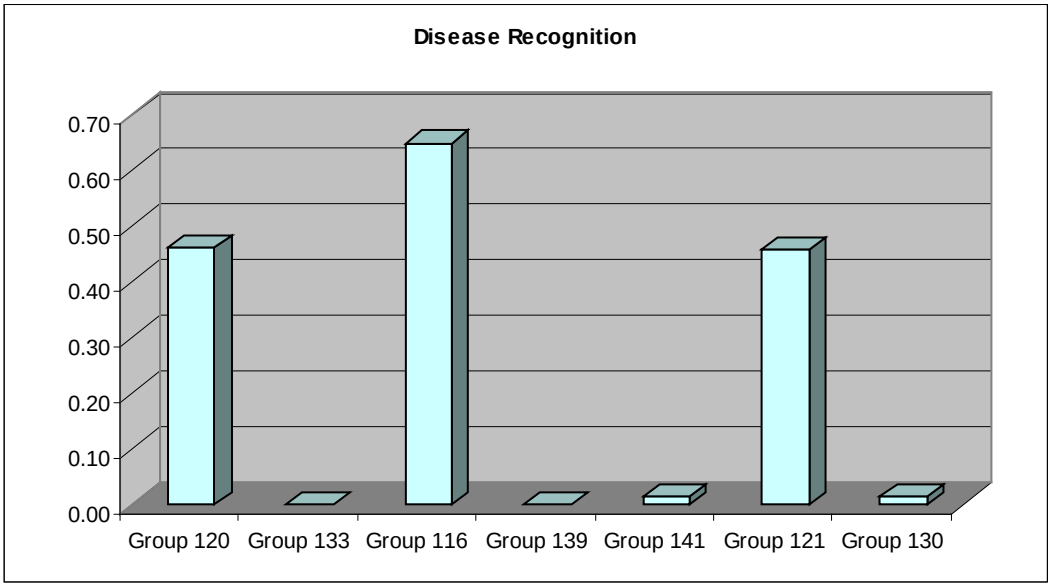
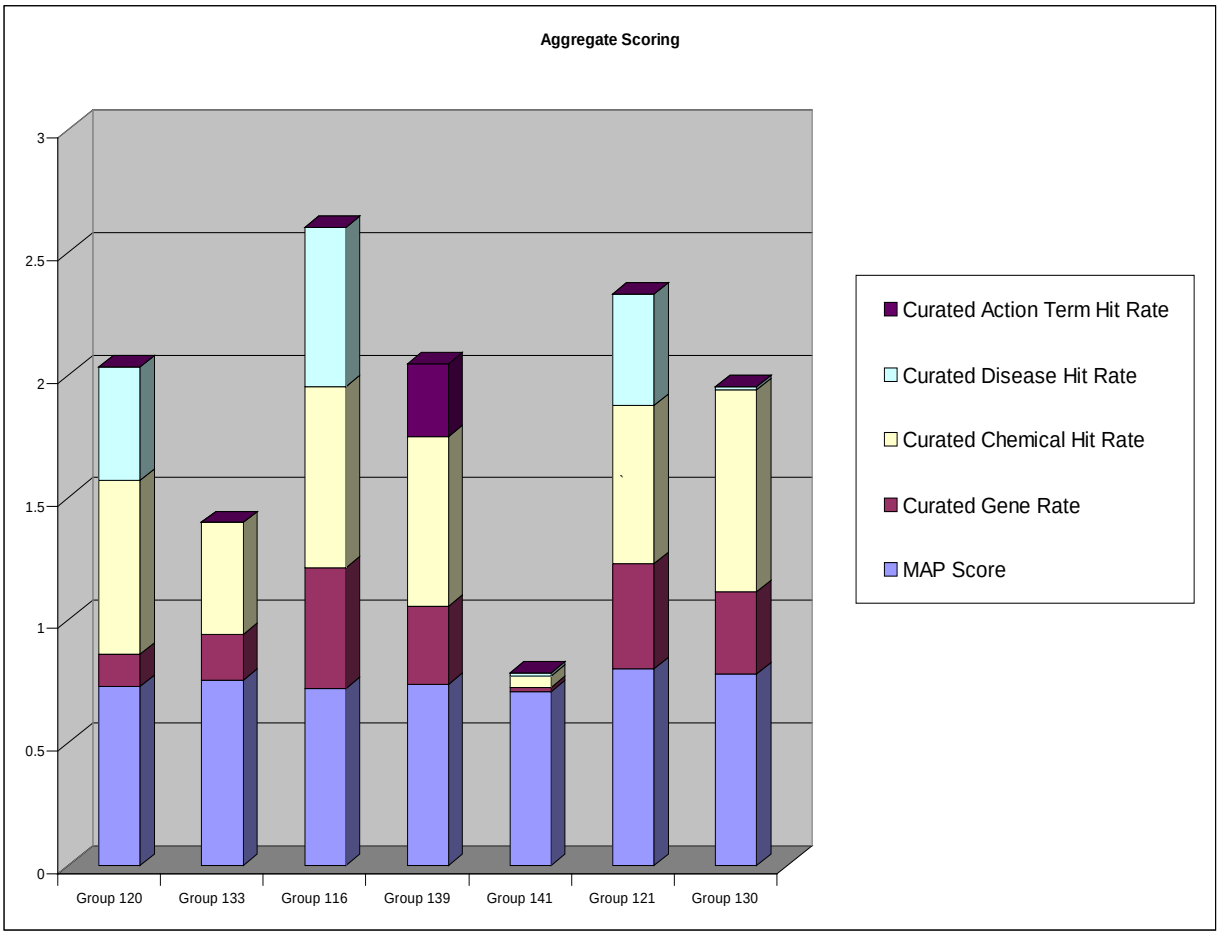

Disease Recognition
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Group 120
Group 133
Group 116
Group 139
Group 141
Group 121
Group 130

Figure 6.


Aggregate Scoring
3
2.5
2
1.5
1
0.5
0
Group 120
Group 133
Group 116
Group 139
Group 141
Group 121
Group 130
Curated Action Term Hit Rate
Curated Disease Hit Rate
Curated Chemical Hit Rate
Curated Gene Rate
MAP Score

System Description for the BioCreative 2012 Triage Task

Sun Kim[1], Won Kim[1], Chih-Hsuan Wei[1], Zhiyong Lu[1] and W. John Wilbur[1, *]

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

*Corresponding author: Tel: 301 435 5926, E-mail: wilbur@ncbi.nlm.nih.gov

Abstract

We developed a machine learning framework based on our prior system for the protein-protein interaction article classification task in BioCreative III. To address new challenges in the Triage task, we explored a named entity identification method for genes, chemicals and diseases. In addition, latent topics were analyzed and used as a feature type to overcome the small size of the training set. For entity annotation and a user interface, a web interface for the Interactive TM task, PubTator, was integrated with article prioritization results.

Article Prioritization for the Triage Task

The BioCreative Triage task is to develop tools or systems to help the literature-based curation workflow used in the creation of the Comparative Toxicogenomic Database (CTD). In particular, this task is focused on prioritizing articles to be curated by evaluating chemical-gene, chemical-disease and gene-disease relationships from text. To participate in this task, we assume the Triage task is an extension of the BioCreative III article classification task (ACT), where protein-protein interaction information was the only concern of prioritizing articles. Since both tasks are data-driven and their goals are to find interaction information among specific entities, we basically follow the same framework (1, 2) developed for ACT. The previous method was successful for the ACT task. However, there are several features we should consider for this new task:

1) Target chemicals are explicitly given for training and test sets.
2) Entities to be identified are now chemical, gene, and disease names, which require new methods for syntactic analysis and named entity recognition (NER).
3) The available training set is very limited. In the Triage task, the numbers of positive and negative examples are only 1,031 and 694, respectively.

To tackle these issues, some features are removed from the previous framework and others are introduced for this task.

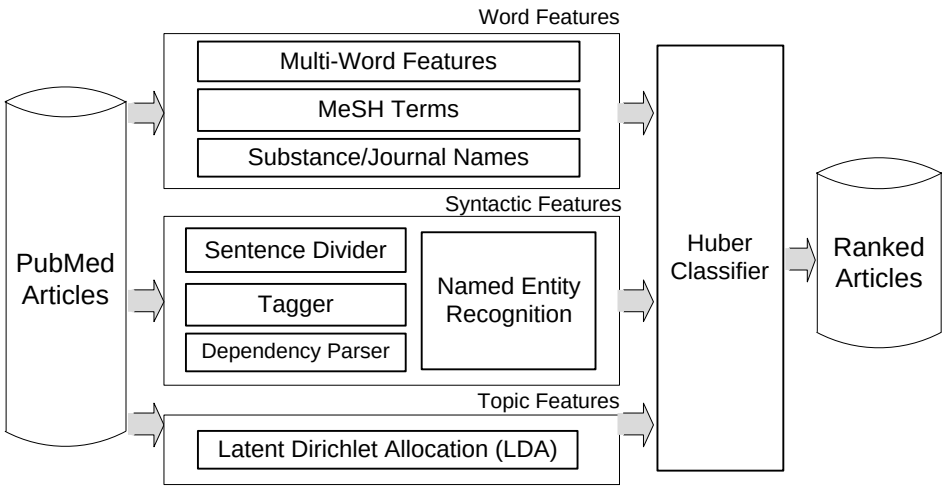


Figure 1. Our article prioritization method.

Figure 1 depicts the overview of our article prioritization method. For input articles, features are extracted in three different ways. One is word features including multi-words, MeSH terms and substance/journal names. The second is syntactic features based on dependency relationships between words. The third is topic features obtained from latent Dirichlet allocation (LDA) (3). After feature extraction procedures, a large margin classifier with Huber loss function (4) is utilized for learning and prioritizing articles. The following subsections describe the three feature types briefly.

Word features from PubMed®

Multi-words are known as n-grams, where n-consecutive words are considered as features. We here use unigrams and bigrams from titles and abstracts. MeSH is a controlled vocabulary for indexing and searching biomedical literature, and MeSH terms are used to indicate the topics of an article. Hence, these terms are included as features. In the Triage task, target chemicals are designated for a set of articles, and journals are treated differently in the CTD rule-based system (5). Therefore, substances and journal names are also extracted from PubMed and used as features.

Syntactic features for identifying entity relationships

This feature identifies interactions or relationships among entities by syntactically analyzing sentences. By using a dependency parser (6), a head word and a dependent word are determined as a two-word combination. Since our goal is to find relationships between two entities, any words indicating relations are likely placed in the head position, whereas their corresponding entities will be placed in the dependent position. Thus, we only consider dependent words as candidate entities, and a NER strategy is used to identify these candidates. For the NER method, we use a vector space approach to modeling semantics (7) and compute our vectors as described in (8). We constructed our vector space using PubMed and an updated version of the SEMCAT database (9), and applied support vector machines to learn how to classify into different entity

types. If two different entity types are found in a sentence, we assume this sentence includes entity-entity relationship information.

Topic features

Along with syntactic features, topic features are newly added to address the Triage problem. There is some evidence that LDA topics can provide features with better generalization properties when there is little training data. We pooled the whole CTD (http://ctdbase.org) and the Triage training set. In our application of LDA, we used the model as put forward in (3) and calculated the model using Markov Chain Monte Carlo (MCMC) simulation as described in (10).

Entity Annotation and User Interface

While entity annotation can be combined with an article prioritization method, our approach does not use fully annotated names for genes, chemicals and diseases. As explained above, the proposed method rather makes a quick decision for single words obtained from dependency parsing. As a result, we currently cannot obtain gene/chemical/disease actors directly from our prioritization system. However, our experimental setup makes individual processes independent. Thus, each module can be replaced with other similar approaches as desired. This applies to our feature selection, machine learning classifiers and even entity/actor annotations.

Since official results should be submitted with actor information as well as prioritized articles, we use PubTator (11) for annotating entities and for providing a web interface for the Triage task. PubTator is a web-based tool which is developed for creating, saving and exporting annotations. PubTator was customized to have a tailored output for combining the results of article ranking and bioconcept annotation.

Experimental Results

Table 1 shows average precisions for different dataset, feature and classifier combinations. The last column is the configuration we used for the Triage task. BC/BC means positive and negative examples from the Triage training set. CTD/BC means positive examples from CTD dataset and negative examples from the Triage training set. Compared to Bayes classifiers (first column), the proposed method improves average precisions up to 5% on average. Note that test examples were always excluded from the training set in either BC/BC or CTD/BC experiments.

Table 1. Article prioritization performance (average precision) for different dataset, feature and classifier combinations.

Dataset	BC / BC			CTD / BC
Feature	Multi-Word Features		Proposed Features	
Classifier	Bayes	Huber	Huber	Huber
2-Acetylaminofluorene	0.7151	0.6812	0.7055	0.6932
Amsacrine	0.5880	0.6676	0.6850	0.7411
Aniline	0.7589	0.7646	0.8000	0.7708
Aspartame	0.3755	0.4520	0.4890	0.5902
Doxorubicin	0.8434	0.8718	0.8689	0.8895
Indomethacin	0.9599	0.9699	0.9761	0.9626
Quercetin	0.9068	0.9176	0.9321	0.9227
Raloxifene	0.7913	0.7940	0.8175	0.7759
Average	0.7424	0.7648	0.7843	0.7933

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

1. Kim, S. and Wilbur, W. J. (2011) Classifying protein-protein interaction articles using word and syntactic features, *BMC Bioinformatics*, **12**(Suppl 8), S9.

2. Kim, S., Kwon, D., Shin, S.-Y., and Wilbur, W. J. (2012) PIE the search: searching PubMed literature for protein interaction information, *Bioinformatics*, **28**(4), pp. 597-598.

3. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, pp. 993-1022.

4. Zhang, T. (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms, *Proceedings of the 21st International Conference on Machine Learning*, pp. 919-926.

5. Summary of curation details for the Comparative Toxicogenomics Database, http://www.biocreative.org/tasks/bc-workshop-2012/Triage

6. Curran, J. R., Clark, S., and Bos, J. (2007) Linguistically motivated large-scale NLP with C&C and Boxer, *Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo)*, pp. 33-36.

7. Turney, P. D. and Pantel, P. (2010) From frequency to meaning: vector space models of semantics, *Journal of Artificial Intelligence Research*, **37**, pp. 141-188.

8. Pantel, P. and Lin, D. (2002) Discovering word senses from text, *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 613-619.

9. Tanabe, L. et al. (2006) SemCat: semantically categorized entities for genomics, *AMIA Annual Symposium Proceedings*, pp. 754-758.

10. Griffiths, T. L. and Steyvers, M. (2004) Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**, pp. 5228-5235.

11. Wei, C.-H., Kao, H.-Y., and Lu, Z. (2012) PubTator: A PubMed-like interactive curation system for document triage and literature curation, *Proceedings of the BioCreative Workshop 2012*.

Ranking of CTD articles and interactions using the OntoGene pipeline

Fabio Rinaldi, Simon Clematide and Simon Hafner
Institute of Computational Linguistics, University of Zurich
{rinaldi,siclemat}@cl.uzh.ch,hafnersimon@gmail.com

Abstract

In this paper we briefly describe the architecture of the OntoGene Relation mining pipeline and its application in the task 1 of BioCreative IV. The aim of the task is to deliver information useful for the triage of abstracts relevant to the process of curation of the Comparative Toxicogenomics Database.

Although the main focus of our text mining research is the extraction of interactions, we decided to participate in the task with the assumption that articles which contain relevant interactions would be relevant themselves.

We use a conventional information retrieval system (Lucene) to provide a baseline ranking, which we then combine with information provided by the relation mining module, in order to achieve an optimized ranking.

1 Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt [14] collects information on all known proteins. IntAct [4] is a database collecting protein interactions. The Comparative Toxicogenomics Database collects interactions between chemicals and genes in order to support the study on the effects of environmental chemicals on health [6]. Most of the information in these databases is derived from the primary literature by a process of manual revision known as "literature curation". Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The work presented here is part the OntoGene project[1], which aims at improving biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [11, 9]. The results of the entity detection feed directly into the process of identification of interactions. The syntactic parser [13] takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

Recently, in the context of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), the OntoGene group has developed a user-friendly interface (ODIN: OntoGene Document INspector) which presents the results of the text mining pipeline intuitive fashion, and allows a deeper interaction of the curator with the underlying text mining system.

In the rest of this paper we first explain how our existing OntoGene relation mining system has been customized for CTD (section 2) and then how it has been integrated with a conventional IR system (Lucene) for the purpose of the Triage task (section 3). We also provide a brief overview of our ODIN curation interface (section 4) and a preliminary evaluation of the results obtained so far (section 5)

[1] http://www.ontogene.org/

2 Methods

In this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions. We describe then in some detail a machine learning approach used to obtain an optimized scoring of candidate interactions based upon global information from the whole CTD.

2.1 Preprocessing and Detection of Domain Entities

In order to solve the triage task, we processed the PubMed abstracts of the referenced articles by the OntoGene Text Mining pipeline.
As shown in our previous work [2], the inclusion of PubMed metadata, such as the list of chemical substances, as well as the annotated MeSH descriptors and qualifiers, improves the detection of important relations and enhances term recognition coverage. Therefore, we added these metadata from the PubMed XML files as a textual list at the end of each abstract. In our OntoGene text mining pipeline, the sentence and token boundaries of the enriched abstracts are identified using the LingPipe framework[2]
Next, we describe our approach to the problem of detecting names of relevant domain entities in biomedical literature (genes, chemicals and diseases for CTD) and grounding them to widely accepted identifiers assigned by the original database. Terms, i.e. preferred names and synonyms, are automatically extracted from the original CTD databases and stored in a common internal format, together with their unique identifiers (as obtained from the original resource). An efficient lookup procedure is used to annotate any mention of a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the reference term – no further disambiguation on concepts is done. For more technical details of our term recognizer, see [8].

2.2 Detection of Interactions

The information about mentions of relevant domain entities (and their corresponding unique identifiers) can be used to create candidate interactions. In other words, the co-occurrence of two entities in a given text span (typically one or more sentences, or an even larger observation window) is a low-precision, but high-recall indication of a potential relationship among those entities. In order to obtain better precision it is possible to take into account the syntactic structure of the sentence, or the global distribution of interactions in the original database. In this section we describe in detail how candidate interactions are ranked by our system, according to their relevance for CTD curation, by exploiting the vast amount of curated articles in the CTD base.
An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only:

$$relscore(e_1, e_2) = (f(e_1) + f(e_2))/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities e_1 and e_2 are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. We know from our previous experiments [9] that giving a "boost" of 10 to the entities contained in the title produces a measurable improvement of ranking of the results. This simple approach can be further optimized if we apply a supervised machine learning method for scoring the probability of a term to be part of an interesting relation. There are two key motivations for scoring concepts based upon relation candidate ranking: First, we need to adapt to highly-ranked false positive relations which are generated by a simple frequency based approach by frequent but uninteresting concepts. The goal is to model some global properties of the curated CTD relations. Second, we want to penalize false positive concepts that our term recognizer detects. In order to deal with such cases, we need to condition the concepts by their normalized[3] textual form t. The combination of a term t and one of its valid entities e is noted as $t:e$.

[2]More information regarding the framework can be found at `http://alias-i.com/lingpipe`.
[3]A normalized textual form of a term consists of the sequence of lower-case alphanumeric characters of all term tokens.

Next we define a predicate $gold(A, e)$ which is true (i.e. 1) for an article A if there is at least one relation in the gold standard where entity e is part of, and false (i.e. 0) otherwise. We estimate the overall probability $P(\, gold(A, e) = 1 \mid t\!:\!e\,)$ with the help of the Maximum Entropy Modeling tool *megam* [3]. For training we use the set of CTD-referenced PubMed articles having not more than 12 manually curated relations[4], additionally removing all articles which are part of the BioCreative training and test set for the respective data set[5].
For unseen normalized terms t, i.e. terms not present in the training data, the maximum entropy classifier would assign a low default probability based on the distribution of all training instances. However, we can specify better back-off probabilities if we take into account the admissible entity/entities e of term t. Our current back-off model works as follows: if the entity e of an unseen t is seen in the article, the averaged probability of all seen term-entity pairs is used. Otherwise, the averaged probability of all entities of the same type as e is used.
The score of an entity e in an article A is the sum of all zoned term frequencies[6] weighted by their gold probability:

$$score(e) = \sum_{t:e \in A} f(t\!:\!e) \times P(\, gold(A, e) = 1 \mid t\!:\!e\,)$$

Having determined the individual score for each entity e, we compute the relation score as the harmonic mean of its component scores:

$$relscore(e_1, e_2) = 2 \times \frac{score(e_1) \times score(e_2)}{score(e_1) + score(e_2)}$$

In preceding work on relation ranking [2], the relation score was taken as a sum of the concept scores. By performing systematic cross-validation experiments on all CTD articles, we noticed that using the harmonic mean improves the results considerably. In order to make the relation scores comparable between different articles we normalize all relation for a given BioCreative data set.

3 Integration with a standard IR system

A conventional IR system is used to provide a baseline document classification. Information derived from the OntoGene pipeline, and from the ranking process described in the previous section, is then added as additional features in order to improve the baseline ranking generated by the IR system. The integration of the various components is performed using mainly JRuby (and some small parts in Java).

3.1 Terminology-aware tokenization

Documents are processed by Lucene in the conventional way, selecting different boost values for title and abstract (10 for title, 3 for abstract, just as in the CTD reference system). The Lucene API is accessed via JRuby. Changes in the boost values did not show any statistically significant change in the MAP scores, because most of the information is in the abstract, not the title. The existence of relevant information in the title typically implies relevant information in the abstract.
The only significant technical change to Lucene preprocessing is the replacement of the "StandardAnalyzer" component (which is the default analyzer for English, responsible for tokenization, stemming, etc.) with our own tokenization results, as delivered by the OntoGene pipeline. The advantage of this approach is that we can flexibly treat recognized technical terms as individual tokens, and map together their synonyms [7]. In order words, after this step all known synonyms of a term will be treated as identical by the IR system.
The "StandardAnalyzer" component is replaced by a simple transformation of the XML output of the pipeline into a format suitable for internal processing by Lucene. In particular tokens and terms as recognized by the pipeline are transformed into Lucene "token" data objects. Whenever a domain entity (denoted by the `Term` element in the XML representation) is found, its words are concatenated to one token. At the same position, a new token with the text of the concept identifier is added to the stream.
As an example:

[4] The threshold of 12 relations is motivated by the observation that the more relations an article has the less probable it is to find them by processing the abstracts only.
[5] This results in 22319 articles for the BioCreative 4 training set, containing 69320 curated relations. For the BioCreative 4 test set, we used 22825 articles with 71064 relations.
[6] As mentioned earlier, occurrences in the title are counted 10 times.

```
<W C="VBN" id="W151" o1="758" o2="767">inhibited</W>
<Term allvalues="MESH_D015232:chem" id="TW152W153"
      matched="prostaglandine2" type="chem">
  <W C="NN" id="W152" o1="768" o2="781">prostaglandin</W>
  <W C="NN" id="W153" o1="782" o2="784">E2</W>
</Term>
<W C="NN" id="W154" o1="785" o2="794">synthesis</W>
```

would be converted to the following (square brackets denote token boundaries):

```
[inhibited] [prostaglandin E2] [synthesis]
            [MESH_D015232]
```

Synonymous terms (as identified by the pipeline) are mapped to their unique identifiers (for this experiment the term identifier provided by the CTD database, which happens to be a MeSH term in the example above). A basic search is conducted by mapping the target chemical to the corresponding identifier, which is then used as a query term to perform a search in Lucene.

3.2 Relation-based query expansion

As described in section 2.2 the OntoGene pipeline is not only used in order to deliver an optimized tokenization, it can also be used to generate candidate interactions, which could be directly used for curation purposes by CTD curators.
Although the definition of the task did not require the participants to deliver candidate interactions, we worked under the assumption that documents which contain relevant interactions would be relevant themselves. When another term is often seen in relation with the target term, it is probably important for the target. This statistical information is used to adjust the ranking of the documents.
The OntoGene pipeline delivers candidate interactions as part of its standard output for each single document. Each interaction is assigned a score in the interval (0,1]. The relations are extracted from all the files in the document set assigned to the target chemical by the organizers. All relations which involve a term equivalent to the target (the target or one of its synonyms) are extracted. The interacting entity (the second term in those interactions) is then added to the search query, for each interaction, giving rise to an expanded query. The additional query terms are weighted according to the normalized score of the original interaction. As an example suppose two documents contain the interactions listed in the first two columns below (document 1 and document 2):

```
document 1:    document 2:       expansion terms:
A C 1          A B 1             C 1    from doc 1
B C 0.7        B D 0.42          B 0.75 from doc 1 (score 0.5) and doc 2 (score 1)
A B 0.5                          D 0.4  from doc 1
A D 0.4
```

If the target term is A, the relations marked in boldface are relevant, which gives us new search terms to be added to the query, listed in the 3rd column with their normalized weights (sum of scores divided by the number of relations). The original target term is given a weight which is above the weight of the relations in order to make it clearly more relevant than any of the added terms. We have experimentally verified on the training data that the using this query expansion process improves the average MAP scores from 0.62225 to 0.694625 (i.e. an improvement of nearly 12%).

4 The ODIN Interface

The results of the OntoGene text mining system are made accessible through a curation system called **ODIN** ("OntoGene Document INspector") which allows a user to dynamically inspect the results of their text mining pipeline. A previous version of ODIN was used for participation in the 'interactive curation' task (IAT) of the BioCreative III competition [1]. This was an informal task without a quantitative evaluation of the participating systems. However, the curators who used the system commented extremely positively on its usability for a practical curation task. An experiment in interactive curation has been performed in collaboration with curators of the PharmGKB database [5, 12]. The results of this experiment are described in [10], which also provides further details on the architecture of the system.

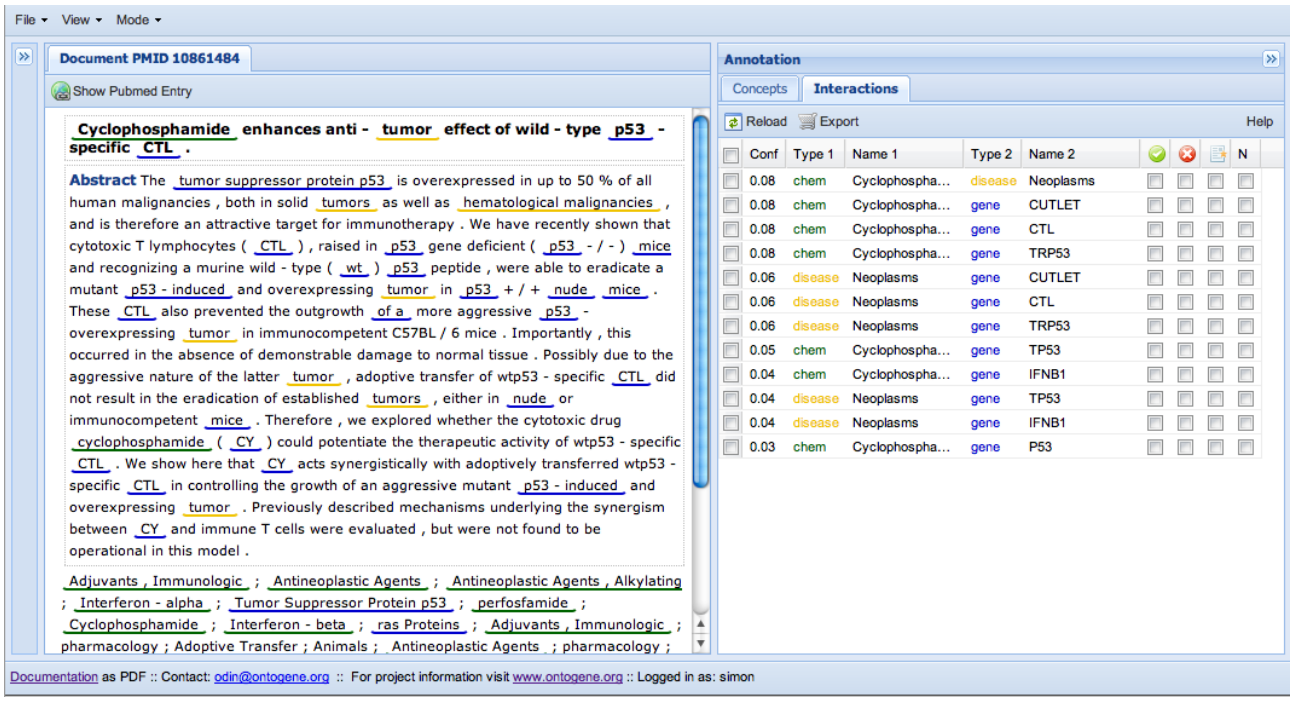


Figure 1: Entity annotations and candidate interactions on a sample PubMed abstract

More recently, we partially adapted ODIN to the aims of CTD curation, allowing the inspection of PubMed abstracts annotated with CTD entities and showing the interactions extracted by our system. We would be interested in providing further customizations according to the needs of the CTD curation process.

5 Evaluation

In order to generally assess the upper limit of our relation recognition system, we evaluated the coverage of the term recognizer on all CTD-referenced articles containing at most 12 curated relations. The table below describes the coverage of term recognition for concepts and relations in experimental data and shows that we find about 3/4 of all entities. However, the upper limits for relation detection are not the same for all relation types. The coverage of relations involving chemicals have substantially lower coverage rates which seems a bit unfortunate for the CTD triage task.

Cat	N	abs	rel
disease	12639	9502	75.18
chemical	38523	30129	78.21
gene	39150	29199	74.58
TOTAL	90312	68830	76.21
dis-gen	6956	5126	73.69
che-dis	12154	8356	68.75
che-gen	52746	34883	66.13
TOTAL	71856	48365	67.31

The table below shows the final results obtained on the training set using the on-line evaluation tool. Due to lack of space and time, we cannot report here a detailed analysis of all intermediate results, which we intend to present at the workshop.

Term	MAP	genes	chemicals	diseases
doxorubicin	0.800	0.167	0.843	0.793
indomethacin	0.936	0.331	0.834	0.725
raloxifene	0.798	0.244	0.818	0.778
amsacrine	0.655	0.603	0.689	0.500
aniline	0.543	0.625	0.561	0.524
2-Acetylaminofluorene	0.643	0.412	0.845	0.421
aspartame	0.365	0.686	0.756	0.720
quercetin	0.853	0.463	0.646	0.653

6 Conclusions

In this paper we have described our approach towards ranking biomedical abstracts for the triage task of the CTD curation process. The peculiarity of the approach is that it gives priority to the identification of candidate interactions, which are then used as additional weighting factors in a conventional IR-based system.
The OntoGene pipeline is capable of delivering all information relevant to CTD curation: entities with their database references, interactions, and interaction terms. In the shared task however, due to insufficient time for customization, we decided to exclude the computation of interaction terms. The results of the system are accessible through an intuitive interactive interface, which we are willing to customize for CTD curation.

Acknowledgments
This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis Pharma AG, NIBR-IT, Text Mining Services, CH-4002, Basel, Switzerland.

References

[1] Cecilia Arighi, Phoebe Roberts, Shashank Agarwal, Sanmitra Bhattacharya, Gianni Cesareni, Andrew Chatr-aryamontri, Simon Clematide, Pascale Gaudet, Michelle Giglio, Ian Harrow, Eva Huala, Martin Krallinger, Ulf Leser, Donghui Li, Feifan Liu, Zhiyong Lu, Lois Maltais, Naoaki Okazaki, Livia Perfetto, Fabio Rinaldi, Rune Saetre, David Salgado, Padmini Srinivasan, Philippe Thomas, Luca Toldo, Lynette Hirschman, and Cathy Wu. Biocreative iii interactive task: an overview. *BMC Bioinformatics*, 12(Suppl 8):S4, 2011.

[2] Simon Clematide and Fabio Rinaldi. Ranking interactions for a curation task. *Machine Learning and Applications, Fourth International Conference on*, 2:100–105, 2011.

[3] Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August 2004.

[4] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucl. Acids Res.*, 32(suppl 1):D452–455, 2004.

[5] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the pharmgkb project. *The Pharmacogenomics Journal*, 1:167–170, 2001.

[6] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):689–692, 2006.

[7] Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Mare Koit, Kadri Vider, and Neeme Kahusk. Terminology as Knowledge in Answer Extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02)*, pages 107–113, Nancy, 28–30 August 2002.

[8] Fabio Rinaldi, Kaarel Kaljurand, and Rune Saetre. Terminological resources for text mining over biomedical scientific literature. *Journal of Artificial Intelligence in Medicine*, 52(2):107–114, June 2011.

[9] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.

[10] Fabio Rinaldi, Gerold Schneider, and Simon Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics, special issue for the Biocuration 2012 conference*, 2012. conditionally accepted for publication.

[11] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3, 2006.

[12] Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. Pharmgkb: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551, 2008. PMID: 18949600.

[13] Gerold Schneider. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich, 2008.

[14] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Research*, 35:D193–7, 2007.

Selection of relevant articles for curation for the Comparative Toxicogenomic Database

Dina Vishnyakova[1,2,*], Emilie Pasche[1,2] and Patrick Ruch[1,3]

[1] Bibliomics and Text Mining (BiTeM) Group: http://bitem.hesge.ch
[2] Division of Medical Information Sciences, University and University Hospitals of Geneva
[3] Information Science Department, HES-SO/University of Applied Science Geneva

*Corresponding author: SIMED; University Hospitals of Geneva; 4, rue Gabrielle-Perret-Gentil; CH-1211 Geneva 14; Tel: +41 22 372 61 99; email: dina.vishnyakova@hcuge.ch

Introduction

We report on the original integration of an automatic text categorization pipeline, so-called ToxiCat (Toxicogenomic Categorizer) to perform biomedical documents classification and prioritization in order to speed up curation of the Comparative Toxicogenomics Database (CTD). The task can be basically described as a binary classification task, where relevance scores are used to ranks a selected set of articles. We design a SVM classifier, which combines four main components: an information retrieval engine for MEDLINE (EAGLi), a biomedical named-entity recognizer based on terminological resources, a gene normalization (GN) service (NormaGene) developed for a previous BioCreative campaign and finally, an ad-hoc keyword recognizer for diseases and chemicals. The main components of the pipeline are publically-available both as web application and web services. The integration performed for the BioCreative competition is available via a web user-friendly interface: http://pingu.unige.ch:8080/Toxicat.

Data and Methods

Data overview

BioCreative 2012 proposes to explore how text mining methods can successfully be applied to practically help biocuration of a large molecular biology knowledge base. The main objective of the Triage-I task is to explore how a set of MEDLINE records, directly retrieved from PubMed using the name of a particular chemical compound can be ranked to prioritize the most relevant articles. In parallel, competitors are also asked to provide additional annotations of interest to maintain the Comparative Toxicogenomics Database (CTD) such the interacting entities (small molecules and gene products) and the pathologies likely to reflect the toxicity of the chemical compound. . We first adopted an analytical view over the data. We selected four chemicals and a sample of 1059 articles out of the training data provided by the organizers. The distributions for the four chemicals are shown in Table 1. Approximately half of these articles in the benchmark contain no information about the chemicals or no information about the genes they are supposed

to link to. Symmetrically, only half of the articles have information about both a gene and a chemical, and only a few have information about diseases. The distribution of entity types in the benchmark is shown in Table 2.

Table 1. Distribution of curated articles for each chemical in the selected sample.

Chemical Name	Number of articles per chemical	Curated articles per chemical in %
Raloxifene	270	60
2-Acetylaminofluorene	178	45,5
Amsacrine	69	53
Quercetimin	542	77

Table 2. Distribution of entities in the selected sample

Entity Name	Number of articles
Chemicals	654
Genes	643
Diseases	28
Co-occurrence of genes and chemicals	602
Input chemical in titles	381

Methods

We designed a SVM classifier for the binary classification of articles (with two classes: curated and not curated). The classifier returns a Boolean value together with a class estimate, which directly expresses the probability to belong either to the positive or the negative class. Features selected to build the classification model are presented in Table 3.

Features can be split within three subsets. The first feature set contains information about MeSH terms of articles extracted from the MEDLINE library, which is locally stored to be indexed by the EAGLi's engine. The use of EAGLi has two main advantages when compared to PubMed's e-Utils: 1. the response time is significantly improved from about one second per PMID for PubMed's e-Utils to an average of 50 ms for EAGLi, which results in an overall processing time

at least one order of magnitude faster; 2; recently published articles are not yet indexed with MeSH, while EAGLi offers the possibility to obtain to automatically index those articles; see [1] for a presentation of the service and [2] for a comparison against similar systems such as MetaMap. According to our observations, curated articles are usually indexed with major descriptors such as pharmacology, toxicity, drug therapy, metabolism, drug effects, chemistry and chemical synthesis. Very often, the indexing with MeSH also normalizes the name of the main chemical with a unique identifier (or preferred term) discussed in the article (i.e. raloxifene or amsacrine).

The second subset of features is obtained via the NormaGene named-entity normalizer [3]. This gene and protein named-entity recognizer was developed for the BioCreative III task to address the gene normalization task [4]. Like other named-entity recognizer, it identifies the boundaries of the gene and protein name but it also attempts to assign a unique identifier at the level of the sequence. Thus, the system attempts to recognize also the species discussed in the article to ultimately link the gene and protein name with a unique sequence. For gene name detection, we must handle lexical ambiguity (homonyms describing two different gene products) and synonymy phenomena. Internally, NormaGene is able to recognize all gene candidates stored in the Gene Protein Synonyms DataBase (GPSDB) [5], as well as all species stored in NEWT [6], which is appropriate to annotate contents for UniProt/SwissProtKB but which exceeds the coverage of CTD. The internal gene and gene product names of NormaGene are therefore reduced to curate CTD. Thus, results returned by NormaGene are compared to the CTD genes controlled vocabulary. If the entities recognized by NormaGene are found in the CTD genes' vocabulary then we extract all synonyms based on t he approved genes ID and match them against the abstract. Indeed, gene and protein identifiers suggested by NormaGene cannot always be explicitly found in the body of the input document as NormaGene uses a generative model, which exploits also functional similarities [7] and not only textual similarities. Gene names used by CTD are imported from EntrezGene. Unlike EntrezGene, a g ene is mostly species-independent, which relaxes the species recognition constrain

Table 3. Selected features for the SVM Classifier.

Features	**Values**
Input chemical compound in the abstract	binary
Input chemical compound in the title	binary
Normalized Input chemical in MeSH terms	binary
Appearance of chemical compounds in the abstract	binary
Frequency of chemical compounds in the abstract	integer
Frequency of main chemical in an abstract	integer

Input chemical detected in first 3 sentences[1]	binary
Appearance of gene/gene product names in the abstract	binary
Frequency of detected genes	integer
Appearance of chemicals and genes in the abstract	binary
Co-occurrence of genes and chemicals in a sentence	integer
Co-occurrence of main chemical and genes in a sentence	integer
Appearance of diseases in the abstract	binary
Journal name relevant for CTD task	binary
Appearance of "pharmacology", "toxicity", "drug therapy", "metabolism", "drug effects", "chemistry" and "chemical synthesis" as the main MeSH terms of the article	binary
Score of overall features	integer(1..100)

The third set of features is an ad-hoc keyword recognizer for diseases and chemicals. This keyword recognizer is based on the controlled vocabularies provided by CTD. We discovered that CTD vocabulary for chemicals contains several chemicals, which seems irrelevant for the curation task. However, the description of CTD vocabularies says that several branches of the original MeSH vocabulary were excluded from CTD's chemical and disease vocabularies [8] because of their weak relevance to CTD tasks. Nevertheless, we discovered that in the vocabularies provided by the organizers, we have all these branches. It was both challenging and risky to decide a priori which branches should be excluded. Thus, we created a Word-Sense Disambiguator (WSD) to filter the detected candidates. This Word-Sense Disambiguator is based on the definitions of the UMLS Semantic Types [9]. The positive semantic types, we are interested were manually defined (e.g.: T114-"Nucleic Acid, Nucleoside, or Nucleotide", T116-" Amino Acid, Peptide, or Protein", T196-"Element, Ion, or Isotope"). This step eliminates non relevant types of chemicals and diseases. In parallel, in order to eliminate common English words from the list of candidates, we created a common English word recognizer based on general-purpose English corpora. Thus, unspecific disease and chemical names were directly discarded.

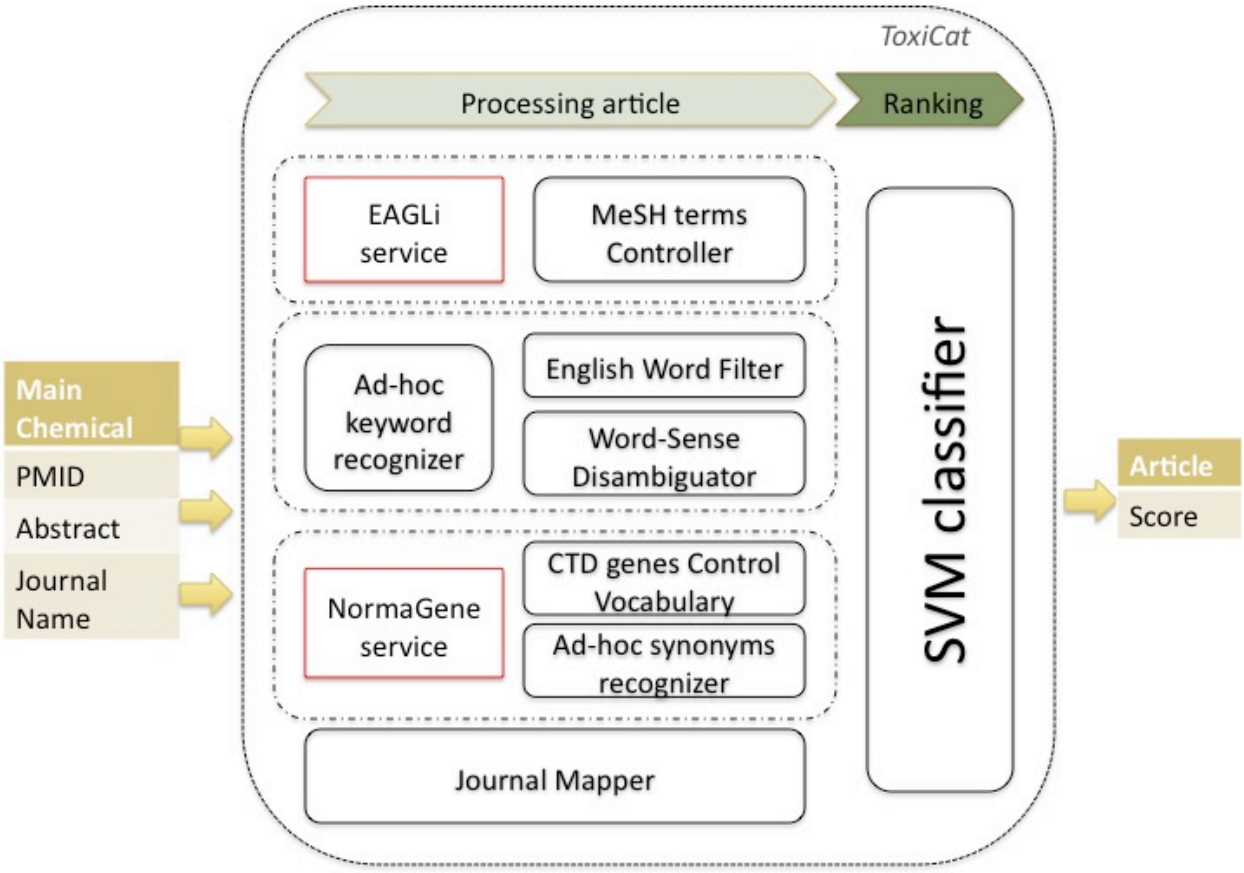


Figure 1. The workflow of ToxiCat.

The general architecture of the ToxiCat workflow is shown in Figure 1. Articles data such as the PMID, the abstract and the journal name are passed to ToxiCat. The PMID is used to query EAGLi's services in order to retrieve all MeSH terms of the article. The abstract of an article is also passed to the ad-hoc keyword recognizer to detect chemicals and diseases candidates. Those candidates are then filtered by the common English word filter and finally by the Word-Sense Disambiguator. In parallel, NormaGene detects the genes' names in the abstract and pass them to the CTD genes Control Vocabulary, which is going to filter out not relevant genes. Remaining gene identifiers are sent to the ad-hoc synonyms recognizer, to detect all synonym names in the abstract. The journal Mapper checks the name of the journal against a list of domain-relevant journal names. Finally, the resulting bag of features is sent to the SVM classifier. The SVM classifier returns a score. This score directly expresses the probability if the article is relevant or not to annotate CTD. Finally, the system's inputs/outputs are shown in the Figure 2 for a couple of input PMIDs and the antibiotic "fluoroquinolone".

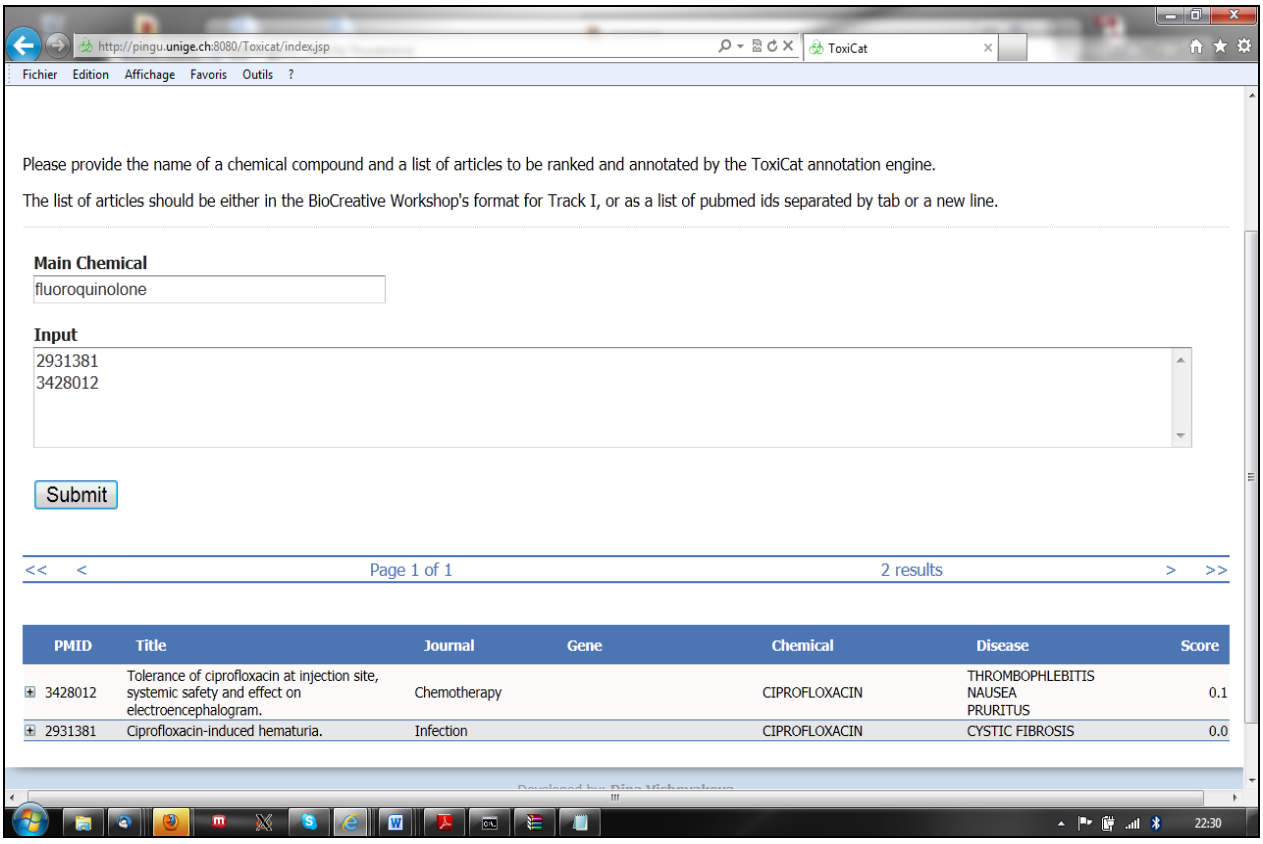


Figure 2: Online page of ToxiCat

Results and Conclusion

Table 4 lists results of our NormaGene system evaluated with BioCreative III data. In Table 4, we display results according to the so-called Threshold Average Precision (TAP-k) on BioCreative III benchmarks: 507 articles curated by the teams who participated in BCIII, 50 manually curated articles, and 50 articles, which were both manually curated and curated by participating teams [10].

Table 4. Results of NormaGene on the cross-species GN task using the BC III benchmarks.

TAP-k	Manual curation/ 50 articles	Mix of manual and teams submissions/ 50 articles	Teams submissions/507 articles
5	0.2236	0.31	0.46
10	0.2324	0.3357	0.46
20	0.2397	0.3357	0.46

From the BioCreative 2012 data, we evaluated the effectiveness of our ad hoc terms recognizer for diseases. Our methods achieved 95% of precision and 92% of recall when tagging diseases in the training sample.

Table 5. Results of ToxiCat for the task-I of BioCreative 2012.

Chemical/Quantity of articles	Intermediate MAP Score	Curated Gene	Curated Chemical	Curated Disease
Urethane/204	0.637	0.08	0.705	0.3
Phenacetin/86	0.831	0.203	0.676	0.5
Cyclophosphamide/154	0.716	0.117	0.747	0.582

In order to tune our binary classifier, we performed ten folders cross-validation and achieved accuracy of 80.5%. We applied the optimal model on the test data and obtained an accuracy of 77%, which suggests some moderate overfitting phenomena. The results of ToxiCat of test data provided by BioCreative 2012 are provided in Table 5.

Table 6. Results of evaluation performed by ToxiCat with different input for gene detection on the test data of task-I of BioCreative 2012.

Chemical/Quantity of articles	Intermediate MAP Score	Curated Gene	Curated Chemical	Curated Disease
Urethane/204	0.632	0.131	0.705	0.3
Phenacetin/86	0.830	0.295	0.676	0.5
Carcinoma/154	0.710	0.191	0.747	0.582

According to results in Table 5, the curated gene score relatively low compared to chemicals and disease scores, which confirms that gene and gene product recognition seems more challenging than recognition of other biomedical entity recognition tasks such as chemicals. From the official results, see Table 5, we excluded some gene candidates, which were highly ambiguous and were associated with several NCBI Identifiers. After the competition, we changed some of the parameters for the gene detection task, and removed the restriction of the overlapping gene names. The recomputed results for the gene detection subtask are found in Table 6. It is obvious that overlapping gene candidates in the final results do improve the “Curated Gene” score but at the same time they decrease the “Intermediate MAP” score. This can be explained by the changes in the classification model, which reduces the ranking score of input articles, when a large amount of gene candidates is found.

Although current results seems suggesting that text mining can effectively help curation tasks by providing access to more relevant contents, it is worth noticing that the effectiveness of ToxiCat is obtained by trading some other dimensions of the components integrated in the pipeline. When

designing the system, we somehow customize a rather generic text processing pipeline (a search engine, EAGLi, a g ene named-entity normalizer, NormaGene, and several terminological resources such as GPSDB…) to answer the specific needs of CTD. Such a step seems both rationale and empirically effective; however it questions the role of the end-user platform. Indeed, if the system must help the professional annotator to curate CTD by basically speeding up annotation, then a system like ToxiCat might be suitable. On the opposite if the system should help curating non us ual contents or novel chemical products, then the system is very likely inappropriate. Ultimately, if the system was to be used as the sole capturing tool for CTD curators, then it may hinder the annotation of new interacting genes, which are not yet listed in CTD as by design non-CTD genes are penalized by the system.

In conclusion, ToxiCat showed competitive performance, in particular for the recognition of chemical compounds. Intermediate MAP score showed that the selected SVM model produced promising results on the test data. In contrast, the identification of pathologies seems nearly as difficult as the recognition of genes and gene products. Further experiments are needed to explain where is the power of the Toxicologenomic Categorizer (ToxiCat) as well as to start explaining the differences observed regarding the recognition power of some of the entity types.

References

1. Ruch P. (2006) Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics. 22(6):658-64
2. EAGLi [http://eagl.unige.ch/EAGLi/]
3. NormaGene [http://pingu.unige.ch:8080/NormaGene]
4. Vishnyakova D, Ruch P. et al. (2011) The gene normalization task in BioCreative III. BMC Bioinformatics, 12(Suppl 8):S2
5. Pillet V, Zehnder M, Seewald AK, Veuthey AL, Petrak J. (2005) GPSDB A new database for synonyms expansion of gene and protein names. Bioinformatics. 21(8):1743-4.
6. NEWT [www.ebi.ac.uk/newt/]
7. Ehrler F, Jimeno A, Geissbühler A, Ruch P (2005) Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot. BMC Bioinformatics, 6(suppl 1):s23.
8. Davis AP, Wiegers TC, Murphy CG, and Mattingly CJ. (2011). The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. Database, Oxford.
9. Jimeno-Yepes A., Mclnnes B., Aronson A. (2011) Collocation analysis for UMLS knowledge-based word sense disambiguation. BMC Bioinformatics, 12(Suppl 3):S4
10. Lu Z., Wilbur W J., et al. (2011) The gene normalization task in BioCreative III. BMC Bioinformatics, 12(Suppl 8):S2
11. Ruch P., Boyer C., Chichester Ch., Tbahriti I., Geissbühler A., Fabry P., Gobeill J., Pillet V., Rebholz-Schuhmann D., Lovis C., Veuthey A-L. (2007) Using argumentation to extract key sentences from biomedical abstracts. I. J. Medical Informatics 76(2-3): 195-200

[i] We tried to use features generated by an argumentative classifier [11] but preliminary experiments were inconclusive.

CoIN: a network exploration for document triage

Yi-Yu Hsu and Hung-Yu Kao*

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

*Corresponding author: Tel: 886 6 2757575 ext 62546, E-mail: hykao@mail.ncku.edu.tw

Abstract

Identifying the interaction type of two biomedical terms is an important task for the curation and text mining in the biomedical field. When different biomedical terms co-occur in a sentence, there are several interaction types between them. In view of these interaction types, biocurators can assess the manner in which two biomedical terms are associated. Co-occurrence Interaction Nexus (CoIN) employs the co-occurrence relations and their network centralities to evaluate the influence of biomedical terms from Comparative Toxicogenomics Database (CTD).

Introduction

Co-occurrence Interaction Nexus (CoIN) is a web-based system that facilitates biocurators to assess articles according to their term correlations among sentences. In recent years, there has been a dramatic increase on the issue of assisting manual curation (1). Manual curation plays an important role in supporting basic analyses for advanced research, and BioCreative 2012 focuses on the integration of biocurations. Hence, CoIN aids for the specific curation task: document triage.

CoIN is designed for searching relevant documents by the relations of co-occurrence from query articles, and CoIN is equipped with network analyses. Although word and syntactical features are the basic components for searching the patterns of domain knowledge, there are still many restrictions on name recognition, in particular the problem of training sets, such as imbalance and shortage. To overcome the barrier of incomplete data, CoIN adopts the heteregeous network to tackle the triage task.

The use of network analyses has experienced different research topics, such as phylogenetics, function predictions, human diseases, and drug developments (2,3). At the same time, the pre-tagging results of CoIN are developed based on the state of the-art named entity recognition tools in BioCreative III (4). Furthermore, we collected dictionary corpora to recognize disease and chemical terms in a sentence level. Therefore, the idea of CoIN is basically generated from the co-occurrence relation of gene, disease, and chemical terms within each sentence of a specific article.

Material and Methods

Curation workflow

For the convenience of biocurators, CoIN allows users to query a gene, disease, or chemical term of a chemical target. As shown in Figure 1, CoIn proceeds to detect gene name by the tagging system and separated the articles into sentences (4). Next, CoIN employs the chemical and disease dictionaries to train a conditional random field (CRF), which is a statistical modeling method often applied in pattern recognition.

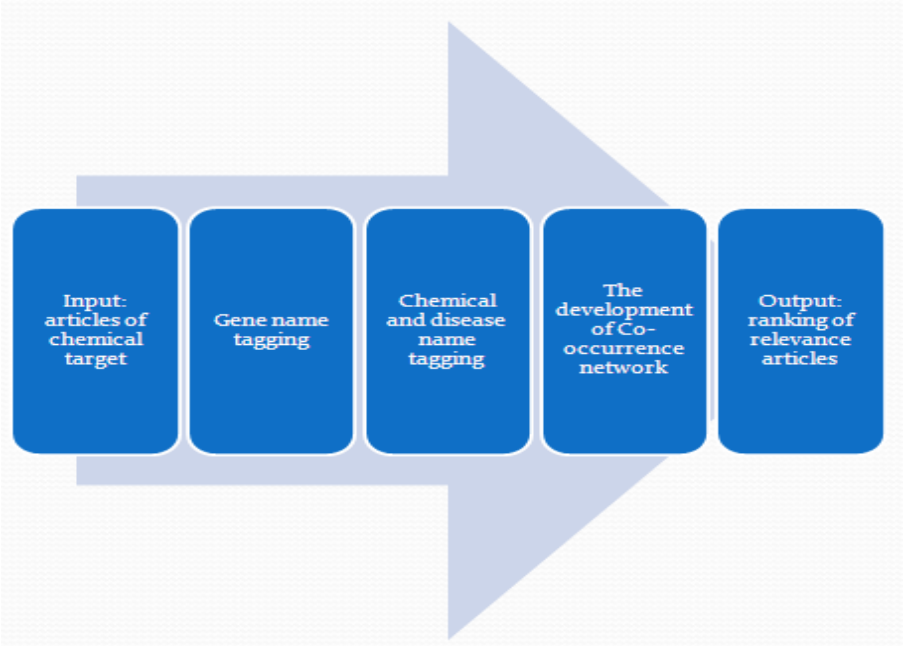


Figure 1: The workflow of CoIN.

After collecting the tagging terms, CoIN calculates the co-occurrence of tagging terms for each sentence, and then the co-occurrence network is constructed using the information of gene-disease, gene-chemical, and chemical-disease interactions, as shown in Figure 2. In the last stage of CoIN, the system provides the normalized co-occurrence frequency, betweenness and PageRank value for prioritizing the list. In the section of text mining methods, we introduce the normalized co-occurrence frequency, betweenness and PageRank respectively.

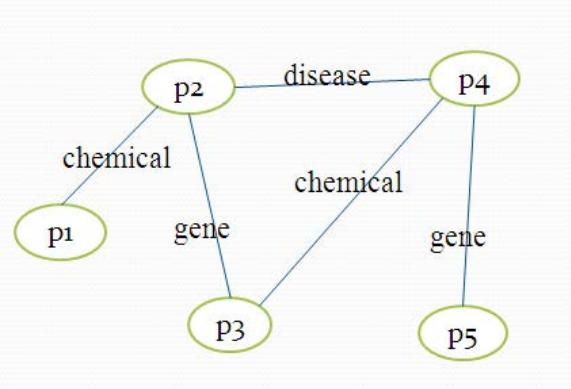


Figure 2: a toy example of co-occurrence interaction network. P_i means PubMed articles. If any sentence shares the co-occurrence of gene, disease, and chemical terms between two articles, the edge is established.

For example, we use the pubmed articles in phenacetin as a resource; otherwise you can also input a gene, disease, or chemical name, as shown in Figure 3. After the computation is finished, we can obtain a ranking list, as shown in Figure 4. In view of the system schema of CoIN, the name recognition process is usually time-consuming and large-scale. However, CoIN provides a quick sorting result to biocurators after the name recognition process is determined. That is, CoIN takes less time in training complex features, but the system returns the ranking result by the network properties of co-occurrence network immediately. However, much interaction data accompanies with many noises, and an overestimation is however caused by the overlapping interactions. In this case, we proposed a normalized co-occurrence frequency to lead biocurators into a easier way for evaluating the quantity of their target. Nevertheless, the normalized co-occurrence frequency is incapable of the propagation information.

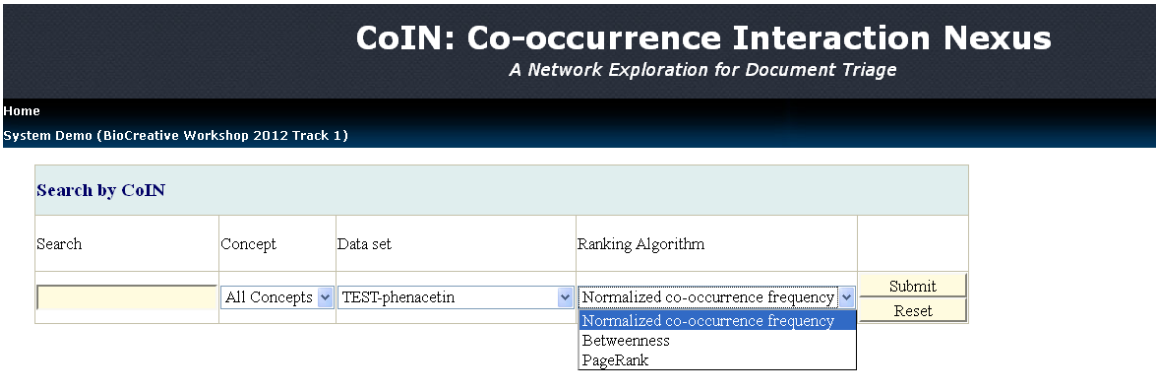


Figure 3: The input screen of CoIN

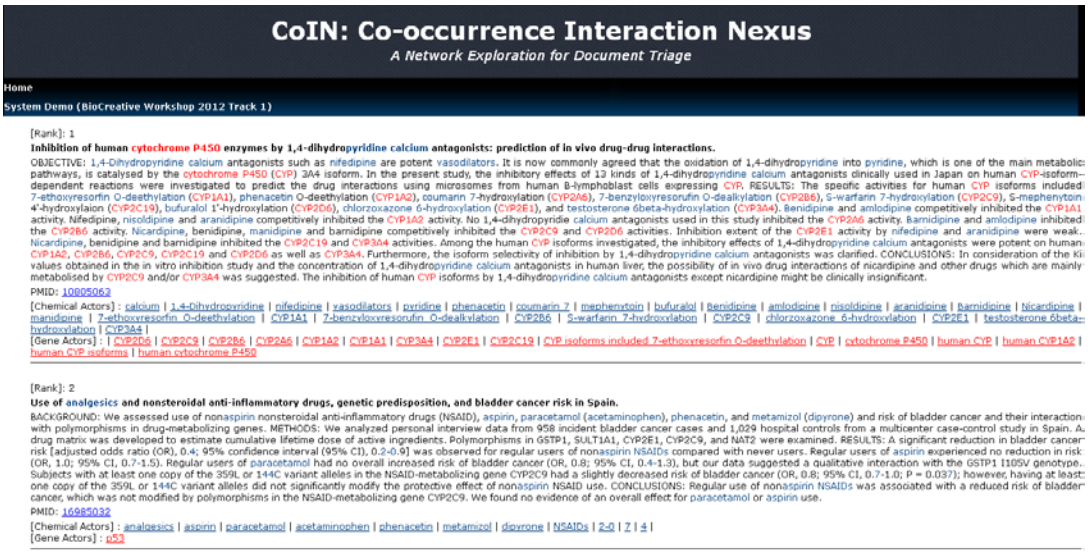


Figure 4: The output screen of CoIN

Text mining methods

In this section, we describe the methods which are applied to CoIN. The ranking algorithms of CoIN are designed for co-occurrence frequency-based and network-based approaches. The co-occurrence frequency approach evaluates the co-occurrence relations when the network-based approach estimates the linking relations. In the following, we simply introduce the normalized co-occurrence frequency because it is derived from co-occurrence frequency.

Normalized co-occurrence frequency

The co-occurrences of gene-disease, gene-chemical, and chemical-disease frequency are usually overestimated, so we normalize the co-occurrence frequency by the standard score z as follows. The standard score z is defined as follows.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where:

x is a raw score to be standardized;
μ is the mean of the population;
σ is the standard deviation of the population.

After the standard score z is calculated, we aggregate the standard score z of gene-disease, gene-chemical, and chemical-disease frequency as the normalized co-occurrence frequency.

Betweenness

A number of studies use topological centralities to make tentative predictions of important vertices in compound networks. In this case, topological centralities are able to measure the global influence of individual proteins. Jeong et al. (5) reported that the protein with high degree are possible to be important proteins in protein-protein interaction (PPI) networks. Yu et al. (6) have demonstrated that a protein with high betweenness centrality are important proteins in yeast PPI networks. More recently, with an increase in network approaches to the study of heterogeneous networks, the accuracy suffers from the incomplete networks and noises. Hence, CoIN focuses on integrating heterogeneous data of gene, disease, and chemical and applying the betweenness centrality to access the articles. The high betweenness centrality means that vertices have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices (7).

PageRank

Google helps users to browse the web pages with keywords quickly, and these pages represent the information of keywords. Therefore, we can organize the important information of keywords by using search engines on t he Internet. PageRank is a famous linking algorithm, which is developed by the founder of Google (8). The PageRank algorithm ranks the page by the linking structure of networks. The PageRank is performed as follows.

$$PR(V_i) = (1-d) + d * \sum_{j \in \ln(V_i)} \frac{PR(V_j)}{\left|Out(V_j)\right|} \tag{2}$$

where:

d is a damping factor, and it is set to 0.85;
$PR(V_i)$ is the PR value of V_i;
$In(V_i)$ is the linking-in number of V_i;
$Out(V_j)$ is the out-link number of V_j.

After computing the PR value of vertices in networks, we can consider that the important vertices with high PR value have more influence than the vertices with low PR value in propagation. CoIN deliberates upon different needs of biocurators, so the system is flexible in customization.

Results and Discussion

We used training data which were released by BioCreative workshop 2012 a nd evaluated the ranking of training data separately by mean average precision (MAP), as shown in Table 1. The result shows that compared with co-occurrence frequency and normalized co-occurrence frequency, the vertices with high betweenness and PageRank value obviously occupy the important positions in a network. Therefore, the vertices with high betweenness and PageRank value are more likely to be important articles, and the

ranking indicates that network-based methods outperform co-occurrence methods. However, when we consider the tagging behavior of biocurator, co-occurrence methods provide more opportunities to curate terms intuitively. That is, co-occurrence methods are possible to have more undiscovered patterns than the others. On the other hand, the overestimated combinations in a sentence also produce noises easily. To avoid the bias, network-based methods are suitable to evaluate a more complex interaction network.

Table 1: Performance of different methods used in CoIN (MAP)

	Normalized co-occurrence frequency	Co-occurrence frequency	PageRank	Betweeness
2-acetylaminofluorence	0.496	0.511	0.601	0.602
Amsacrine	0.661	0.687	0.574	0.601
Aniline	0.676	0.657	0.884	0.880
Aspartame	0.465	0.443	0.434	0.479
Doxorubicin	0.692	0.724	0.675	0.672
Indomethacin	0.945	0.949	0.911	0.911
Quercetin	0.809	0.822	0.835	0.821
Raloxifene	0.698	0.703	0.634	0.634
Average	0.680	0.687	0.693	0.700

Acknowledgments The authors are grateful to Intelligence Knowledge Management laboratory for building the prototype system and James Hsu for helpful advice.

References

1. Neveol, A., Islamaj Dogan, R. and Lu, Z. (2011) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J Biomed Inform*, **44**, 310-318.
2. Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141-147.
3. Furney, S., Alba, M.M. and Lopez-Bigas, N. (2006) Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics*, **7**, 165.
4. Wei, C.H. and Kao, H.Y. (2011) Cross-species gene normalization by species inference. *BMC Bioinformatics*, **12 Suppl 8**, S5.
5. Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41-42.
6. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. and Gerstein, M. (2007) The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Comput Biol*, **3**, e59.
7. Freeman, L. (1977) A set of measures of centrality based upon betweenness. *Sociometry*, **40**, 35-41.
8. Brin, S. and Page, L. (1998), *Seventh International World-Wide Web Conference (WWW 1998)*, Brisbane, Australia.

DrTW: A Biomedical Term Weighting Method for Document Recommendation

Jiun-Huang Ju, Yu-De Chen, Jung-Hsien Chiang*

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

*Corresponding author: Tel: +886-6-275-7575 ext.62534, E-mail: jchiang@mail.ncku.edu.tw

Abstract
Extraction of the interactions of gene, chemical and disease is important to further understand the underlying biological processes of human. Although the literature repositories such as PubMed® provide such knowledge to researchers, the overwhelming volume of these digital records makes people hard to get up-to-date on biomedical discoveries. Hence, how to retrieve the relevant articles accurately and efficiently has been a pressing need for improvements in literature review. This work describes a biomedical term weighting method to automatically analyze and rank articles for document recommendation. Experiments evaluated by BC test system show that the proposed approach obtains an overall MAP (Mean Average Precise) of 0.785. DrTW is available at http://140.116.247.56/~biocreative/.

Introduction
The development of bioinformatics in recent years has been getting more attention for further understanding the underlying biological processes and led to the rapid growth of biomedical literature. The U.S. National Center for Biotechnology Information (NCBI, hosted by the U.S. National Library of Medicine) has indexed more than 21 million abstracts for biomedical literature in its freely-access literature repositories: PubMed. Consequently, the numerous biomedical discoveries are buried in the millions of these literatures. Although the repositories provide a literature search service to people, one could suffer from the time-consuming search. To cope with the daunting task, we have developed DrTW which employed a number of text mining techniques to perform a document recommendation using a biomedical term weighting method.

Material and Methods
The principle of the proposed approach includes 3 stages: 1. Sentence segmentation, 2. Named entities recognition and 3. Biomedical term weighting method. Moreover, we also provide users a friendly web-based interface which offers the information of abstract-text, gene, chemical, disease and relation. Fig. 1 illustrates the underlying algorithm in detail.

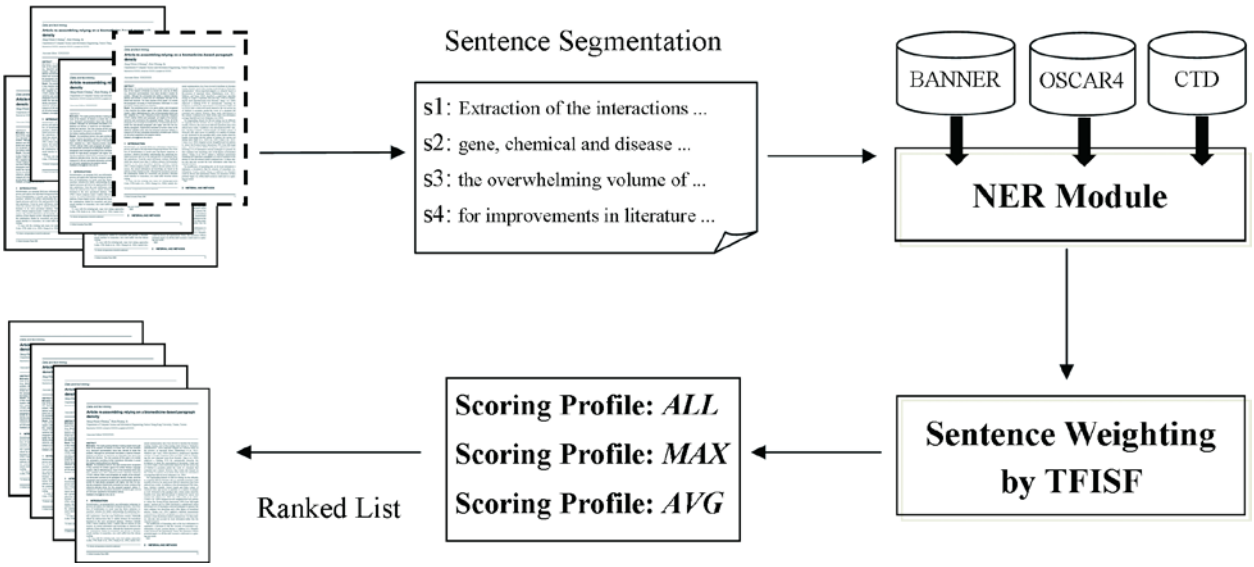


Figure 1. System workflow of DrTW.

Sentence Segmentation

The most important step of pre-processing is the sentence segmentation for each article. We used the NLP (Natural Language Processing) tool, LingPipe (1), to divide each article into several individual sentences. After sentence segmentation, the following stages will be performed to all sentences iteratively.

Named Entities Recognition

For CTD (Comparative Toxicogenomics Database) data curation, we focus on gene, chemical and disease entities recognition using a hybrid strategy. That is, there is a vast room for improvement in using only single NER (Named Entities Recognition) module. In order to enhance the performance of NER, we also take into account the CTD controlled vocabularies. Herein, all of the recognized terms are called "*keywords*".

- *Gene*
  We used a gene name recognizer, BANNER (2), and a lexicon, CTD controlled vocabularies (3) for genes, to identify gene names. BANNER is a state-of-the-art NER tool and a machine learning system based on CRF (conditional random fields) which shows comparable results. Since the CTD controlled vocabularies are curated manually, the performance of NER module could be increased thanks to the CTD genes added.

- *Chemical*
  Similarly, we utilized a chemical recognizer, OSCAR4 (4), and a lexicon, CTD controlled vocabularies (3) for chemicals, to identify chemical names. OSCAR4 employs a Maximum Entropy Markov Model to recognize the likely chemical entities.

- *Disease*
  Due to the lack of off-the-shelf disease entities recognizer, we can only use the CTD controlled vocabularies (3) for diseases to match the disease entities. Furthermore, disease

entities recognition achieved a low precise because of the delicate variations of disease terms.

Biomedical Term Weighting Method

In the field of information retrieval, term frequency-inversed document frequency (TF-IDF) is commonly used to estimate how important a term is in a corpus. However, we modified the TF-IDF to be applied to a sentence-level. The modified term-weighting method is described as follows.

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- $n_{i,j}$ expresses the frequency of term *i* in sentence *j*
- $\sum_k n_{k,j}$ represents the number of all of the terms in sentence *j*

$$\mathrm{ISF}_i = \log_2 \frac{|S|}{\left|\{j : t_i \in s_j\}\right|}$$

- $|S|$ denotes the number of sentences in an article
- $\left|\{j : t_i \in s_j\}\right|$ indicates the number of sentences containing term *i*

$$\mathrm{TFISF}_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{ISF}_i$$

- $\mathrm{TFISF}_{i,j}$ shows the modified term-weighting methods in this system. The higher $\mathrm{TFISF}_{i,j}$ a term gets, the more important the *keyword* means

For all of *keywords* in an article, the term weights were computed and summed according to the aforementioned TFISF to weight each of the sentences. We then designed 3 scoring profiles for ranking the articles as described below. Note that the profile *MIN* will not be discussed due to the unsatisfied result.

1. *ALL*: Summation of the scores of all sentences.
2. *MAX*: The score of the sentence with maximum weight.
3. *AVG*: The average of the scores of all sentences.

Results and Evaluation

To verify the usefulness of DrTW, we evaluated the system using BC'12 Track I online-testing website (https://bc.ctdbase.org/) on 1300 randomly selected articles obtained by BC'12 Track I training set that consists of 8 different target chemicals. The measurement used is MAP (5). Table 1 depicts that the performance of proposed system for different scoring profiles and target chemicals, and table 2 describes the system performance on BC'12 Track I testing set.

Table 1. Comparison of MAP for different scoring profiles and target chemicals.

Target	2-AAF	amsacrine	aniline	aspartame	doxorubicin	indomethacin	quercetin	raloxifene	ALL
Scoring Profile	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP
ALL	**0.731**	0.637	0.646	0.335	**0.799**	0.955	**0.855**	**0.732**	**0.784**
MAX	0.702	**0.645**	**0.661**	0.333	0.759	**0.975**	0.838	0.698	0.754
AVG	0.702	0.639	0.657	**0.344**	0.758	0.957	0.853	0.725	0.776

Table 2. System performance on BC'12 Track I testing set.

Target Chemical	The MAP using profile-ALL
urethane	0.605
phenacetin	0.834
cyclophosphamide	0.78

Interface Description

We constructed a web interface for demonstrating our results. The web site was built based on PHP and YUI framework. In the start page, user have to choose a target chemical and click the submit button (as shown in Fig. 2). Then, the user will be referred to the result page that exhibits the relevant articles in ordered rank (as displayed in Fig. 3). The *keywords* will be colored in red if the word represents a chemical name. Besides, detail information will be displayed when "[Show detail info]" button is clicked. The detail information we provided including the identified gene, chemical, disease and relation which are highlighted in different colors, respectively (as illustrated in Fig. 4). Moreover, user can click each of the *keywords* for linking out to the CTD database for more specialized information.

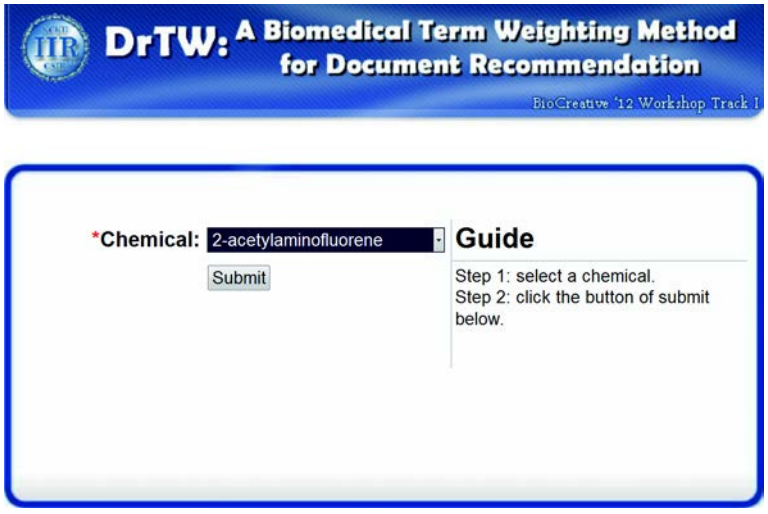


Figure 2. The web interface of DrTW.

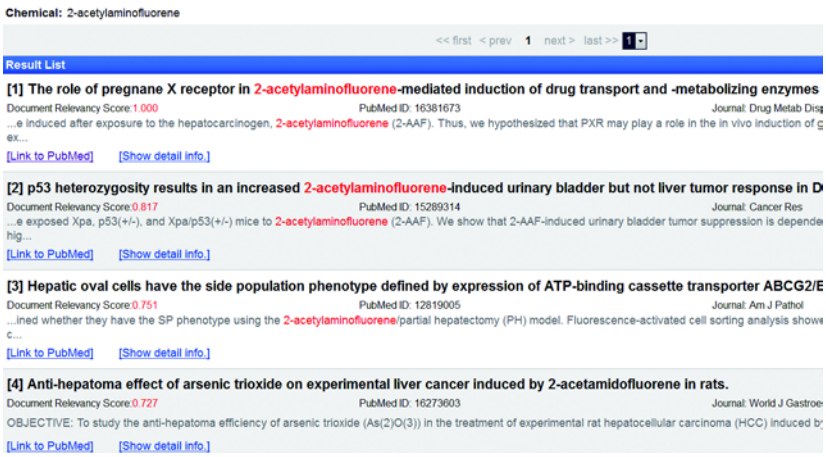


Figure 3. The result page of DrTW.

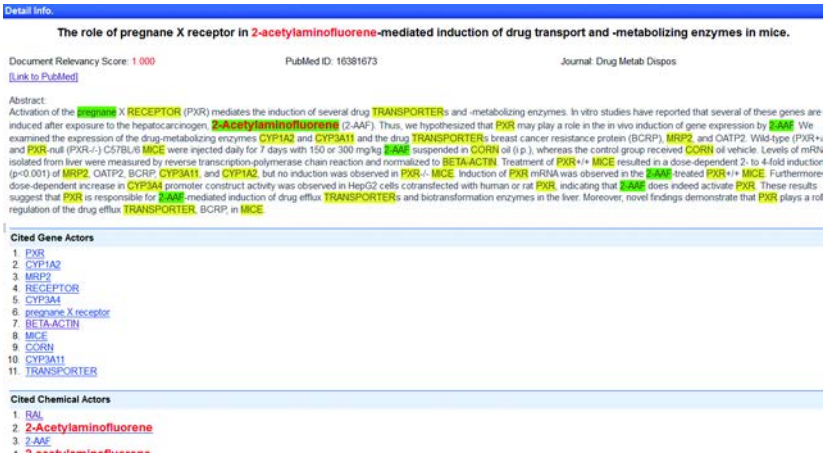


Figure 4. The detail information for result page of DrTW.

System Implementation

The core system of DrTW was implemented in Perl that computed the weight of articles and produced the ranking results for web interface. All programs were carried out on a Linux operating system with Intel Xeon 2.0 GHz processor (Quad-Core) and 8GB RAM. The curated data was stored in MySQL database management system, and the web service was based on an Apache server.

References

1. Alias-i. (2008) LingPipe 4.1.0 http://alias-i.com/lingpipe/ (Accessed February 19, 2012).

2. Leaman,R. and Gonzalez,G. (2008) BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition. *Pac Symp Biocomput*, **13**, 652-663.

3. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A. *et al*. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research*, **37**, 786-792.

4. Jessop,D.M., Adams,S.E., Willighagen,E.L. *et al*. (2011) OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, **3**, 41.

5. Voorhees,E.M., and Harman,D.K. (2005) TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA: MIT Press.

4. Jessop,D.M., Adams,S.E., Willighagen,E.L. *et al*. (2011) OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, **3**, 41.

5. Voorhees,E.M., and Harman,D.K. (2005) TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA: MIT Press.

C_2HI: a Complete CHemical Information decision system

Chao-Hsuan Ke[1], Tsung-Lu Michael Lee[2], Jung-Hsien Chiang[1*]

[1]Department of Computer Science and Information Engineering, National Cheng Kung University, [2]Department of Information Engineering, Kun Shan University, Tainan, Taiwan

*Corresponding author: Tel: +886-6-275-7575 ext.62534, E-mail: jchiang@mail.ncku.edu.tw

Abstract

C_2HI is a web service that integrates different information extraction techniques to extract and provide rankings for chemical-gene-disease associations among biomedical articles. The major objective of this system is to simplify the job of curating any association in between chemical, gene, and disease. We define those abstracts, which contain chemical-gene-disease associations, as curatable abstracts. This system supports great user interface and functions so that users can simply insert a chemical name and the C_2HI web service will return a comprehensive list of abstracts associated with the specific chemical. In addition, strong chemical-gene and chemical-disease relations are highlighted in the abstracts.

To evaluate the performance of the system, the experiment was performed by using a CTD training data set which consists of manually curated molecular interactions and relationships from 1,724 documents. The mean average precision (MAP) performance score achieves 0.787, while the highest MAP score is 0.922 for chemical *indomethacin* and the lowest score is 0.444 for *aspartame*.

Availability: C_2HI is freely available at http://140.116.247.40/C2HI/.

Material and Methods

The ranking process of C_2HI is consisted of two major components, the named entity recognition (NER) module and interaction ranking evaluation (IRE) module. In NER module, we adopted Comparative Toxicogenomics Database (CTD) (1) as dictionaries to recognize three types of entity names (chemical, gene and disease) and machine learning-based methods to identify complex chemical and gene names in articles.

The workflow of constructing C_2HI is shown in Figure 1. It includes two major modules: (A) named entity recognition module and (B) interaction ranking evaluation module. C_2HI also provides a user-friendly web interface for users to query and view the identified results. The detail of each module is described in the following subsections.

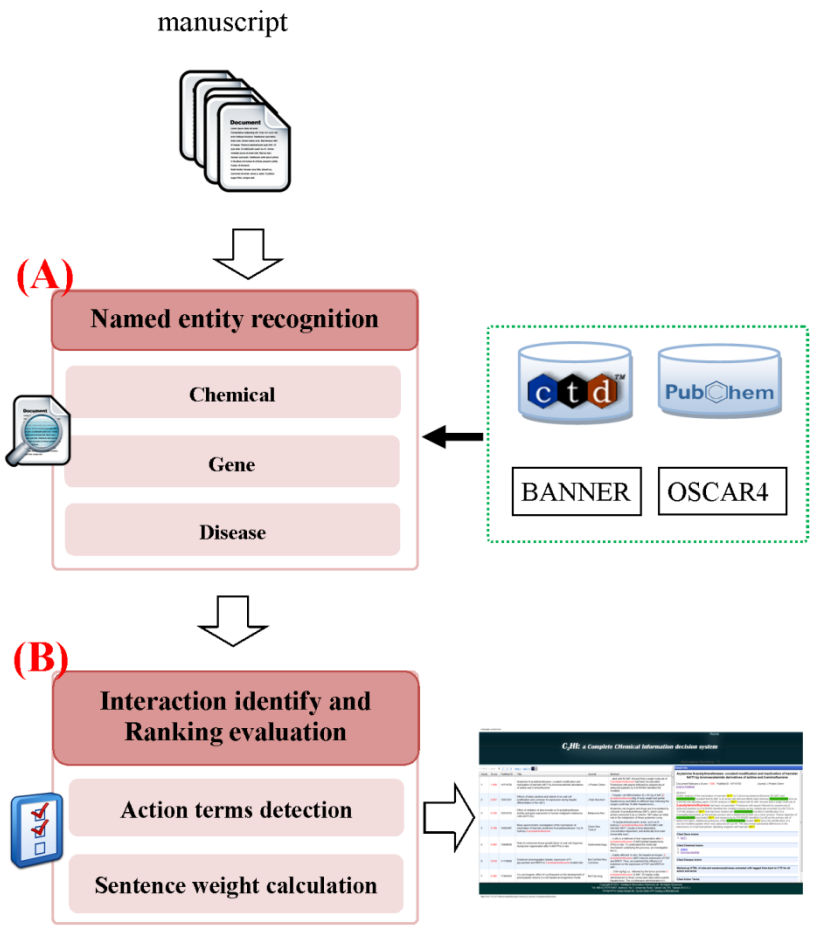


Figure 1. System modules of C_2HI

Named entity recognition

The system firstly splits full-text into sentences and marks the biological entity names in sentences based on the named entity recognition techniques. In order to enhance the performance in entities recognition, we used a combination measure that composing entities recognizer and a lexicon in chemical and gene names.

Chemical names recognition Similarly, we used a chemical recognizer, OSCAR4 (2), and a vocabulary, CTD Chemical database (1) (named *CTD Chemicals*), to recognize chemical entities. OSCAR4 employs a Maximum Entropy Markov Model (MEMM) to recognize chemical names; it is effective for separating chemical entities in chemistry related publications than using dictionary alone (3). Therefore, if a term is identified as the chemical name by OSCAR4 or identical to a chemical name/synonym in *CTD Chemicals*, it is marked as a chemical entity.

Gene names recognition We adopted BANNER (4) as gene name recognizer, and CTD Gene database (named *CTD Genes*) as lexicon, to recognize gene names. BANNER is based on a machine learning model which utilizes conditional random fields and a rich feature set widely surveyed from the literature; while *CTD Genes* is a manually constructed database of chemical-gene–disease interactions. Similar to chemical recognition process, if a term was identified as

gene entity by BANNER or matched to a gene name/synonym in the CTD Genes database, it is marked as a gene entity.

Disease names recognition In disease recognition, we directly adopted CTD disease data (named *CTD Diseases*) as vocabulary to recognize the disease names.

Interaction identify and ranking evaluation

To rank chemical-gene–disease interactions among evaluated articles, the main goal is to provide the most relevant score associated with chemical-gene and chemical-disease relations. We formulate the task of extracting chemical interaction as a ranking problem. For Track I task, we defined a new scoring function to evaluate each article. The method relies on di fferent frequencies of discriminating words (biological entities) between the curatable and non-relevant curatable articles. (5)

Here give the score to each article according to the description of entities. In article, firstly splits abstract into sentences and denoted $S=\{s_i\}$ where s_i is the i_{th} sentence. For each sentence, if there is gene was recognized in s_i, than $Score_g$ set as 5 point. In others, the chemical or disease was recognized in s_i, than $Score_c$ and Scored also set as 5 point respectively. Otherwise, if there is no any entities were marked in s_i, than the value of s_i is set as 0. F inally, $\text{Article}_{\text{score}}$ is the total number, that summarizing the score value for each sentence in abstract and divide by the number of sentences in abstract (m is the number of sentences in abstract). An abstract is scored by summing the weights of each of its sentence scores.

$$\text{Score}_g = \begin{cases} if\ gene\ was\ recognized\ in\ s_i, & 5 \\ else & , \ 0 \end{cases} \quad (1)$$

$$\text{Score}_c = \begin{cases} if\ chemical\ was\ recognized\ in\ s_i, & 5 \\ else & , \ 0 \end{cases} \quad (2)$$

$$\text{Score}_d = \begin{cases} if\ disease\ was\ recognized\ in\ s_i, & 5 \\ else & , \ 0 \end{cases} \quad (3)$$

$$s_i = \text{Score}_g + \text{Score}_c + \text{Score}_d \quad (4)$$

$$\text{Article}_{score} = \frac{s_i}{\sum_{i=1}^{m} No.of\ sentences\ in\ abstract} \quad (5)$$

Evaluation and Result

To evaluate the system, we performed a well-known method called mean average precision (MAP) (6) in the field of information retrieval. In our evaluation procedure, we upload different target chemical from training corpus to the system, and the results are illustrated in Table 1. In addition, the performances of testing data (*urethane*, *phenacetin* and *cyclophosphamide*) are illustrated in Table 2.

Table 1. The target chemical results in Training data

Target Chemical	Intermediate MAP Score	Curated Gene Text Mined Hit Rate	Curated Chemical Text Mined Hit Rate	Curated Disease Text Mined Hit Rate	Curated Action Term Text Mined Hit Rate
2-acetylaminofluorene	0.676	0.124	0.338	0.000	0.000
amsacrine	0.673	0.194	0.308	0.000	0.000
aniline	0.620	0.236	0.159	0.000	0.000
aspartame	0.444	0.314	0.430	0.000	0.000
doxorubicin	0.779	0.058	0.634	0.000	0.000
indomethacin	0.922	0.112	0.542	0.000	0.000
quercetin	0.847	0.151	0.337	0.000	0.000
raloxifene	0.737	0.085	0.372	0.000	0.000

Table 2. The target chemical results in Testing data

Target Chemical	Intermediate MAP Score	Curated Gene Text Mined Hit Rate	Curated Chemical Text Mined Hit Rate	Curated Disease Text Mined Hit Rate	Curated Action Term Text Mined Hit Rate
urethane	0.667	0.105	0.452	0.000	0.000
phenacetin	0.875	0.306	0.231	0.000	0.000
cyclophosphamide	0.722	0.151	0.687	0.000	0.000

Interface description

In C_2HI, it provides eight given chemical entities, which are available in the selection menu.

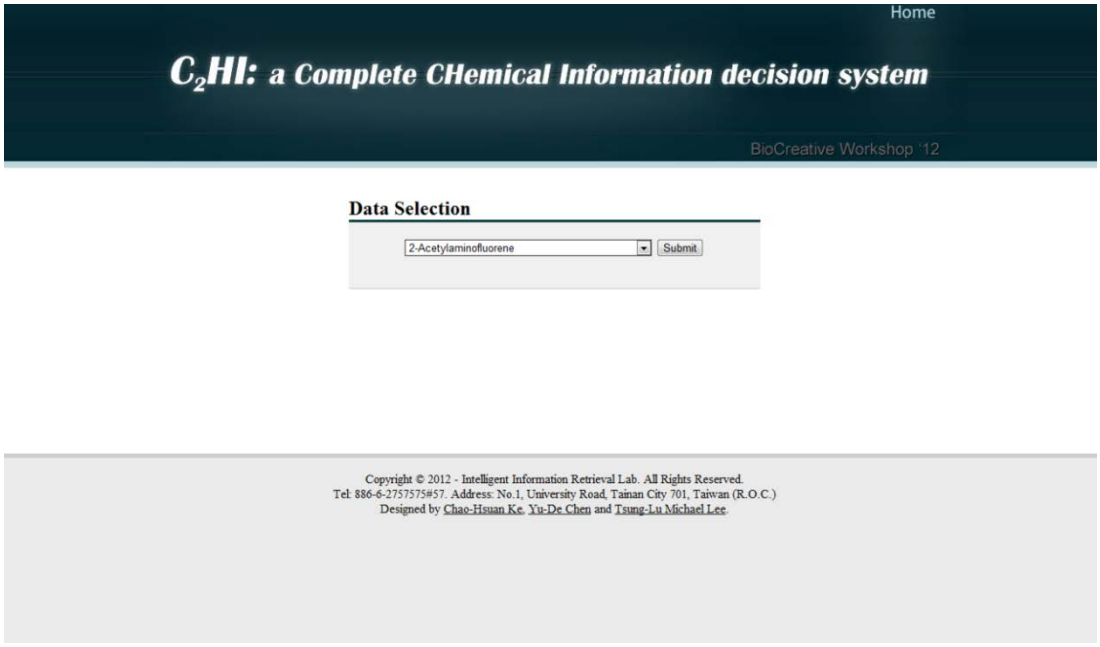


Figure 2. The web interface of C_2HI

The C_2HI interface is designed to provide user-friendly interface to users. In Figure 3, we demonstrated an example of a query result. It listed all the articles which are focused on '*2-Acetylaminofluorene*'. C_2HI displays the extracted results with PubMed ID and provides a link to check detailed information for every highlighted name entity, and show a weight score for each article. In addition, C_2HI also provides a convenient retrieval of each entry by chemical, gene and disease name; it is effective help user to find the target articles.

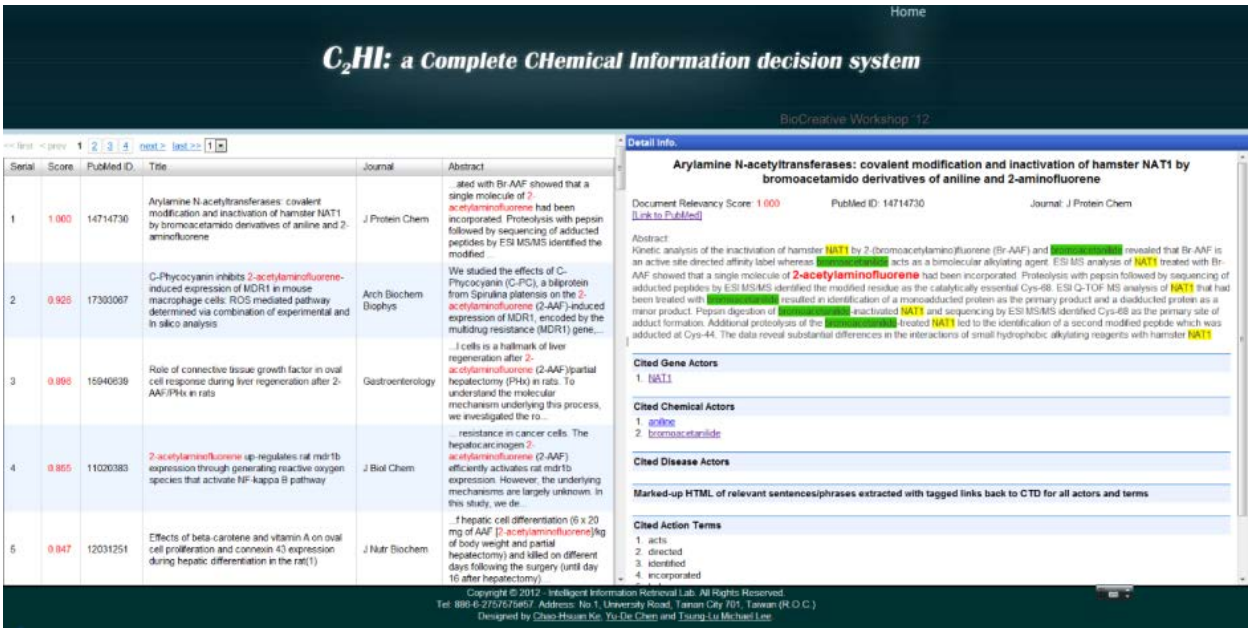


Figure 3. Example output of ranked articles for target chemical '*2-Acetylaminofluorene*'

System implementation

The underlying code of the application and web pages were programmed using Java 6, PHP 5.2.6 respectively, and tested on a n Apache/2.2.11 server with a MySQL database (5.0.75 version). All the system is running in a Linux operating system with two Intel Xeon 2.5 GHz processor (Quad-Core) and 2048MB RAM. In evaluating system performance, we tested 1724 training data sets, system spent 1 hour 42 minutes for NER process. However, it took only 14 seconds for calculating article weights.

Funding

This work was supported by National Science Council, Taiwan (Grants NSC99-2221-E-006-127-MY3, NSC100-2627-B-006-011).

References

1. Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., Mattingly, C. J. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Research*, ***37***, D786-D792.

2. Jessop, D., Adams, S., Willighagen, E., Hawizy, L., Murray-Rust, P. (2011) OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, ***3***, 41.

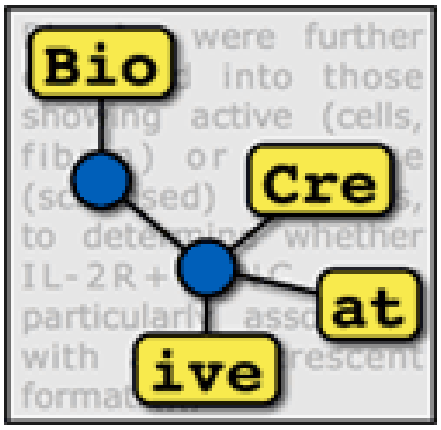
3. Hettne, K., Williams, A., van Mulligen, E., Kleinjans, J., Tkachenko, V., Kors, J. (2010) Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *Journal of Cheminformatics*, ***2***, 3.

4. Leaman, R., Gonzalez, G., in *Book BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition*, ed., ed. by Editor, City, **2008**, Chap. Chapter, pp. 652-663.

5. Suomela, B., Andrade, M. (2005) Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics*, ***6***, 75.

6. Voorhees, E., Harman, D. (2005) TREC: Experiment and Evaluation in Information Retrieval. Cambridge. *MA: MIT Press*, 21-52.

Track 2



Overview of BioCreative Curation Workshop Track II: Curation Workflows

Zhiyong Lu[1] and Lynette Hirschman[2,*]

[1]National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, Bethesda, MD, 20817
[2]The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730
*Corresponding author: Tel: 617-893-9608 Email: lynette@mitre.org

Abstract

Manually curating data from the biomedical literature is a rate-limiting factor for many expert curated databases. Despite the continuing advances in biomedical text mining and its promise in assisting manual curation, few existing text mining tools have been successfully integrated into production literature curation systems such as those used by the expert curated databases. To close this gap and better understand all aspects of literature curation, we invited submissions of written descriptions of curation workflows from expert curated databases in the BioCreative 2012 workshop Track II. In response to our call, we received seven submissions from expert curated databases. We reviewed these to identify commonalities across these workflows as well as some areas of contrast; we also searched for common ontologies and controlled vocabularies used across databases and we summarized the current and desired uses of text mining for biocuration. Compared to a similar survey in 2009, our results show that in 2012 significantly more databases are already using text mining in some parts of their curation workflows. In addition, our study finds that text-mining aids for gene indexing, document triage, and ontology term annotation are the features of greatest interest to the expert curators.

Introduction

BioCreative 2012 Workshop Track II on "Curation Workflows" is an outgrowth of experiences at an earlier workshop on "Text Mining for the Biocuration Workflow" held at the 3rd International Biocuration Conference, April 2009. In preparation for that workshop, the workshop organizers interviewed curators and elicited workflows for eight expert curated biological databases (1), with the goal of better understanding where text mining might be most usefully inserted into the curation workflow. This turned out to be a useful activity: it enabled the curators to make explicit their procedures, and it provided the text mining developers with an overall context and possible insertion points for text mining modules. This activity was useful for another reason: it turned out that – at a detailed level – the workflows differed quite a bit, even among model organism

databases. There were differences in the scale and complexity of the curation activities, the volume of literature to be curated, the sources of the literature to be curated, the prioritization process for curation, the resources available for curation, and the types of entities curated.

The positive feedback we received from curators from the 2009 workshop led us to propose a track for the BioCreative 2012 Curator Workshop devoted explicitly to collecting workflows from multiple biological databases (BioCreative 2012 Curator Workshop Track II, hereafter Track II).

Our work is also related to several prior studies on the integration of text mining tools into the biocuration workflow. In (2,3), the authors described text-mining applications for assisting the curation of the Mouse Genome Informatics (MGI) resource and the Comparative Toxicogenomics Database (CTD), respectively. More recently, Krallinger and colleagues (4) reported an overview of current text-mining methods for linking ontologies and protein-protein interactions to the biomedical literature, from their BioCreative experience (5,6).

Methods

The Track II call for papers asked curation teams to produce a document describing their curation process starting from selection of articles for curation (as journal articles or abstracts) and culminating in database entries.

As part of the track materials, we provided an outline identifying issues that would be useful to text mining developers who are seeking to produce algorithms and tools to assist the curation process (shown in Table 1).

Outline of the Description of a Curation Workflow	
Introduction	a. Overall philosophy: what information is captured and from what sources? b. What use is being made of this information or is envisioned for this information? c. What is the current workflow of the operation, and where are automated methods used?
Encoding methods	a. How is the information captured to make it machine readable? b. What entities are involved and how are they entered in the database? c. What relationships are involved and how are they symbolized? d. What standardized or controlled vocabularies are used?

	e. Give examples of a variety of data elements and how they appear in the database.
Information access	a. When a curator runs into a problem or a difficult case what kind of information is needed to solve it? b. What kind of internet searching is used most often in difficult cases? Dictionary? Wikipedia? Other database?
Use of text mining tools	a. What text mining tools do you currently employ in your workflow and what problems do these algorithms solve for you? b. What problems do you have that are not currently solved, but which you think could be amenable to a text mining solution (i.e., for which steps text mining could overcome current bottlenecks in the existing pipeline)?

`

Table 1: Outline of issues in describing the curation workflow

We received eight submissions to this track, of which seven described workflows of existing expert curated databases, including

- AgBase (agricultural plants and animals)
- FlyBase (fruit fly)
- MaizeGDB (maize)
- MGI (mouse)
- TAIR (Arabidopsis)
- WormBase (C. elegans)
- Xenbase (frog)

Based on these submissions, we identified commonalities across the workflows as well as some areas of contrast. Table 2 below lists three basic stages of processing that were common across curated databases, as well as some sub-stages (1).

Curation Stage	Sub-stage	Description
Sources	0	Sources of papers to be curated
Paper Selection	1	Triage to prioritize articles for curation
	2	Indexing of biological entities of interest
Full Curation	3	Curation of relations, experimental evidence
	4	Extraction of evidence within document (sentences, images)
	5	Check of record

Table 2: Stages in the curation workflow

Results

The workflows showed commonalities across the three stages identified in Table 2, as well as differences. Table 3 summarizes some of these comparisons.

Curation Stage	Commonalities	Differences
Source	- PubMed search (abstracts) - Full-text articles (pdf)	- Number of papers to be curated - Acceptance of sources outside of PubMed (e.g. author submission)
Paper selection (Triage)	- Manual process by humans - Primarily based on abstract - Assignment of curation priorities - Identification of genes/proteins	- DB specific selection criteria (e.g. species, gene/function, novelty) - Identification of DB-specific curation requirements, e.g., experiment evidence
Full curation	- Gene (function) centric - Use of full text - Use of controlled vocabularies and ontologies - Identification of experimental evidence - Contacting authors when needed	- Annotating DB specific bio-entities, e.g. anatomy, cell type, - Annotating database/species specific entities and relationships - Annotating images (Xenbase)

Table 3: Commonalities and differences in the curation workflow stages

All of the databases encoded a variety of biological entities using standard vocabularies and ontologies. Table 4 identifies (a subset of) common types of biological entities curated in the various databases. In particular, all of the databases used the Gene Ontology (7) to encode information about genes. In several cases, the workflow submitted to Track II described only a specific slice of a larger curation process, so that the full curation process for some of these databases (MGI in particular) may be considerably broader than what is captured in Table 4.

Ontologies	AgBase	TAIR	MGI	Xenbase	MaizeGDB	FlyBase	WormBase
Gene (7)	X	X	X	X	X	X	X
Plant (8)	X	X			X		
Sequence (9)			X			X	X

Table 4: Common ontologies used across multiple curation databases.

Finally, the Track II call for papers asked the database curators to identify where they used text mining/natural language processing in their current workflow, and where they would like to see it used. All of databases were already using text mining, and six of the seven databases were using Textpresso (10) to search for specific classes of entities and/or to pre-annotate certain classes of concepts (11). Some of the current and future/desired uses are summarized in Table 5. There was strong interest in having enhanced text mining capabilities to recognize and assign ontology terms, particularly the three branches of GO, including extension to gene function and biological process, which are both quite challenging. (Textpresso has a capability to assign GO cellular component terms, which was being used in a number of databases). There was also strong interest in better use of text mining to identify and prioritize documents for curation (the triage process).

	Specific use cases of text mining tools
Current	• Finding gene names and symbols (gene indexing) • Querying full text with Textpresso • Annotating GO cellular component terms
Future/Desired	• Improving gene indexing results • Performing document triage • Recognizing other biological concepts especially ontological terms (e.g. GO) • Capturing complex relations such as gene regulation

Table 5: Current uses of text mining and desired uses

Discussion and Conclusions

One striking change from the previous survey in 2009 is that today, all seven systems are already using text mining in at least some parts of their workflow. This may, in part, be due to the heavy representation of model organism databases among the submissions to this track. There is a sophisticated suite of open source software tools available for use by model organism databases through GMOD (http://gmod.org). In addition, Textpresso is being increasingly widely used, particularly in this community. Moreover, Textpresso's capabilities are being extended, in response to the needs of the MODs (model organism databases). The workshop will offer an excellent opportunity to understand why the Textpresso model is succeeding, and what lessons can be learned from its success.

It is encouraging to see the wider uptake of text mining, particularly in the MOD community. However, several nagging questions remain: Are these tools good enough to enable curators to keep up with the flood of data? How much do they help? Are these the right tools and the right insertion points to ease the "curation bottleneck"?

Using these workflow descriptions, can we now begin to quantify where curator time is spent? And can we begin to build some more sophisticated models of the costs and benefits of bringing tools into the workflow? Can we start to build realistic models of time spent on development/adaptation of tools to a specific database, as well as time spent training curators to use the tools?

In conclusion, we have analyzed and reviewed curation workflow descriptions from seven independent curation groups, based on which we have identified both unique and common aspects of literature curation among groups. Moreover we have identified several possible insertion points for text mining to simplify manual curation. At the BioCreative IV workshop in 2013, we will (begin to) address some of the remaining questions by working in close partnership between the biological database curators and the text mining tool developers.

Acknowledgements

We would like to thank other BioCreative 2012 organizers for helpful discussion and the track II teams for their participation. This research is funded by the US National Science Foundation grant DBI-0850319 (LH) and the N.I.H. Intramural Research Program, National Library of Medicine (ZL).

References

1. Hirschman, L., Burns, G.A.P.C., Krallinger, M.*, et al.* (2012) Text Mining for the BioCuration Workflow. *Database (Oxford)*, accepted.
2. 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.
3. Wiegers, T.C., Davis, A.P., Cohen, K.B., Hirschman, L., Mattingly, C.J. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*, **10**, 326.
4. Krallinger, M., Leitner, F., Vazquez, M.*, et al.* (2012) How to link ontologies and protein-protein interactions to literature: text-mining approaches and the BioCreative experience. *Database (Oxford)*, **2012**, bas017.
5. Krallinger, M., Vazquez, M., Leitner, F.*, et al.* (2011) The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, **12 Suppl 8**, S3.
6. Arighi, C.N., Lu, Z., Krallinger, M.*, et al.* (2011) Overview of the BioCreative III Workshop. *BMC Bioinformatics*, **12 Suppl 8**, S1.

7. Ashburner, M., Ball, C.A., Blake, J.A.*, et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-9.
8. Jaiswal, P., Avraham, S., Ilic, K.*, et al.* (2005) Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comparative and functional genomics*, **6**, 388-97.
9. Eilbeck, K., Lewis, S.E., Mungall, C.J.*, et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, **6**, R44.
10. Muller, H.M., Kenny, E.E., Sternberg, P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, **2**, e309.
11. Van Auken, K., Jaffery, J., Chan, J., Muller, H.M., Sternberg, P.W. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, **10**, 228.

WormBase Literature Curation Workflow

Kimberly Van Auken[1*], Bieri T[2], Cabunoc A[3], Chan J[1], Chen WJ[1], Davis P[4], Duong A[3], Fang R[1], Grove C[1], Harris TW[3], Howe K[4], Kishore R[1], Lee R[1], Li Y[1], Muller HM[1], Nakamura C[1], Nash B[2], Ozersky P[2], Paulini M[4], Raciti D[1], Rangarajan A[1], Schindelman G[1], Tuli MA[4], Wang D[1], Wang X[1], Williams G[4], Yook K[1], Hodgkin J[5], Berriman M[6], Durbin R[6], Kersey P[4], Spieth J[2], Stein L[3], Sternberg PW[1, 7].

[1]Division of Biology, California Institute of Technology, Pasadena, CA, [2]The Genome Institute, Washington University School of Medicine, St Louis, MO, [3]Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, ON, Canada, [4]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [5] Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford, UK, [6] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [7]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA

*Corresponding author: Tel: (609) 937-1635, E-mail: vanauken@caltech.edu

Abstract
WormBase (http://www.wormbase.org) is a model organism database containing data about *C. elegans* and other nematodes. The WormBase literature curation workflow typically begins by downloading bibliographic information from PubMed followed by subsequent steps of pdf acquisition, data type flagging, entity recognition, and fact extraction. Using a combination of manual, semi-, and fully automated approaches including community curation, Perl scripts, Support Vector Machines (SVMs), and the Textpresso (http://textpresso.org/) information retrieval system, WormBase curators annotate over 30 different data types to support *C. elegans*-based biomedical research.

Introduction
Overview
WormBase (http://www.wormbase.org) is a model organism database that curates data about *C. elegans* and other nematodes (1). Although WormBase is largely a gene-centric database, it warehouses, in addition to genomic sequence and gene function curation, information about additional aspects of nematode biology, including anatomy, reagents, researchers, and publications. Currently, WormBase releases a new version of the database six times a year.

Use of Information
The WormBase user community is composed of nematode biologists, as well as scientists from a variety of other research communities. Information in WormBase is intended to assist both day-to-day laboratory experimentation as well as computationally intensive data mining. Curated data is accessible to users via individual web pages, data sets available for download from an ftp site, and query interfaces such as WormMart (for further details, see 1). WormBase also exports annotations to other databases and projects, such as the Gene Ontology (GO) Consortium (2).

Current Workflow

The WormBase literature curation workflow is diagrammed below. Papers relevant to WormBase curation are processed through automated and manual triage (data type flagging) methods. Text mining and manual curation are used for fact extraction.

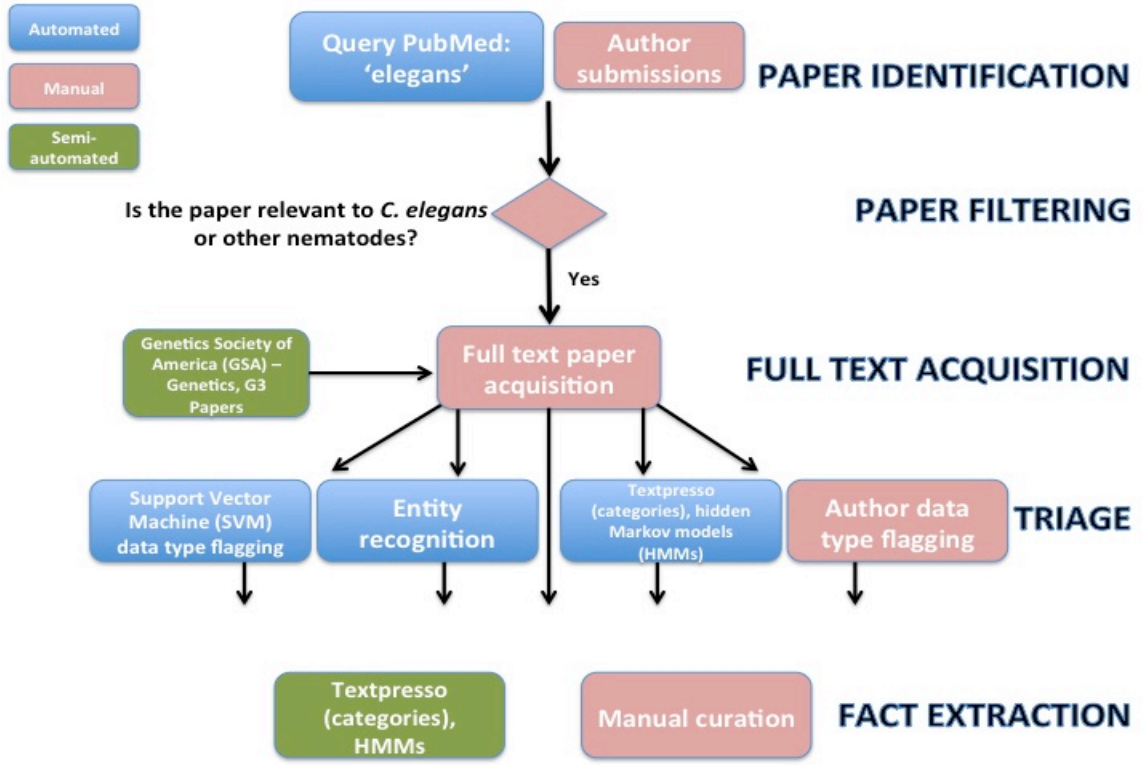


Figure 1. The current WormBase literature curation workflow.

Workflow: Paper Identification to Full Text Acquisition

Paper Identification: PubMed Searches

Papers curated for WormBase typically enter the curation pipeline via automated, daily PubMed searches using the keyword 'elegans'. Bibliographic information from papers thus identified is presented to a curator for manual approval via a web-based form. In most cases, papers can be approved or rejected based upon the content of the abstract but in some cases, curators access the full text of the paper before making a final decision. At this stage, approved papers are designated 'primary' or 'not primary' to indicate whether the paper is likely to contain primary experimental data.

Paper Identification: Author Submission

WormBase also receives non-PubMed-indexed articles directly from authors. In these cases, a curator assesses the relevance of the paper to WormBase before manually entering bibliographic information.

Paper Acquisition: Genetics Society of America (GSA)

In collaboration with the GSA and Dartmouth Journal Services, WormBase receives pre-publication access to *C. elegans* articles accepted by the journals Genetics and G3. The full text of these articles, and their corresponding digital object identifiers (DOIs), are submitted to

WormBase from the GSA via a web service. Full bibliographic information is subsequently retrieved via the PubMed pipeline. Genetics and G3 articles are curated, in part, via a text markup pipeline that adds hyperlinks to entities within the paper to WormBase web pages, a process that can also serve as a curation flagging mechanism (3).

Full Text Acquisition

We strive to obtain the full text of every paper included in WormBase. PDF acquisition is performed manually and the full text of all papers is archived in our curation database.

Workflow: Triage: Entity Recognition and Data Type Flagging

Entity Recognition

Entity recognition, such as for genes, variations, transgenes, antibodies, molecules, and human diseases, is an essential aspect of WormBase curation. In cases where the entity in question conforms to standard *C. elegans* nomenclature, such as genes, variations and transgenes, pattern matching using regular expressions is used for identification. In other cases, e.g. molecules, lists of entities mined from databases such as ChEBI (4), are used for curation.

Data Type Flagging: Support Vector Machines (SVMs)

SVMs are currently used at WormBase to flag papers for 10 different data types (for details on the methodology, see 5). SVM analyses are performed on the full text of articles and results presented to curators on a web page that lists predicted positive and negative papers. Positive papers are further classified as being of high, medium, or low confidence.

Data Type Flagging: Textpresso Category Searches and Hidden Markov Models (HMMs)

The Textpresso information retrieval system (6) is also used to identify papers for curation. Using both manual and automated pipelines, keyword and/or category searches on the full text of papers are used to identify documents containing data types of interest. Curators use existing Textpresso categories or may design new, curation task-specific categories. Searches can be restricted to specific paper sections, maximizing the likelihood that search results are most relevant to data curation. In addition, as described in more detail below, HMMs are also used to identify papers for annotating to the GO's Molecular Function (MF) ontology.

Data Type Flagging: Manual Flagging

Curators and authors are able to flag papers for specific data types using a web-based form. Corresponding authors are contacted via e-mail shortly after their publication is incorporated into the WormBase curation database. The email message contains a link to a form where authors can flag the data types in their paper and, optionally, provide experimental details.

Workflow: Fact Extraction

Overview

Curators extract experimental details either via fully manual curation or by semi-automated methods using the results of Perl scripts, Textpresso category searches, or HMMs. Annotations are made from the full text of articles (not from abstracts), and consist largely of creating and characterizing entities (e.g., sequence change in a variation) and then linking two or more entities in a biological relationship, often using controlled vocabularies (e.g., gene A regulates gene B with respect to process P in cell type C). Free text descriptions are also used to create concise gene function descriptions or capture some experimental details.

Manual Curation

In many cases, curators manually enter data into a curation database using web-based forms (see below). With over 30 different data types curated, manual annotation covers a very broad

range of experimental results and uses a number of different controlled vocabularies and ontologies. A table describing the different data types and methods used for their curation is available at: http://wiki.wormbase.org/index.php/WormBase_Literature_Curation_Workflow

Semi-automated Curation: Textpresso Category Searches and HMMs

In addition to manual curation, results of Textpresso category searches are also used for fact extraction, specifically genetic and physical interactions and GO Cellular Component Curation (CCC) (7). In some cases, the Textpresso searches are performed on a subset of SVM-positive papers to prioritize high confidence papers for curation. In the case of genetic interactions and GO CCC, Textpresso sentences are presented in the context of a curation form. For GO CCC, the curation entity and, where possible, suggested annotations based on previous curation, are pre-populated on the form. HMMs are currently used to curate enzymatic and transporter activities for GO's MF ontology. Here, curators are presented with sentences ranked according to the probability that they contain relevant information. In some cases, sentences identified by the HMM are not sufficient to make an annotation, but indicate that an experiment has been performed. In those cases, the curator consults the full text to determine what annotation can be made.

Workflow: Links to Biomedical Literature

WormBase data models contain tags that link curated data directly to the reference used for curation. Reference information is displayed on the website including links out to PubMed.

Encoding Methods

Making Information Machine Readable

Much of the information gathered from literature curation is captured in web-based curation forms. One form, the Ontology Annotator (OA), is used to curate 12 different data types. The OA is based on the Phenote (http://phenote.org/) curation tool developed by the Berkeley Bioinformatics Open-Source Projects (http://berkeleybop.org/). Information entered via the OA is stored in a PostgreSQL database maintained in a Linux operating system environment. Key features of the OA include easily annotating using ontologies, the ability to view information about other curated objects, auto-completion of ontology searches, and drop-down menus for short lists. Prior to each build of the WormBase database, data is exported in file formats conforming to the ACeDB database underlying part of the WormBase web site.

Entities, Relationships and their Representation in the Database

Data in WormBase is represented as distinct classes, e.g., genes, RNAi experiments, and molecules, that store information about specific instances of that class. Each class has a corresponding data model that organizes information, denoted by tag names, relevant to representing that class in the database. Tag names may reflect a relatively straightforward concept such as 'NCBI Taxonomy ID' or a more complex relationship between two entities such as 'Affects phenotype of'. Tag values may be a database identifier from WormBase or another database, integers, or free text. Evidence for the value contained within a tag, such as a reference, may be associated directly with the tag value or may instead be included as part of the database object used to populate the tag.

Standardized and Controlled Vocabularies

WormBase curators use a number of ontologies, including the Gene Ontology (2), Sequence Ontology (8), Phenotype Ontology (9), Cell and Anatomy Ontology (10), and Life Stage Ontology (10), for data type curation. Internal controlled vocabularies are also used for data

curation to capture, for example, details about antibody production, such as the organism in which an antibody was produced and whether the antibody is tissue-specific.

Information Access

Curation Difficulties and their Resolution

Curation difficulties arise in two general ways. First, information presented in a publication may not be sufficient to link the data unambiguously to a WormBase database object requiring curators to contact authors for additional information. If no further information is available, data may be assigned to the most general entity possible or may not be captured in WormBase. Second, information presented in a paper may not be sufficient to confidently assign annotations without additional background knowledge, requiring curators to consult previous publications or WormBook (11) or online resources such as GO or Wikipedia.

Use of Text Mining Tools

Overview: Text Mining Tools Currently in Use

WormBase uses several different text mining tools, ranging from Perl scripts for pattern matching, to SVMs (5), Textpresso searches (6, 7) and HMMs, to aid curation. A brief summary of each method and how it is used is presented below.

Scripts for Entity Recognition and Automated Downloads

Perl scripts are used for a variety of tasks, including pattern matching for entity identification, automation of downloads from external sites, such as PubMed, conversion of pdfs to plain ASCII text for Textpresso mark up and indexing, and image extraction from journal articles.

SVMs for Data Type Flagging

SVMs are used for data type flagging. Currently, SVMs flag papers containing the following data types: antibodies, genetic interactions, physical interactions, gene regulation, RNAi experiments, variation phenotypes, overexpression phenotypes, expression patterns, variation-associated sequence changes, and gene model corrections.

Textpresso Category Searches for Triage and Fact Extraction

Textpresso category searches on the full text of papers are used to identify papers for curation (triage) as well as to identify sentences for fact extraction. In some cases, Textpresso searches are performed as part of a two-step process in which papers classified by SVM as positive for a given data type are then subject to Textpresso searches. Also, predicted SVM negative papers can be searched using Textpresso categories to help identify potential false negatives (e.g. SVM negative, but high-scoring Textpresso papers).

HMMs for Fact Extraction

We recently developed an HMM for identifying sentences describing enzymatic and transporter activities. After several rounds of training, the HMM was applied to the entire *C. elegans* corpus and the results are being used to both curate to the GO's MF ontology and assess annotation metrics.

Further Development of Text Mining Tools

Although text mining applications have helped tremendously with our workflow, additional data types might benefit from text mining and further improvements can be made in the precision and recall of existing methods. For example, although we currently use Textpresso and HMMs for GO CC and MF annotation, we have not yet employed text mining for GO Biological Process (BP) annotation. Text mining might also help yield more sophisticated

triage approaches that consider existing curated information before flagging a paper. Such an approach would allow curators to prioritize curation of truly novel experimental results.

Funding

This work was supported by the US National Human Genome Research Institute [Grant no. HG02223 and Grant no. HG004090] and the British Medical Research Council [Grant no. G070119]; PWS is an investigator with the Howard Hughes Medical Institute.

Acknowledgements

K Van Auken gratefully acknowledges helpful comments on the manuscript from C Grove, K Howe, HM Muller, MA Tuli, G Williams, and K Yook.

References

1. Yook K, Harris T, et al. (2012) WormBase 2012: more genomes, more data, new website. Nucleic Acids Res. 40(Database issue):D735-41.
2. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. Nucleic Acids Res. 40(Database issue):D559-64.
3. Rangarajan A, Schedl T, Yook K, Chan J, Haenel S, Otis L, Faelten S, De Pellegrin-Connelly T, Isaacson R, Skrzypek MS, Marygold S, Stefancsik R, Cherry JM, Sternberg PW, Muller HM. (2011) Toward an interactive-article: integrating journals and biological databases. BMC Bioinformatics. 12:175.
4. de Matos P, Adam N, Hastings J, Moreno P, Steinbeck C. (2012) A database for chemical proteomics: ChEBI. Methods Mol Biol. 803:273-96.
5. Fang R, Schindelman G, Van Auken K, Fernandes J, Chen W, Wang X, Davis P, Tuli MA, Marygold S, Millburn G, Matthews B, Zhang H, Brown N, Gelbart WM, Sternberg PW. (2012) Automatic categorization of diverse experimental information in the bioscience literature. BMC Bioinformatics. 13(1):16
6. Müller HM, Kenny EE, Sternberg PW. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2(11):e309.
7. Van Auken KM, Jaffery J, Chan J, Müller HM, Sternberg PW. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) cellular component curation. BMC Bioinformatics. 10:228.
8. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6(5):R44.
9. Schindelman G, Fernandes JS, Bastiani CA, Yook K, Sternberg PW. (2011) Worm Phenotype Ontology: integrating phenotype data within and beyond the *C. elegans* community. BMC Bioinformatics. 12:32.
10. Lee RYN and Sternberg PW. (2003) Building a cell and anatomy ontology of *Caenorhabditis elegans*. Comp Funct Genomics. 4:121-6.
11. Girard LR, Fiedler TJ, Harris TW, Carvalho F, Antoshechkin I, Han M, Sternberg PW, Stein LD, Chalfie M. (2007) WormBook: the online review of *Caenorhabditis elegans* biology. Nucleic Acids Res. 35(Database issue):D472-5.

Literature curation workflow at The Arabidopsis Information Resource (TAIR)

Donghui Li*, Robert Muller, Tanya Z. Berardini, Eva Huala

Department of Plant Biology, Carnegie Institution for Science, Stanford, CA

*Corresponding author: Tel: 650 325 1521, E-mail: donghui@stanford.edu

Abstract
TAIR (The Arabidopsis Information Resource) is a model organism database for *Arabidopsis thaliana*. Here we describe our manual literature curation workflow. The current workflow can be divided into two phases: paper triage and curation. Structured controlled vocabularies such as Gene Ontology and Plant Ontology are used to convert free text information in the literature into a machine-readable format. We also describe our curation platform and the use of text mining tools in our workflow.

Introduction
TAIR (The Arabidopsis Information Resource, http://www.arabidopsis.org) is the primary database for *Arabidopsis thaliana*, a model organism for plant biology research (1, 2). TAIR strives to be a centralized, curated gateway to *Arabidopsis* biology, research materials and community members. Data available from TAIR includes the complete *Arabidopsis* genome sequence along with gene structure, gene function information, metabolism, gene expression, genome maps, genetic and physical markers, publications, and information about the *Arabidopsis* research community. In addition, seed and DNA stocks information and ordering for the Arabidopsis Biological Resource Center (ABRC) are fully integrated into TAIR.

TAIR is a curated database; data are processed by Ph.D.-level biocurators who ensure their accuracy. TAIR data come from a variety of sources including manual curation of published literature and sequence data, computational pipelines for annotating gene structure and function, integration of data from other biological databases and resources (GenBank, ABRC, Gene Ontology Consortium etc.) and submissions from the research community. TAIR also provides researchers with an extensive set of data visualization and analysis tools.

This article describes our workflow for manual literature curation only; workflows for other aspects of our curation effort are not included. We have recently developed a semi-automated curation system for protein subcellular localization using Textpresso in collaboration with the Wormbase team at Caltech (http://www.wormbase.org/). A brief introduction on this project is also provided.

Results
Overall curation philosophy and information captured

TAIR aims to extract comprehensive *Arabidopsis* gene-related information from published literature. The information we capture from the literature includes:

- gene function (e.g. molecular function, biological process, subcellular localization);
- gene expression;
- allele/polymorphism (this document uses these two terms interchangeably);
- phenotype;
- gene symbols, full names;
- a textual description of the gene.

We no longer curate protein-protein interaction from the literature since we can import this type of data from other resources dedicated to this task (e.g. BioGRID). Due to decreased funding for manual literature curation, we can only curate a subset of the *Arabidopsis* papers published each year. Following are considered high-priority papers:

1) literature of any age pertaining to genes assigned by the Gene Ontology Consortium (as a member of the GOC, TAIR actively participates in the GOC-coordinated annotation projects),
2) literature describing the characterization of previously undescribed ('novel') genes,
3) genes that do not have any Gene Ontology annotations at all,
4) recent literature from high impact factor journals

Information extracted from the literature is integrated into the TAIR database, which now serves as a central access point for *Arabidopsis* data. For each gene, we maintain a regularly updated locus detail page (locus is another way to refer to a gene). TAIR's locus detail page represents the most comprehensive starting point for a researcher to find out what is known about an *Arabidopsis* gene which forms the basis for generating testable hypotheses. We also provide a suite of tools for users to query and analyze the various types of data available at TAIR. To see an example of the locus detail page, go to http://www.arabidopsis.org/servlets/TairObject?name=AT3G52910&type=locus

The current workflow of the operation

Figure 1 illustrates our manual literature curation workflow. The operation can be divided into two phases: triage and curation. At the beginning of each month, we download and prioritize all incoming new papers before moving on to the next step of curation. Each month we search and download from PubMed for articles that contain 'arabidopsis' in the title, abstract or keywords. This is done using a script. Gene names are then automatically extracted from the downloaded abstracts for manual verification by curators. In addition, new gene symbols that are not in our database are identified by a script using a regular expression and are added to the database for curator verification. A priority ranking (high, medium, normal) is automatically assigned to each paper based on journal-impact factor. Curators then review each abstract to determine whether the article contains *Arabidopsis* gene-related information. If the answer is "No", the article is not curated. For papers that contain Arabidopsis gene-related information, the curators then verify the

automatically generated gene-paper links and make corrections when necessary. If the paper describes characterization of a novel gene, the priority is changed to 'first'.

A curator then checks out a first priority paper for curation. Our curation tool allows multiple curators to work simultaneously. This checkout step allows us to better manage the curation task. Upon finishing curation, the curator marks the paper as Scanned (i.e. curated).

Encoding methods

We make extensive use of structured controlled vocabularies such as Gene Ontology (GO) and Plant Ontology (PO) to convert the free text information in the literature into a machine readable format. The resulting annotations are entered into a relational database allowing for query and analysis. Each annotation has a reference allowing us to trace back to the original data source.

The following section describes the data elements curated and controlled vocabularies used.

Genes: The *Arabidopsis* community has developed a nomenclature system where each gene is assigned a unique AGI (Arabidopsis Genome Initiative) locus identifier in a standardized format (e.g. AT5G46330). TAIR is now the central agency responsible for assigning *Arabidopsis* locus identifiers. Many genes also have other names in the literature (e.g. AT5G46330 is commonly known as FLAGELLIN-SENSITIVE 2 or FLS2). TAIR maintains a comprehensive gene name controlled vocabulary file (e.g. AT5G46330-FLS2- FLAGELLIN-SENSITIVE 2-reference, AT5G63580-FLS2- FLAVONOL SYNTHASE 2-reference). During the curation process, we manually verify that the gene name reported in the article is mapped to the correct AGI locus identifier.

Gene function: We use Gene Ontology vocabularies for molecular function, biological process and cellular component to annotate gene function.

Gene expression: We use Plant Ontology vocabularies for plant anatomy and plant growth and developmental stages to annotation gene expression.

Allele/polymorphism: We use sets of controlled vocabulary terms developed at TAIR to capture allele information (e.g. polymorphism type, mutagen) from the literature. We also allow a curator-created free text description to be attached to a polymorphism.

Germplasm: Germplasm refers to strains with unique genotypes. We use sets of controlled vocabulary terms developed at TAIR to capture germplasm information such as species variant, related alleles etc. We also allow a free text description to be attached to a germplasm.

Phenotype: Phenotype is currently annotated as free text. We are working with the community to develop and adopt a phenotype ontology for curating phenotype information.

We curate the following data relationships:
gene-gene function: Each annotation includes the following components:

- gene
- relationship type (linking the gene and the GO term) e.g. involved in, located in
- GO term describing gene function
- evidence type describing how the annotation to the term is supported. We use the GO evidence codes (http://www.geneontology.org/GO.evidence.shtml) in conjunction with TAIR's evidence description controlled vocabulary.
- reference
- annotated by and date

Example: AT5G46330 | involved in | defense response by callose deposition in cell wall (GO:0052544) | inferred from mutant phenotype: biochemical/chemical analysis. | Clay, et al. (2008) | The Arabidopsis Information Resource/ 2009-04-10.

To view examples on the TAIR website, go to:
http://www.arabidopsis.org/servlets/Search?action=search&type=annotation&tair_object_id=2170483&locus_name=AT5G46330

gene-gene expression: Same as above except that Plant Ontology terms are used.

gene-allele-germplasm-phenotype: Each annotation includes the following components:

- allele name and aliases
- gene name
- polymorphism information (polymorphism type, polymorphism site, inheritance etc.)
- a textual description
- phenotype
- reference

Example: fls2-17 | AT5G46330.1 | substitution, exon, recessive | no description available | Mutant seedlings treated with 10μM flg22 peptide (strong growth inhibitor) display shoot and root growth similar to that of wildtype Ler. | Gomez-Gomez, et al. (2000).

To view the complete record, go to:
http://www.arabidopsis.org/servlets/TairObject?id=500245330&type=polyallele

Curation tools

We use PubSearch as our main literature curation system. PubSearch is a web-based literature curation tool that allows curators to search and annotate genes to keywords from articles. It is based on a MySQL relational database for the back-end, and Java Servlet and Java Server Pages running in a Tomcat container for the API and front-end applications. PubSearch was initially developed by the GMOD project

(http://gmod.org/wiki/PubSearch); we have made extensive improvement to this tool recently with new features such as community annotation processing.

The TAIR web application system is a series of web applications running on a two-node, clustered JBoss application server and Apache web server. The application interface consists of Java Server Pages and Java Server Faces pages along with a comprehensive set of CGI-based tools (BLAST, patmatch, and so on). The Java servlets share cached data using a memcached distributed caching server. The data resides in an Oracle relational database. Our curation workflow is supported by extract-transform-load (ETL) pipelines that move data between the PubSearch and TAIR databases on a regular basis. TAIR also hosts a comprehensive GBrowse browser that runs on top of a MySQL database.

Most papers contain all the information we need to make annotations. When a curator runs into a difficult case, e.g. a paper mentions an *Arabidopsis* gene by its symbolic name (non-AGI locus identifier) and this name can be mapped to several AGI identifiers, additional tools and/or databases searches are required. In this example, the curator may do a BLAST search to find the correct locus identifier based on DNA/protein sequence information in the paper; alternatively, the curator could rely on information at TAIR or other databases (PubMed) to disambiguate the gene name. When we need to consult additional papers, we often use TAIR's Textpresso full text search tool to quickly identify relevant information from TAIR's full text literature corpus (27812 full text papers as of December 2, 2011). Occasionally we contact the author for clarification.

Use of text mining tools

In our current literature curation workflow, we use an algorithm to automatically extract gene names from the abstracts. A gene-reference link is also automatically generated during this process. This is then followed by a manual verification step to confirm the gene-reference link is valid. Manual extraction of gene name from literature is a tedious process; this algorithm therefore greatly improves efficiency.

Recently, in collaboration with the Wormbase team, we have developed a procedure to automatically extract protein subcellular localization information from full text literature using the Textpresso text mining tool (3, 4). In this approach, the entire *Arabidopsis* full text literature corpus is processed by Textpresso, sentences that contain *Arabidopsis* gene names, protein subcellular localization data, as well as assay related words are extracted and GO terms are suggested. A curator then manually validates each suggested annotation.

Manual literature curation by professional curators produces accurate and consistent annotations. Yet manual extraction of gene function information from the literature is a labor-intensive process. For example, ~3000 *Arabidopsis* publications are currently published in peer-reviewed journals each year. TAIR's curation team can only cover a fraction of these papers (about 1/3). There is a strong incentive to develop highly effective text mining system to automate the extraction of gene function information from

the research literature using standard, community-developed ontologies of biological concepts. This translates into the demand for tools capable of publication retrieval, entity recognition (gene name, gene function, expression, polymorphism, phenotype etc.) and ontology mapping. A user-friendly curation interface is highly desirable.

References

1. Swarbreck, D., Wilks, C., Lamesch, P., *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*. **36**, D1009-1014.

2. Lamesch, P., Berardini, T.Z., Li, D., *et al*. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. **40**, D1202-1210.

3. Müller, H.M., Kenny, E.E., Sternberg, P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. **2**, e309.

4. Van Auken, K., Jaffery, J., Chan, J., *et al*. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*. **10**, 228.

Funding

This work was supported by the National Science Foundation (grant DBI-0850219) and the National Institutes of Health National Human Genome Research Institute (NHGRI) (grant 5P41HG002273-09 for gene function curation, partial). Additional support for gene function curation comes from the TAIR sponsorship program (see http://arabidopsis.org/doc/about/tair_sponsors/413 for a complete list of sponsors).

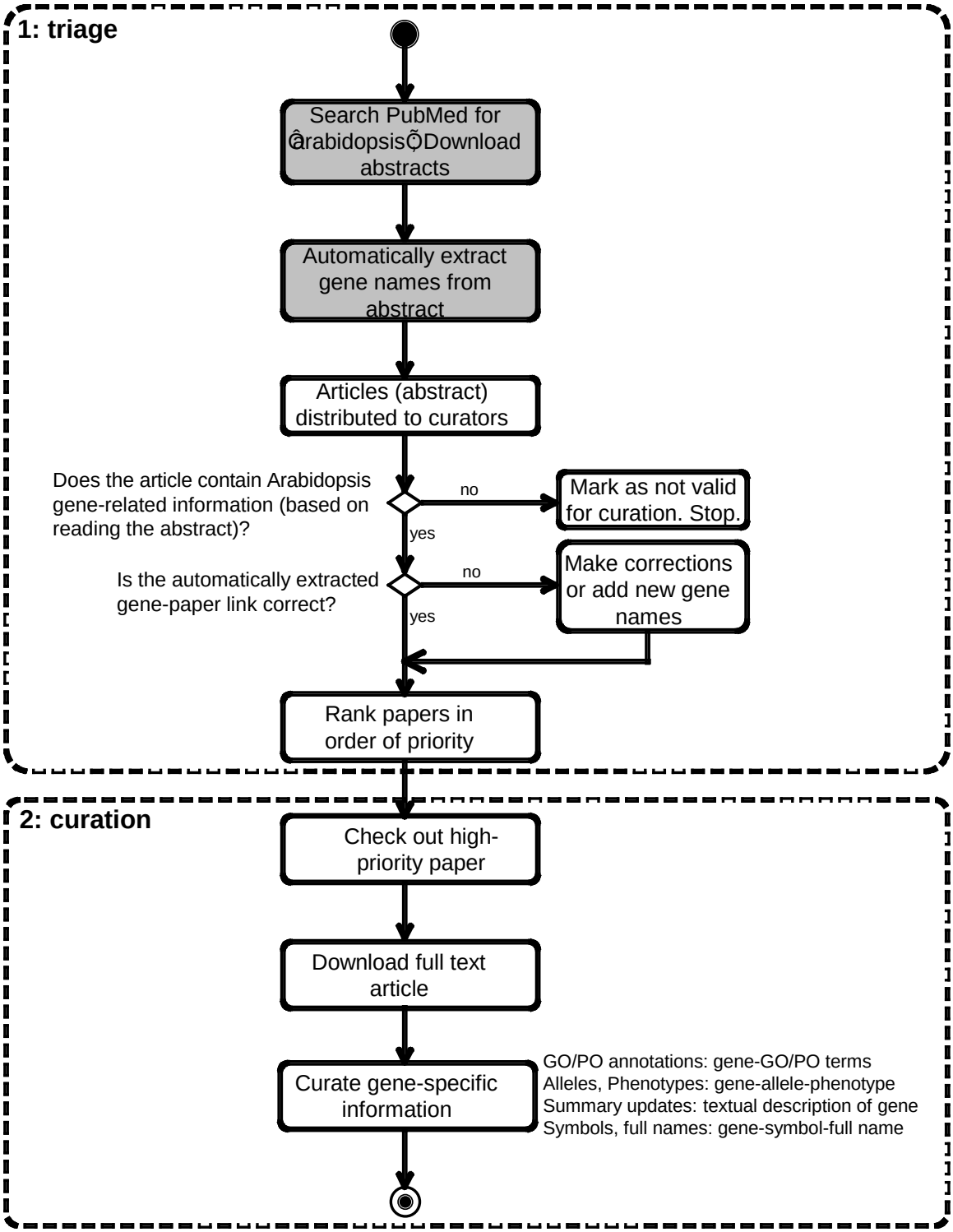


Figure 1. TAIR literature curation workflow. Shaded area indicates automated steps. Full text download is partially but not completely automated.

BioCreative Track II: Summary of Curation Process
for *one component* of the
Mouse Genome Informatics Database Resource
www.informatics.jax.org

INTRODUCTION

MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease. The MGI team curates the biomedical literature (14,000 publications per year) and normalizes and integrates sequence and functional data about mouse genetics and genomics from almost 50 other database and informatics resources. MGI organizes curation teams around particular aspects of the domain including sequence data, phenotypes, embryonic expression data, comparative and functional information, mouse tumorigenesis, and mouse models for human diseases. MGI utilizes multiple bio-ontologies and is the authority for mouse gene and strain nomenclature. Overall, almost 20 professional curators are employed by MGI in addition to the engineers and biomedical analysts who build and manage the system. In 2011, the projects contributing to this resource are:

Mouse Genome Database (MGD) Project: MGD includes data on gene characterization, nomenclature, mapping, gene homologies among mammals, sequence links, phenotypes, disease models, allelic variants and mutants, and strain data.

Gene Expression Database (GXD) Project: GXD integrates different types of gene expression information from the mouse and provides a searchable index of published experiments on endogenous gene expression during development.

Mouse Tumor Biology (MTB) Database Project: MTB integrates data on the frequency, incidence, genetics, and pathology of neoplastic disorders, emphasizing data on tumors that develop characteristically in different genetically defined strains of mice.

Gene Ontology (GO) Project at MGI: The Mouse Genome Informatics group is a founding member of the Gene Ontology Consortium (www.geneontology.org). MGI fully incorporates the GO in the database and provides a GO browser.

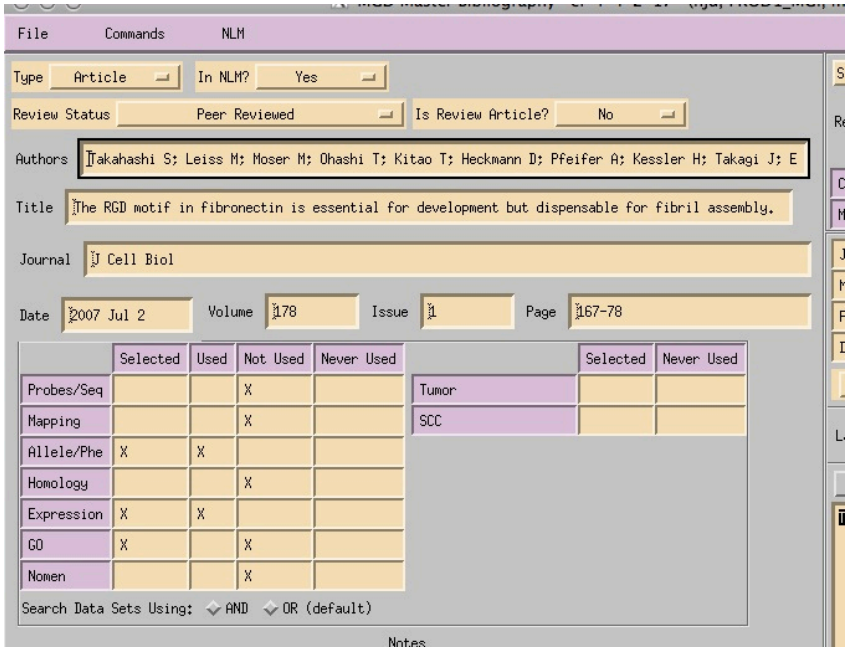
MouseCyc Project at MGI: The MouseCyc database focuses on Mus musculus metabolism and includes cell level processes such as biosynthesis, degradation, energy production, and detoxification. It is part of the BioCyc (http://www.biocyc.org/) collection of pathway databases created at SRI International.

Here we outline the workflow process for one component of the MGI data acquisition and integration process - that associated with the **Gene Ontology Project at MGI**. MGI assigns functional annotations (GO terms) to genes and protein products via automated methods and manual curation. Electronic-based annotation strategies include mapping and translating data from the Enzyme Commission, Swiss-Prot, InterPro, Riken, rat and human ortholog experimental data and others. Curation of these data sets includes review and resolution of Quality Control Reports generated through the process of data loading and comparisons to existing data in MGI. For the purpose of this BioCreative Track II, we will not discuss the automated integration methods, but rather concentrate on the manual literature curation, which is a vital source of experimental mouse functional data. While we focus on the GO component of MGI in this description, the manual literature curation process is very similar for other MGI components that curate literature.

The subcomponents of the literature curation process include:

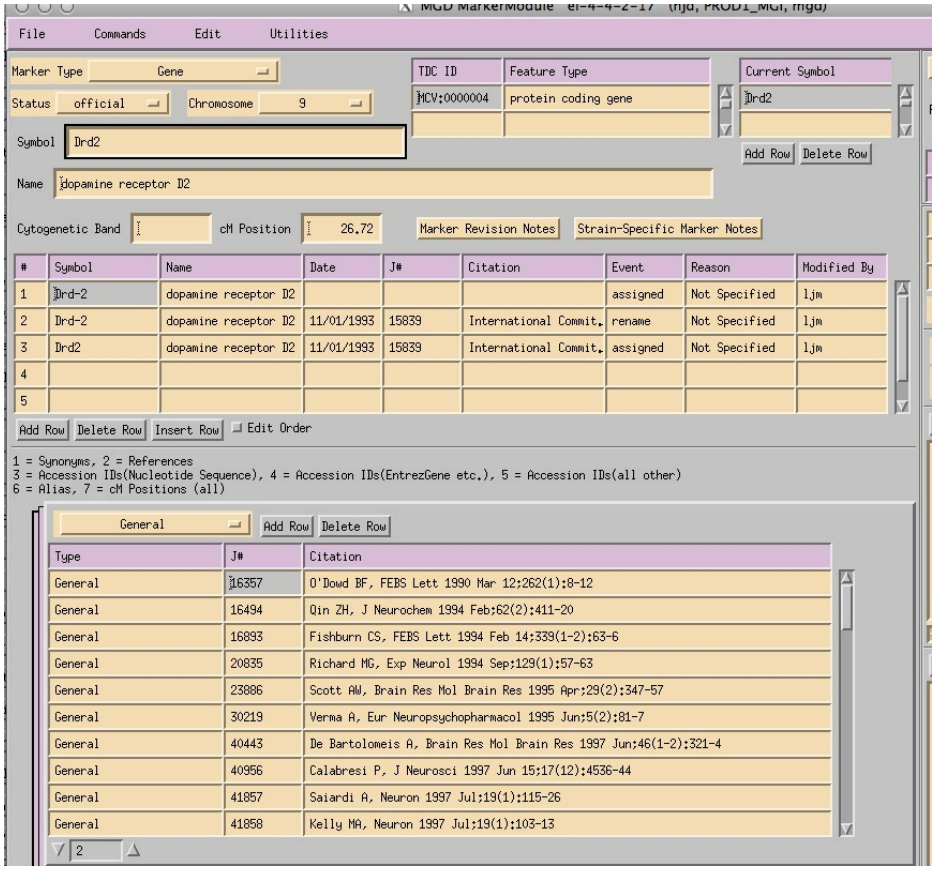
a) *identifying and obtaining relevant scientific literature*
Papers containing data on mouse genes and proteins are first identified using tools such as Quosa (see below). Curators then determine whether the experiments described are of a suitable nature to be used for GO annotation (*i.e.*, do the experiments aid in determining the normal function of the gene?).

b) *adding the publications to the MGI system through the Editorial Interface (EI)*
Papers selected as containing data appropriate for GO annotation are entered into a “master bibliography” module.



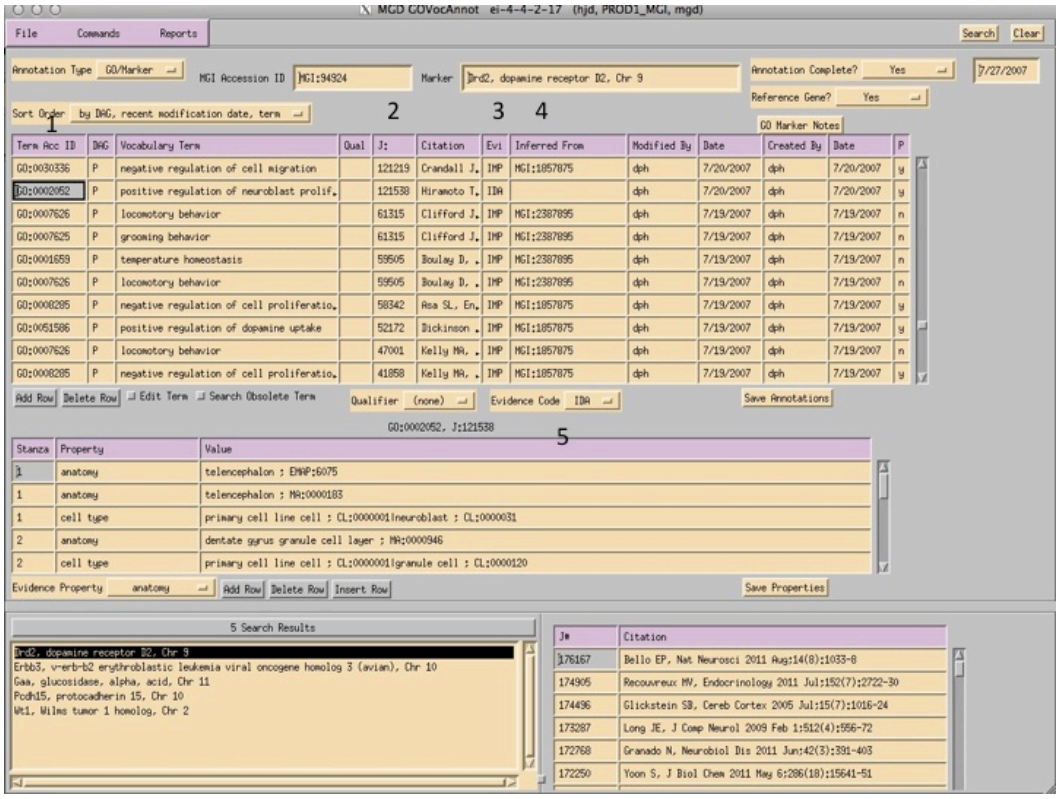
This EI screenshot shows the record for a paper that has been selected for alleles / phenotypes, embryonic expression, and GO. The paper has not yet been curated for GO (indicated by an X in the selected and X in Not Used).

c) *indexing the papers to determining the genes being studied*
A paper is then associated with the genes discussed within by adding the genes to the paper's attributes.



Note, when a paper discusses several genes, not all of them may be objects for direct GO annotation. For example, a paper describing the effects of a knock out of a particular gene may use analysis of other gene products as markers of particular processes being affected, but the annotation to involvement in the process would only be made to the gene being knocked out. Currently, the tables marking the topical areas that the paper is selected for are not directly tied to the genes that the paper is associated with.

d) *creating a GO annotation using the EI module for GO*



A curator at MGI uses the GO EI module shown above to enter annotations. The interface is gene centric. Individual protein isoforms or modified forms can be indicated using annotation extensions called “properties”, where an id for a specific isoform can be indicated. After reading a paper, the curator selects the appropriate GO id and enters it (1). Next, the reference number is added (2), as well as the evidence code (3). If required by the type of evidence code, additional information is added to the inferred from column (4). Once the annotation is saved, information about the cell type that the experiment was done in, or the specific isoform, or tissue, is entered into the annotation properties section (5).

e) tracking metrics and quality control measures to set priorities for upcoming work.
GO annotation metrics in MGI are monitored daily. Annotations are tracked based on a variety of criteria such as annotation source (MGI curation or data load) and evidence (experimental or predictive, such as via orthology or domain). Since the gene ontology changes frequently, the most current version is downloaded from the GO site daily. If a GO term has become obsolete, a report is generated flagging the annotations that use that term for a curator to update. Additionally, we use the master bibliography tables and the GO annotations to keep track of various areas that need focus, such as “genes with no GO annotation but have papers that are selected for GO but not used”.

We envision a time when papers will be automatically identified and tagged using NLP tools and served up to curators ready for final review of the annotations followed by submission to MGI.

ENCODING METHODS

1. *How is information captured to make it machine-readable?*
   GO annotations are shared with the GO Consortium (GOC) via a gene association file (GAF). This is a tab-delimited file that contains most of the elements of a GO annotation as outlined in the GO EI section above. This file is available on either the GOC web site or from the MGI FTP site. This file and the gene ontology vocabulary file are used as input for many analytical tools such as GO Term Finder. Instruction for construction of a GO GAF file are found in GO documentation at http://www.geneontology.org/GO.format.gaf-2_0.shtml.

2. *What entities are involved and how are they entered into the database.*
   Entities involved in the MGI literature annotation stream for functional annotations include the relevant source reference and the genes or proteins being studied. The reference details are entered into the EI Master Bibliography as described above. In addition to references that can be tagged with PubMed IDs, the MGI system tracks data sources using internal references provided either by MGI to describe certain data loads and processes, or by GOC to describe the generation of certain inferential annotations. These internal references have persistent IDs, titles and abstracts.

3. *What relationships are involved and how are they symbolized?*
   All data assertions in MGI are supported by evidence and citation to the source of the information. For assertions that are associated with controlled vocabularies such as the GO, links are provided to vocabulary browsers that provide the relationships between the assertion and other knowledge in that domain.

4. *What standardized or controlled vocabularies are used.*
   For Gene Ontology annotation, we start, of course, with the GO ontology file, which is loaded nightly into the MGI database so that the terms are available within our editorial interface. Additionally, we also use the Cell Ontology, the Mouse Embryonic and Adult Anatomy Ontologies, the Evidence Code Ontology, and PSI-mod to add additional information to an annotation. Currently these ontologies are not loaded into the database.

5. *Give examples of a variety of data elements and how they appear in the database.*
   For GO annotations, MGI provides web-based table, text, and graphical views.

A small portion of the table summary for GO annotations to the *Drd2* (dopamine receptor D2) gene is shown below. Columns include the GO aspect, term name, evidence code, inferred from IDs, and reference. Note that at present the data in the GO EI properties attributes are not accessible in these web views. This information is also available in the GAF file discussed above.

Molecular Function	dopamine receptor activity, coupled via Gi/Go	IMP	MGI:1857875	J:67550
Molecular Function	dopamine receptor activity, coupled via Gi/Go	IMP	MGI:2387895	J:41857, J:65767
Molecular Function	dopamine receptor activity, coupled via Gi/Go	IMP	MGI:3576789	J:97838
Molecular Function	dopamine receptor activity, coupled via Gi/Go	ISO	P14416	J:164563
Molecular	dopamine receptor activity, coupled via Gi/Go	ISO	P61169	J:155856

Using this table and associated information, MGI provides an automatically generated text description of the GO annotations. A snippet of this section is shown below for the *Drd2* gene.

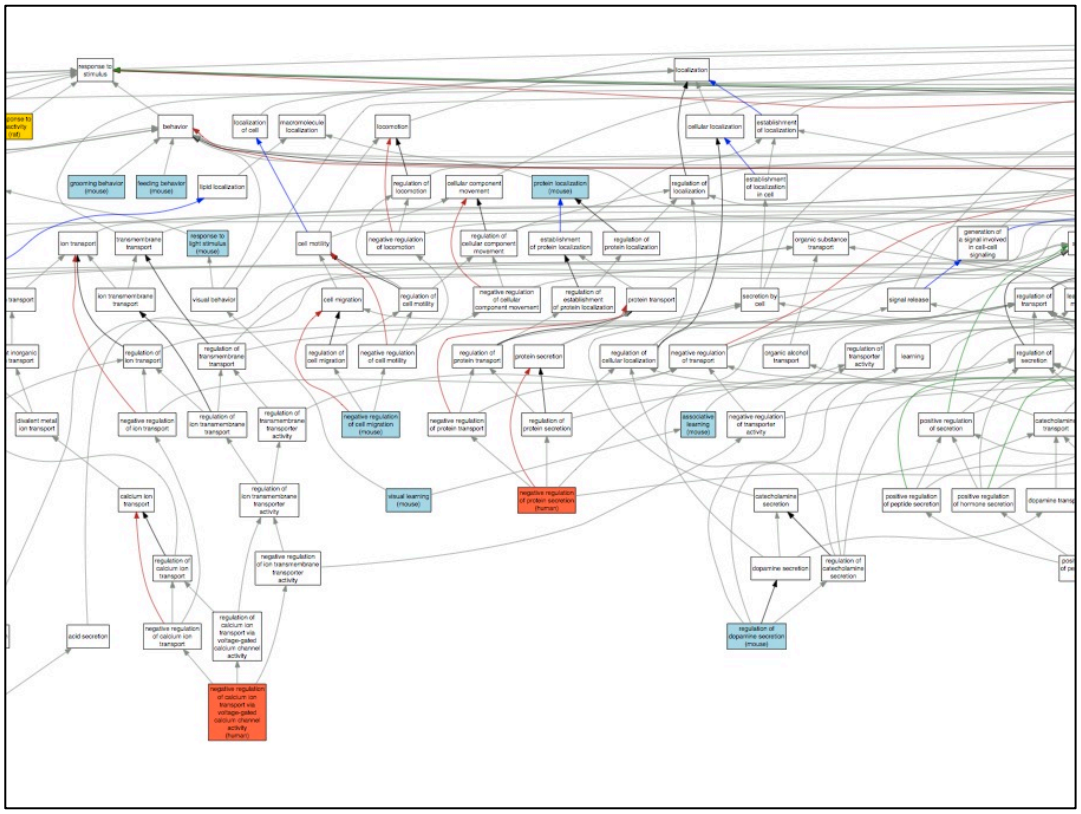
Symbol	**Drd2**
Name	**dopamine receptor D2**
ID	MGI:94924

GO Annotations as Summary Text (Tabular View) (GO Graph)

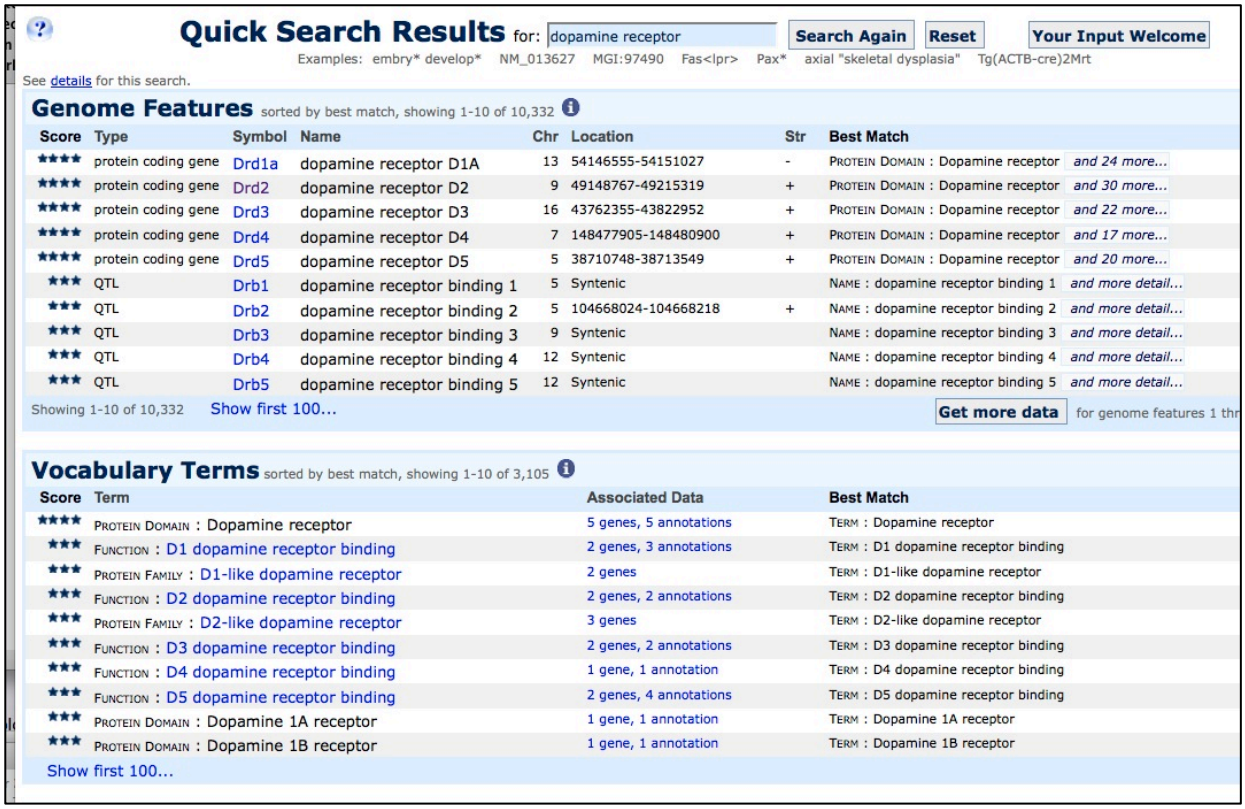
GO curators for mouse genes have assigned the following annotations to the gene product of Drd2. (This text reflects annotations as of Monday, October 03, 2011.) This gene has been selected for comprehensive annotation as part of the Reference Genome Annotation Project of the Gene Ontology Consortium. MGI curation of this mouse gene is considered complete, including annotations derived from the biomedical literature as of July 27, 2007. If you know of any additional information regarding this mouse gene please let us know. Please supply mouse gene symbol and a PubMed ID.

Mouse Genome Informatics Scientific Curators (J:155856; J:164563) suggested that, because of a sequence or structural similarity to UniProt:P61169 (dopamine receptor D2) and UniProt:P14416 (dopamine receptor D2), Drd2 has drug binding activity. Mouse Genome Informatics Scientific Curators (J:155856; J:164563) also suggested that, because of a sequence or structural similarity to UniProt:P61169 (dopamine receptor D2) and UniProt:P14416 (dopamine receptor D2), Drd2 has dopamine receptor activity, coupled via Gi/Go activity. Research by Qin ZH et al and Goldberg MS et al suggests by a direct assay that Drd2 has dopamine receptor activity, coupled via Gi/Go activity. Research by Saiardi A et al, Yamaguchi H et al, Chen JF et al, and Bozzi Y et al using the alleles $Drd2^{tm1Ebo}$, $Drd2^{tm1Mok}$, and $Drd2^{tm1Low}$ suggested that Drd2 has dopamine D2 receptor activity. Mouse Genome Informatics Scientific Curators (J:155856; J:164563) suggested that, because of a sequence or structural similarity to UniProt:P61169 (dopamine receptor D2) and UniProt:P14416 (dopamine receptor D2), Drd2 has dopamine D2 receptor activity. Research by Javitch JA et al, Demotes-Mainard J et al, Ratty AK et al, Baik JH et al, Montmayeur JP et al, Guiramand J et al, Xu R et al, Wang Y et al, and Xu M et al suggests by a direct assay that Drd2 has dopamine D2 receptor activity. Research by Risinger FO et al, Palmer AA et al, and Cunningham CL et al using the allele $Drd2^{tm1Low}$ suggested that Drd2 is involved in the behavioral response to ethanol. Furthermore, research by Ralph RJ et al,

MGI also provides a graphical display of GO annotations from the GO detail page for each gene. A section of this display for *Drd2* gene is shown below.



A user can obtain GO related information using either simple or complex queries. For example, by typing the words *dopamine receptor* in the MGI “quick search box” (below) and clicking the Search button, MGI brings back a list of genes whose names match the query, and also GO (and other) vocabulary terms used by MGI where *dopamine* or *receptor* occurs in the term name or its synonym (e.g., Protein Domain: Kappa opioid receptor or Phenotype: abnormal synaptic dopamine release). The “Associated Data” column provides links to the annotations associated to the vocabulary terms.



Quick Search Results for: dopamine receptor — Search Again | Reset | Your Input Welcome

Examples: embry* develop* NM_013627 MGI:97490 Fas<lpr> Pax* axial "skeletal dysplasia" Tg(ACTB-cre)2Mrt

See details for this search.

Genome Features sorted by best match, showing 1-10 of 10,332

Score	Type	Symbol	Name	Chr	Location	Str	Best Match
****	protein coding gene	Drd1a	dopamine receptor D1A	13	54146555-54151027	-	PROTEIN DOMAIN : Dopamine receptor and 24 more...
****	protein coding gene	Drd2	dopamine receptor D2	9	49148767-49215319	+	PROTEIN DOMAIN : Dopamine receptor and 30 more...
****	protein coding gene	Drd3	dopamine receptor D3	16	43762355-43822952	+	PROTEIN DOMAIN : Dopamine receptor and 22 more...
****	protein coding gene	Drd4	dopamine receptor D4	7	148477905-148480900	+	PROTEIN DOMAIN : Dopamine receptor and 17 more...
****	protein coding gene	Drd5	dopamine receptor D5	5	38710748-38713549	+	PROTEIN DOMAIN : Dopamine receptor and 20 more...
***	QTL	Drb1	dopamine receptor binding 1	5	Syntenic		NAME : dopamine receptor binding 1 and more detail...
***	QTL	Drb2	dopamine receptor binding 2	5	104668024-104668218	+	NAME : dopamine receptor binding 2 and more detail...
***	QTL	Drb3	dopamine receptor binding 3	9	Syntenic		NAME : dopamine receptor binding 3 and more detail...
***	QTL	Drb4	dopamine receptor binding 4	12	Syntenic		NAME : dopamine receptor binding 4 and more detail...
***	QTL	Drb5	dopamine receptor binding 5	12	Syntenic		NAME : dopamine receptor binding 5 and more detail...

Showing 1-10 of 10,332 Show first 100... Get more data for genome features 1 thr

Vocabulary Terms sorted by best match, showing 1-10 of 3,105

Score	Term	Associated Data	Best Match
****	PROTEIN DOMAIN : Dopamine receptor	5 genes, 5 annotations	TERM : Dopamine receptor
***	FUNCTION : D1 dopamine receptor binding	2 genes, 3 annotations	TERM : D1 dopamine receptor binding
***	PROTEIN FAMILY : D1-like dopamine receptor	2 genes	TERM : D1-like dopamine receptor
***	FUNCTION : D2 dopamine receptor binding	2 genes, 2 annotations	TERM : D2 dopamine receptor binding
***	PROTEIN FAMILY : D2-like dopamine receptor	3 genes	TERM : D2-like dopamine receptor
***	FUNCTION : D3 dopamine receptor binding	2 genes, 2 annotations	TERM : D3 dopamine receptor binding
***	FUNCTION : D4 dopamine receptor binding	1 gene, 1 annotation	TERM : D4 dopamine receptor binding
***	FUNCTION : D5 dopamine receptor binding	2 genes, 4 annotations	TERM : D5 dopamine receptor binding
***	PROTEIN DOMAIN : Dopamine 1A receptor	1 gene, 1 annotation	TERM : Dopamine 1A receptor
***	PROTEIN DOMAIN : Dopamine 1B receptor	1 gene, 1 annotation	TERM : Dopamine 1B receptor

Show first 100...

INFORMATION ACCESS

1. *In a difficult case of curation, what kind of information is needed to solve it?*
   In general, markup for GO annotation can be very challenging. The concepts captured by the GO are not limited to simple text matching of terms, but also inferences made via the relationship between the terms, as well as interpretation of what experimental assays can be used as evidence. For example, it is not clear how experiments that can be interpreted as evidence for a gene product's involvement in the regulation of a process could be accurately assessed by anything other than a human curator with background in experimental biology.

2. *What kind of internet searching is used in difficult cases? other databases? wikipedia*?
   Internet searching per se is not really an issue unless a curator needs to obtain additional background in the subject matter of the paper in question. Quite often textbooks or other PubMed articles might be consulted.

USE OF TEXT MINING TOOLS

1. The QUOSA (http://www.quosa.com) application is used to query PubMed for recent papers containing data about the mouse and determining which component of the database will be curated from the paper (GO, expression, mutant alleles, phenotypes, mapping, tumor). QUOSA provides efficiencies for identifying and downloading appropriate literature and supplementary materials from PubMed.

2. PROMINER (http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/products/prominer.html) is used to assist in the gene indexing process. Utilizing official nomenclatures and synonym lists for mouse/rat/human gene names and gene symbols, PROMINER marks up papers for review by a curator, who then associates genes-to-papers in the MGI EI.

3. We are currently working closely with Dr. Gully Burns, Cartic Ramankrishnan, and others at USC to incorporate the SciKnowMine system into MGI curation system. The goal is to replace the triage process such that SciKnowMine can identify relevant papers, and assign appropriate curation areas to that paper (e.g. GO, gene expression, phenotype) based on content. Papers will be ranked for relevance based on knowledge gained from the current MGI Master Bibliography, which includes a literature corpus of over 170,000 references.

The Xenbase Literature Curation Process

Jeff B Bowes[1,*], Kevin A Snyder[1], Christina James-Zorn[2], Virgilio G Ponferrada[2], Chris J Jarabek[1], Bishnu Bhattacharyya[1], Kevin A Burns[2], Aaron M Zorn[2] and Peter D Vize[1]

[1]Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada, [2]Division of Developmental Biology Cincinnati Children's Research Foundation and University of Cincinnati Department of Pediatrics, College of Medicine. Cincinnati, OH 45229, USA

*Corresponding author: Tel: 403 220 2824, E-mail: bowes@ucalgary.ca

Abstract

Xenbase (www.xenbase.org) is the model organism database for *Xenopus tropicalis and Xenopus laevis,* two frog species, used as model systems for developmental and cell biology. Xenbase curation processes centers on associating papers with genes and extracting gene expression patterns. Papers with the keyword "*Xenopus*" are imported into Xenbase and split into two curation tracks. In the first track, papers are automatically associated with genes and anatomy terms, images and captions are semi-automatically imported, and gene expression patterns found in those images are manually annotated using controlled vocabularies. In the second track, PDFs of the same papers are downloaded and indexed by a number of controlled vocabularies and made available to users via the Textpresso search engine and text mining tool.

1. Introduction

a) Overview

Xenbase (www.xenbase.org) (1, 2) is the model organism database for the frog species, *Xenopus tropicalis and Xenopus laevis.* Xenbase is the central source for genomic, gene expression, literature and other experimental data on *Xenopus.* Xenbase also acts as a clearinghouse for *Xenopus* gene nomenclature.

b) Applications

Xenopus is a powerful model system for both developmental and cell biology. Developmental biologists use *Xenopus* embryos to study gene function during development and to model human congenital diseases, while cell biologists use *Xenopus* eggs and oocytes for exploring the mechanistic basis of central processes such as cell division. Xenbase also makes *Xenopus* data accessible to researchers using other model organisms by using gene symbols based on the symbols for human orthologs and by providing links to orthologous genes in humans and similar major vertebrate model organisms (e.g., mouse and zebrafish).

c) Literature Curation Workflow

Currently gene expression patterns are the best supported data type in Xenbase. Our curation workflow focuses on associating papers with genes and anatomy terms, and extracting gene expression patterns. The *Xenopus* literature corpus contains more than 42,000 papers with about 1,500 new papers added per year. Gene expression data are often presented in papers in the form of image evidence- typically images of embryos stained to display the distribution of an mRNA in the various tissues of the organism. Therefore, the principal Xenbase curation workflow centres on images and captions.

The Xenbase literature curation workflow is described in figure 1 and discussed below. (The following 16 steps correspond to the numbered items in figure 1)

1. Every week an automatic process searches for new or updated papers in Pubmed whose abstract, title or metadata contains the keyword "*Xenopus*".

The paper is then imported into two parallel curation processes.

Xenbase Curation Process

The primary process, focussing on annotation of gene expressions and image data, is as follows:

2. Each paper's title, abstract and metadata are imported into Xenbase's DB2 relational database through which they are made available on Xenbase (www.xenbase.org/literature/literature.do) and to the literature curation pipeline.
3. An automatic process identifies mentions of genes and anatomy terms in the abstract and title based on gene symbols and anatomy terms as well as synonyms (see section 2a).
4. Curators select papers for curation based on the presence of gene expression evidence and whether Xenbase may reproduce images either through open access licensing or special permissions.
5. Curators initiate a process that automatically scrapes images from journal websites.
6. Curators choose which images have relevant information (e.g., gene expression) and import those images.
7. Captions for imported images are automatically annotated to identify genes and anatomy terms.
8. Curators assign content types to both papers and images from the paper. This may be types of content we currently curate (e.g., gene expression) or content types we plan to curate in the future once support is added to Xenbase (e.g., phenotypes).
9. The curator annotates image captions for gene expression patterns via a custom web-based curation system. In the future, this curation is likely to be expanded to include additional data types such as phenotypes, gene regulation, transgenics, and protein localization.
10. The curator updates the status to indicate whether or not curation is complete.

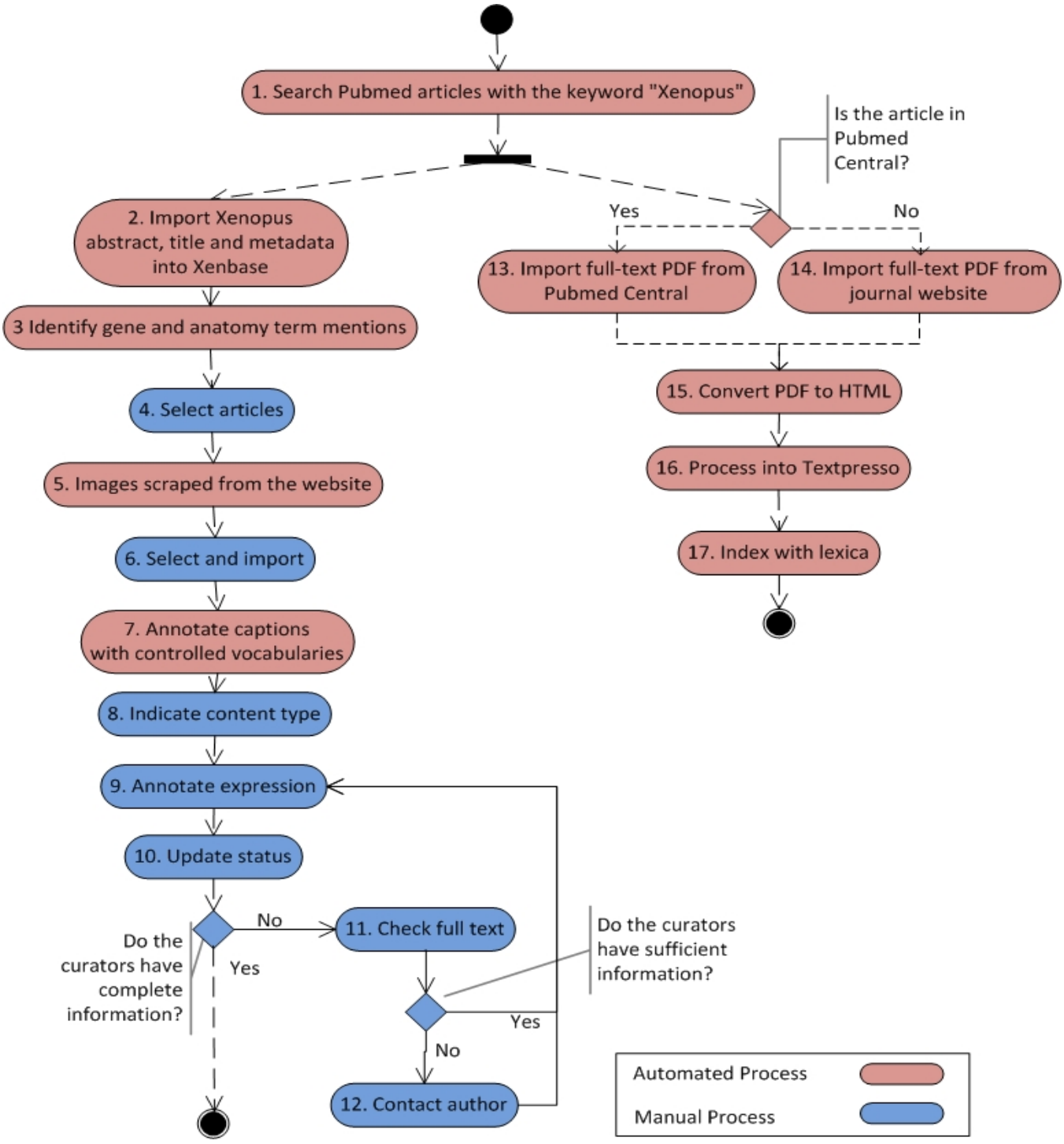


Figure 1: Xenbase Literature Curation Workflow

11. If there were insufficient data in the image captions to complete the annotation, the curator will proceed to the paper text to obtain the required information. (See section 3.)
12. If the curator is still missing information, they may contact the author. (See section 3.)

Textpresso Process

In the second process, the full texts of papers are imported into Xenbase's Textpresso (3) text mining and search tool (see section 4a).

13. If the paper is in Pubmed Central, its full text is downloaded from Pubmed central.
14. If the paper is not in Pubmed Central, the PDF for the journal is imported from the journal web site.
15. PDF documents are converted to XML.
16. Documents are processed into Textpresso (see section 4a).
17. Documents are indexed using Textpresso controlled vocabularies.

2. Encoding Methods

a) Data Entities and Controlled Vocabularies

Genes: We capture which genes were expressed in an experiment. Because *Xenopus laevis* is pseudotetraploid, having two versions of each gene (paralog-a and paralog-b), we also must consider which version of each gene was expressed. These are identified via a controlled vocabulary of gene symbols and synonyms drawn from the Xenbase gene catalog. Xenbase gene symbols are used by NCBI Genbank for *Xenopus* gene naming and are based on the names of human orthologs. Synonyms represent historical or alternative identifiers for genes. This is especially important as the same gene is often referred to by more than one name in archival literature. Xenbase allows any user to enter synonyms for genes.

Clones: Xenbase attempts to identify which cDNA clones were used in *in situ* hybridization experiment. These are identified via references to NCBI accession numbers.

Antibodies: For immunohistochemistry experiments various data on the antibody are captured.

Species: We capture which of the *Xenopus* species, *laevis* or *tropicalis,* was used in the experiment.

Anatomy Items: Anatomy items (i.e., tissues, organs, cell types), where genes are expressed, are represented using the Xenopus Anatomy Ontology (XAO) (4). The XAO is structured as a directed acyclic graph (DAG) and the XAO is constructed using best practice outlined by the OBO foundry allowing it to be interrelated to Anatomy Ontologies from other model organisms.

Development Stages: Nieuwkoop and Faber (NF) development stages, defined in (5), have long been the accepted standard for delineating periods of development in *Xenopus*. NF stages are also included in the XAO.

Content Type: Xenbase uses a small ontology structured as a DAG to describe different content types of papers. These include antibodies, *in situ* hybridizations and phenotypes.

Other: Numerous additional controlled vocabularies are used by the Textpresso. Some examples are gene regulation relationship, morpholino and antibody terms. These are all structured as single level controlled vocabularies with synonyms.

b) Data Relationships

The core relationship that we currently capture in literature is which gene is expressed in which place, at what time (development stage) and for which species. Further metadata on these experiments such as specific clone or antibody identity used are also captured.

c) Sample data elements

- www.xenbase.org/literature/article.do?method=display&articleId=39749 provides an example of a curated article. Note that names of genes and anatomy items are hyperlinked in the abstract and image captions. (To view captions click "show captions".) Clicking on an image will open a box showing the image, caption and a table of curator-entered annotations.
- www.xenbase.org/gene/expression.do?method=displayGenePageExpression&geneId=484814&tabId=1 is the expression tab of the sox3 gene page. Click an image under the "Summary" or "Literature Images" section to see sample annotations.
- The XAO can be downloaded from obofoundry.org/cgi-bin/detail.cgi?id=xenopus_anatomy

3. Information Access and Curation Bottlenecks

The main information access issues that curators experience are as follows:

Missing Information: The authors have not entered pertinent information on the experiment. For example, the authors may not have specified which species of *Xenopus* was used, which *X. laevis* gene paralog was tested or the Genbank accession number of the clone used in the experiment. In the case of antibodies there may be missing information on its construction. This requires first searching the entire text of the paper, literature search of secondary references describing the missing data and then, if necessary, contacting the author.

Anatomy term not in XAO: There are cases where the author has described expression in a tissue that is not currently represented in the XAO. This requires the curator to research the tissue, determine whether the term is a synonym of an existing XAO term or whether it is a valid term missing from the XAO. Xenbase curators are continually expanding the XAO and keeping it in sync with other model organism ontologies. Curators will define the new terms and integrate them into the XAO in relation to existing terms. This may involve examining anatomy ontologies for other vertebrates such as mouse or zebrafish or through consultation with other ontology development teams such as National Center for Biomedical Ontology (NCBO)(www.bioontology.org) , Uberon (www.uberon.org), the common amphibian anatomical ontology (www.amphibanat.org) or Phenoscape (www.phenoscape.org).

4. Use of text mining tools

a) Current Tools

Link Matching: Our link matching tool identifies gene and anatomy term mentions in titles, abstracts and captions. This uses a combination of inverted indices and regular expressions to match gene symbols, anatomy terms and their synonyms. In the case of genes, synonyms may be added by any user. Terms with common homonyms can be entered into a table of exclusions that are ignored by the matching process, to reduce false positives. Identified genes or anatomy terms are hyperlinked to gene and anatomy pages, respectively. This tool comes into play at both steps 3 and 7 in our curation workflow diagram. Although this is a fairly basic approach, the identification of synonyms is a particularly important step allowing this tool to associate genes and anatomy terms from different papers, despite the plethora of alternative gene names and variant anatomical descriptions used in the scientific literature.

Textpresso: We have implemented Textpresso, a biological text search and mining tool. Textpresso is used to index the full text of documents, and in particular, index the corpus by controlled vocabularies. Textpresso also segments the paper into sections such as abstract, body, discussion, materials, results and citations. Currently, Textpresso is used as an advanced query tool (www.xenbase.org/cgi-bin/textpresso/xenopus/home) that allows users to return documents or sentences matching particular criteria. Users can pose questions such as “return sentences with two genes and a regulation term” or “return sentences containing a gene mention and a Morpholino from the materials section”.

In the near future we plan to expand Textpresso’s application to identify papers that contain information on antibodies and/or morpholinos for particular genes. This information will be presented to users on Xenbase gene pages.

b) Unsolved Problems

There are a number of problems where text mining could be applied to improve our curation workflow.

At steps 3 and 7, our current technology does an effective job of finding gene and anatomy term mentions, especially by taking advantage of synonyms. However, we are aware that more effective text mining methods for capturing this information exist. Furthermore, because of numerous ways used to describe a range of development stages (e.g., stages 1,2,3 and 4; 1-5; St. 1 to 5; etc.) our current methods do not capture this information. Furthermore, it would be useful to capture other entities such as Gene Ontology terms or NCBI accession numbers.

At step 4 of the process, classifiers that could identify the content of papers using our content type ontology could help in triaging papers for curation.

Much more ambitiously, at step 7, actual gene expression relationships could potentially be captured from captions. If false positives were kept to a reasonable level, extracted relationships could be approved or edited by curators, increasing the efficiency of the process. This would be a difficult

problem to solve. If this was made possible, extracting other relations such as gene regulations (i.e., gene *a* regulates gene *b)* or phenotypes (e.g., knocking out gene *a* results in phenotypes 1, 2 and 3) would also be valuable.

At step 16, Textpresso attempts to segment papers into sections. However, it does this poorly. It would be very useful to segment papers well. For example, we have found that associating papers with gene mentions produces many false positives as paper titles in the paper's references mention genes that are unrelated to the work described in the paper. Being able to properly distinguish between the references section and other sections of a paper would allow us to more accurately associate papers with genes by excluding genes referenced in the citations section. Extending the markup of papers to identify paragraphs and part-of-speech tagging would also be extremely useful.

As current curation in Xenbase is very image and caption centric it may miss information found only in the full body text of papers. Beyond gene expression, other types of curatable information may not be presented via images (e.g., gene regulation). While limited curation resources still preclude examining every paper, we have considered developing a pipeline that would use Textpresso to extract sentences from papers that may contain useful biological information (gene expression, antibodies, phenotypes, morpholinos, etc.). This could also be implemented with other text mining tools. The interface would be designed to allow curators to approve, reject or edit extracted facts and zoom out from sentences to surrounding text to further assess facts in the data.

5. Funding

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health [5P41HD064556]

6. References

1. Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, Noumen E, Pollet N, Vize PD. (2008) Xenbase: A Xenopus biology and genomics resource. *Nucleic Acids Res*., **36** (Database issue), D761-7.

2. Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, Zorn AM, Vize PD.(2010) Xenbase: Gene expression and improved integration. *Nucleic Acids Res*., **38** (Database issue), D607-12.

3. Muller HM, Kenny EE, Sternberg PW. (2004) Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*., **2(11)**, e309.

4. Segerdell E, Bowes JB, Pollet N, Vize PD. (2008) An ontology for Xenopus anatomy and development. *BMC Dev Biol*. **8**, 92.

5. Nieuwkoop PD, Faber J. (1994) Normal table of *Xenopus laevis* (Daudin). Garland, New York.

Summary of the FlyBase-Cambridge Literature Curation Workflow

Peter McQuilton* and the FlyBase-Cambridge Literature Curation team
FlyBase, Department of Genetics, University of Cambridge, CB2 3EH, UK.

*Corresponding author: Tel: 0044 (0) 1223 333963, Email: pam51@gen.cam.ac.uk

1. Introduction

1a. Overview

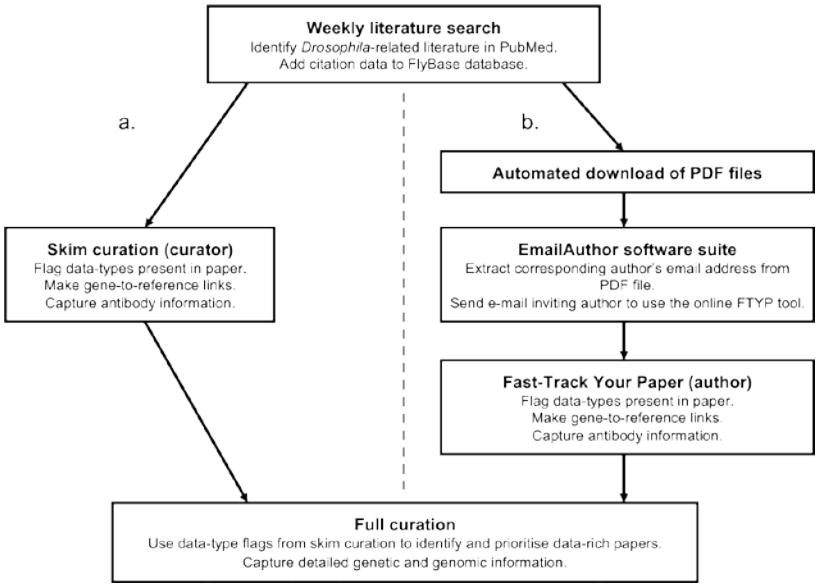
FlyBase (www.flybase.org) is the leading database for genomic and genetic information on the model organism *Drosophila melanogaster* and other members of its clade. The FlyBase web site currently contains 2.5 million pages covering 19 different data classes and 20 Drosophila reference sequenced genomes. Major sections include phenotypic, gene expression and interaction data. In addition, FlyBase maintains millions of links to other key resources such as strain and clone repositories, expression databases and sequence databanks.

The FlyBase project is spread over 4 sites, each with its own remit. At FlyBase-Cambridge, we focus on the manual curation of genetic entities (e.g. genes, mutant alleles, transgenic constructs) and their associated phenotypes from the published Drosophila literature. We also annotate genes with Gene Ontology (GO) terms. Data are curated into simple text file templates, submitted to the central FlyBase database, and subsequently made available for browsing and searching on the FlyBase website.

1b. Curation Workflow

We perform a weekly search of NCBI-PubMed for Drosophila-related publications and add the basic citation details to the FlyBase database. We estimate that approximately 2,150 primary research papers are published in English on Drosophila each year, which is more than can be curated in depth by our curators and so have had to introduce methods to prioritize papers for curation.

Each Drosophila-related paper initially undergoes rapid 'skim curation'. T his involves (i) associating the main genes with the paper, and (ii) attaching internal flags to papers to indicate the type(s) of data they contain, such as a newly characterized gene, or the generation of a novel mutant allele or transgene. We call this second step 'triaging'.



Recently, we have successfully engaged the Drosophila community in skim curation. First, we added a 'Fast-Track Your Paper' (FTYP) tool to the FlyBase website that allows users to effectively skim curate their own (or other's) papers. W e have subsequently taken this process further by developing software that automatically emails the corresponding author of newly published Drosophila paper directly inviting them to use our FTYP tool.

In addition to the above, we have recently begun work on a text-mining method to triage uncurated papers. More information on this can be found in section 4a.

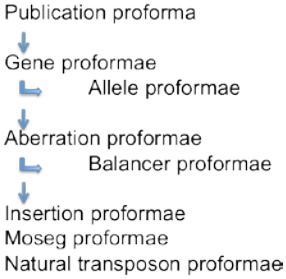
After triage, papers are placed into a list for 'full curation'. Those papers with the largest number of flags are considered the most data-rich and are curated first. Using this approach, each paper is prioritized based solely on its relevance to FlyBase.

In 'full curation', literature curators extract genetic entities and related molecular and phenotypic data from the Results (text, tables and figures) and Methods sections of a paper. Additional types of data include Gene Ontology (GO) terms, genetic interactions, allelic interactions, aberrations, allele classes, and Construct data.

2. Encoding methods

2a. Overview

Data from a paper are added to structured proforma using a range of valid symbols (and recorded synonyms), controlled vocabularies (see table) and free text. These structured text file templates are known as proformae. Several proformae are



bundled together to form a 'curation record' for that paper. Each curation record undergoes quality control checking and is then parsed into the FlyBase chado database.

2b. Description of Entities captured, their relationships and symbolization

We capture data from the literature in curation records composed of structured proforma arranged in a hierarchy, starting with a single publication proforma, followed by gene proformae and nested allele proformae where necessary, then by transgene insertion proformae, and natural transposon proformae. These proformae represent the report pages found on the FlyBase website. Those generated from data curated in FlyBase Cambridge are shown in table 1, along with the ID prefix we use and example entities. The proformae are opened, edited, and saved in a standard text editor.

All data curated from a paper are explicitly attributed to that paper in the database. Multiple relationships exist between the curated data types. For example, the *dpp*^{EP2232} **allele** is caused by the *P{EP}dpp*^{EP2232} **insertion** and disrupts the *dpp* **gene**.

Table 1 Reports in FlyBase.

FlyBase Report	*Example*	*Unique ID*
Aberration	Df(3R)BSC678	FBab0045744
Allele	dppEP2232	FBal0089340
Balancer	TM6B	FBba0000057
Gene	dpp	FBgn0000490
Insertion	P{EP}dppEP2232	FBti0010414
Natural Transposon	P-element	FBte0000037
Recombinant Construct	pGaTN	FBmc0000380
References	St. Pierre and McQuilton, P. (2009)	FBrf0213427
Transposon	P{UAS-dpp.GFP.T}	FBtp0013609

Each proforma is composed of a number of fields, which are split into four main types:

I. Symbol/ID fields
II. Free text description fields
III. 'Soft CV fields' – where we use a limited set of terms to describe a feature in a semi-controlled manner
IV. Controlled Vocabulary fields – where we generate lines composed of controlled vocabulary terms from a number of different ontologies.

2c. Controlled Vocabulary Usage

Along with the use of controlled symbols and unique IDs we use a number of controlled vocabularies (in the OBO file format) to annotate genes and alleles. Those used in literature curation are shown in table 2.

Table 2. Controlled Vocabularies used in Literature Curation.

Ontology	*Example term*	*Term ID*
fly_anatomy.obo	dMP2 neuron	FBbt:00001602
fly_development.obo	pupal stage P6	FBdv:00005353
Term Qualifier	nutrition conditional	FBcv:0000714
Phenotypic Class	smell perception defective	FBcv:0000404
Sequence Ontology (SO.obo)	engineered_foreign_gene	SO:0000281
Publication Ontology	paper	FBcv:0000212
Origin of mutation	in vitro construct	FBcv:0000455
Allele Class	Amorphic allele	FBcv:0000688
GO Cellular Component	Golgi cisterna membrane	GO:0032580
GO Molecular Function	acetyl-CoA transporter activity	GO:0008521
GO Biological Process	mRNA processing	GO:0006397

2d. Examples

Examples of filled-in proforma fields are shown below. The text between the exclamation mark (!) and the colon (:) is the proforma field label, and the text entered after the colon is the curated data.

Symbol/ID fields:

! G1a. Gene symbol to use in FlyBase *a :ftz

! G1b. Gene symbol(s) used in reference *i :DmFtz

Free text description field:

! GA7a. Phenotype [free text] *k :Expression of @ftz[LRAAA.YPPWLK.Scer\UAS]@ in developing imaginal discs under the control of @Scer\GAL4[Dll-md23]@ transformed antennae to complete legs with five distinguishable segments and a malformed A3 segment.

Soft CV fields:

! G28b. Source for merge/identity of [SoftCV] *q : Source for merge of: NT1 Spz2

! GA12a. Nature of the lesion, wrt GA11 - nt/aa changes [SoftCV] *s :Amino acid replacement: A112N.

Controlled Vocabulary fields:

! GA56. Phenotype (class | qualifier(s)) [CV] *k :neuroanatomy defective | heat sensitive

! GA17. Phenotype (anatomy | qualifier(s)) [CV] *k :adult leg | ectopic antenna
mechanosensory bristle | supernumerary

3. Information access

The most common problem encountered during curation is an ambiguous genetic entity (gene, mutant allele, transgene etc.). This situation can arise when a unique identifier (such as a FBgn or CG number for genes), or an accurate and explicit

reference for a mutant or transgenic line is not given. Ambiguity is a particular problem in the following cases:

- a generic symbol/name is used (e.g. '*Actin*', *UAS-Notch*)
- a symbol/name is used that is a synonym for a different gene (e.g. '*ras*' is the current FlyBase symbol for the '*raspberry*' gene (FBgn0003204) but is often used in the literature to refer to the '*Ras85D*' gene (FBgn0003205)
- the FlyBase symbol for the gene only differs in case from that of another FlyBase gene (e.g. *dl* (*dorsal*, FBgn0260632) is a different gene from *Dl* (*Delta*, FBgn0000463))

These ambiguities can usually be resolved by:

- Searching for associated details about the entity in the paper (e.g. use of a specific mutant allele can identify the gene being discussed)
- Consulting the supplementary files for additional details
- Performing a BLAST search using any sequence data present in the paper or supplementary files
- Executing an in-house script that reports entities used by a specified author in previously curated papers
- Emailing the corresponding author for clarification.

In addition to FlyBase, other internet sites used to resolve ambiguities include stock centers (e.g. Bloomington Drosophila Stock Center- http://flystocks.bio.indiana.edu/, Vienna Drosophila RNAi Center- http://stockcenter.vdrc.at) and the various databases incorporated within the National Center for Biotechnology Information (NCBI).

If the ambiguity cannot be resolved, then a curator will either associate a generic/unspecified entry for that entity with the paper, or else omit the entity and add a (non-public) note to the curation record explaining the situation.

4. Use of text-mining tools

4a. Automatic triaging of papers by datatype using SVM

In collaboration with WormBase, we use SVM (Support Vector Machine) methods to triage primary research papers into categories based on our skim/author curation flags. We run the SVM on the full text of a paper, extracted from the published PDF. So far, we have trained the SVM to triage papers for new alleles, new transgenes and gene renames. We have a number of other triage flags for which we haven't yet trained the SVM that may be more amenable to other text-mining methodologies.

4b. Possible uses of Text-mining in our curation workflow

We can envisage text-mining being used at multiple points within our curation workflow:

Triaging of papers - Further to our current SVM work, we can imagine text-mining being used to categorize papers based on broad GO terms or gene function. GO terms are not covered by our triage flags. We are investigating the use of Textpresso towards this end (see Van Auken et al. BMC Bioinformatics. 2009 PMID:19622167.).

Symbol extraction – Similar to our work collaborating with WormBase and the GSA (see references), we hope to use text-mining to extract genetic entity symbols from literature.

A text-mining equivalent of our skim curation – Combining symbol extraction for genes, alleles and insertions with a text-mining triaging method, we hope to be able to fully automate the skim curation process.

Text mark-up – Using text-mining to suggest Controlled Vocabulary terms such as GO terms for specific genes from particular regions of text, along with gene and allele symbols and data triage flags. This would speed up manual curation of a paper.

5. Recent FlyBase References

FlyBase 101 - the basics of navigating FlyBase. McQuilton P, St Pierre SE, Thurmond J; the FlyBase Consortium. *Nucleic Acids Res.* 2011 Nov 29. [Epub ahead of print] PMID:22127867.

Toward an interactive article: integrating journals and biological databases. Rangarajan A, Schedl T, Yook K, Chan J, Haenel S, Otis L, Faelten S, DePellegrin-Connelly T, Isaacson R, Skrzypek MS, Marygold SJ, Stefancsik R, Cherry JM, Sternberg PW, Müller HM. *BMC Bioinformatics*. 2011 M ay 19;12:175. doi: 10.1186/1471-2105-12-175. PMID:21595960.

Inside FlyBase: biocuration as a career. St Pierre S, McQuilton P. *Fly (Austin).* 2009 Jan-Mar;3(1):112-4. Epub 2009 Jan 6. PMID: 19182544.

A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Mungall CJ, Emmert DB; FlyBase Consortium. *Bioinformatics*. 2007 Jul 1;23(13):i337-46. PMID: 17646315.

Incorporating text-mining into the biocuration workflow at the AgBase database.

Lakshmi Pillai[1,2*], Catalina O. Tudor[3*], Philippe Chouvarine[2], Carl J. Schmidt[4**], K Vijay-Shanker[3**], Fiona McCarthy[1,2**]

1. Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University.
2. Institute of Genomics, Biocomputing and Biotechnology, High Performance Computing Collaboratory, Mississippi State University.
3. Department of Computer and Information Sciences, University of Delaware.
4. Department of Animal and Food Sciences, University of Delaware.
* & ** indicate joint authorship

Abstract

AgBase provides annotation for agricultural gene products using the Gene and Plant Ontologies (GO & PO), as appropriate. Unlike model organism species, agricultural species have a body of literature that does not just focus on gene function; as a result early biocuration workflow analysis at AgBase determined that biocurators spend up to 50% of their biocuration time identifying suitable papers to annotate. To solve this problem we use text-mining to identify literature for curation. Since AgBase provides biocuration for multiple species, our Gene Prioritization (GP) interface ranks gene products for annotation based upon user requests and their presence in commonly used microarrays. Biocurators select the top ranked gene and mark annotation for these genes as 'in progress' or 'completed'; links enable biocurators to move directly to our Biocuration Interface (BI). Our BI includes all current GO annotation for gene products and is the main interface to add/modify AgBase curation data. The BI also displays eGIFT results for each gene product. eGIFT is a web-based, text-mining tool that associates informative terms (iTerms) and sentences containing them, with genes. iTerms are ranked based on a score which compares the frequency of occurrence of a term in the gene's literature to its frequency of occurrence in documents about genes in general. Moreover, iTerms are linked to GO terms, where they match either a GO term name or synonym. Since most agricultural species do not have standardized literature, eGIFT searches all gene names and synonyms to associate papers with genes. Since many of the gene names can be ambiguous, eGIFT applies a disambiguation step to remove matches that do not correspond to this gene. Another additional filtering process is applied to retain those abstracts that focus on the gene rather than mention it in passing. eGIFT also links an iTerm to sentences mentioning the term, allowing biocurators to rapidly scan the relation between the iTerm and the gene. The BI is also linked to our Journal Database (JDB) and as literature based annotations are added, the corresponding journal citations are stored in JDB. Just as importantly, biocurators also add to the JDB citations that have no GO annotation. The AgBase BI also supports bulk annotation upload to facilitate our IEA annotation of agricultural gene products and all annotations must pass standard GO Consortium quality checking prior to release in AgBase.

Introduction

Amongst databases and resources that provide literature-based biocuration there are two broad approaches for targeting biocuration. In the first approach, biocurators regularly triage all published literature to identify papers that are likely to contain information to be biocurated. This approach works particularly well where the literature is focused to a well-defined set of journals and there is a larger research community. In the second approach, biocurators target certain gene sets and, for each gene in this set, do comprehensive literature searches to identify all annotation for this gene. The advantage of this approach is that this approach can target well studied gene sets and biocurators are able to provide a comprehensive annotation set for a gene or gene product. Naturally, these approaches are not exclusive; as biocurators from different databases collaborate to provide co-ordinated and consistent annotation, biocurators may change their biocuration approaches to suit their needs.

The AgBase database provides functional data for agricultural researchers via both sequence-based functional annotation and manual biocuration of published literature (1). Our literature annotation is currently focused on providing Gene Ontology (GO) annotation for agricultural gene products from chicken, cow, corn and cotton. To do this we utilize GO Consortium best practices and procedures while adapting our biocuration process to the specific needs required for agricultural literature. This body of literature differs from that of model organism species in that it includes a large body of work addressing general production issues, genetic markers and trait characterization; this data does not typically contain information that can be biocurated to the GO. In addition to identifying gene products that have been studied directly in agricultural species, we must also identify which of these gene products are likely to have literature that will yield GO annotations. Moreover, we wish to identify genes that are important for the agricultural research community and provide GO annotation to support further functional modeling. As a result we do not attempt to provide detailed, literature-based functional annotation for every gene product but rather focus on annotating prioritized gene sets for the agricultural species on which we work.

Biocuration Workflow

This paper describes the pipeline we developed to prioritize our annotation effort, ensure that we rapidly and accurately identify literature for GO annotation, capture annotation and literature details, do quality checking and generate gene association files (Figure 1). The individual components of this pipeline are described in more detail in the following sections. We are now in the process of extending this pipeline to include simultaneous annotation to the Plant Ontology (PO), as appropriate.

Gene Prioritization (GP) Interface. AgBase biocurators target manual biocuration using a Gene Prioritization interface that ranks genes based upon user requests or presence on microarrays. Gene information is loaded from the NCBI Entrez Gene database and a GP interface is generated for each species for which we provide literature-annotation. (Note that an exception to this rule is the cotton GP interface – since the cotton genome is not yet sequenced and the literature for agriculturally important cotton gene products spans four taxa, we currently provide the cotton GP based upon gene products for these four species as a single interface.) In each case genes are ranked using a simple scoring system where they are assigned a count for each time they have a gene product that occurs on commonly used microarrays. When

researchers request annotations via the AgBase Community Requests & Submissions page, the count is also incremented. Researchers who request annotations are provided with a view only link to access the relevant GP list so that they can track their request in the queue. Prioritization based upon a mixture of functional genomics platforms and researcher requests enables GO annotation for each species to reflect community needs and current interests. In addition, biocurators can also add annotation requests (for example, as part of a collaborative project to annotate defined gene sets).

Biocurators with access to the AgBase biocuration pipeline access the GP interface and select the species on which they are working. The GP interface shows the ranked list of genes, gene names, their current count and links from the GP to the corresponding gene product in the main biocuration interface. To ensure that each biocurators can track not only their own annotation but also ongoing biocuration by others in the group, a biocuration status menu is displayed. The status menu can be set to display "annotation in progress" or "annotation complete" and the name of the biocurators (based upon login) is displayed. The biocurator selects the top ranked genes, sets the menu to indicate that they are working on this gene and uses the gene product link to go directly to the corresponding gene product in the main biocuration interface. Once the gene is marked as complete, its count is zeroed and the date recorded.

Each GP list is updated as part of the regular AgBase database update procedure. Genes with a count that become obsolete are flagged for biocurators analysis and reassignment, if possible. This process helps us to ensure that the AgBase mappings from array IDs to gene product accessions remain current.

Biocuration Interface (BI). The main BI may be accessed directly or via the GP interface, enabling biocurators to also add GO annotations for gene products that do not have a specific GP list. Agricultural literature that we biocurate often contain functional data for more than one gene product or species and it is our policy to annotate all GO functional information from papers that we manually curate.

Our BI is specifically designed for GO annotation and is updated to ensure that we are able to capture additional data fields mandated in the new GO gene association file format (gaf 2.0). With each update we load all manual GO annotations provided by the GO Consortium as well as computationally-derived GO annotations (Inferred from Electronic Annotation or IEA annotations) for those species that we specifically target our manual biocuration to. By loading species-specific IEA annotations provided by EBI GOA biocurators are able to see at a glance the likely GO terms that they may expect to find in their literature searches. By loading manual annotations for other species, we are able to transfer GO annotations amongst species where there is evidence of sequence or structural similarity.

We use standard quality checking scripts developed by the GO Consortium to prompt biocurators during the annotations process; biocurators are prompted to ensure that all mandatory fields are completed, they are completed in the expected format, GO IDs are not obsolete and so forth. Information about the biocurator and the date of annotation are auto-filled. Since we are annotating species for which there are no model organism databases or reference gene set and we do not restrict our annotation to proteins from UniProtKB, biocurators may also add gene

product record pages using accessions from NCBI if there is no UniProtKB record available or in cases where the protein record is not appropriate (e.g. functional RNAs). Another feature that we found necessary for our biocuration pipeline was the ability to add literature records not found in PubMed. Since PubMed does not contain all agricultural functional literature, biocurators may also add references from agricultural literature collections such as Agricola, the National Agricultural Library Catalog. GO Consortium guidelines allow this practice, although where possible we prefer to use PubMed IDs. We also include a free text comment field for biocurators to capture additional, pertinent information. Most typically this is used to note a new GO term request, or until the recent change in the GO gaf, to capture links to other ontology terms.

eGIFT Integration. Another novel tool that we use to focus our manual biocuration effort is the extracting Genic Information From Text tool (eGIFT) (2). eGIFT searches PubMed to identify literature associated with specific genes and associates informative terms, called iTerms, and sentences containing them, with these genes. We link iTerms to GO terms and display these in the AgBase BI. Integrating eGIFT with our BI enables AgBase biocurators to rapidly identify publications for GO annotation.

We incorporate eGIFT into our annotation pipeline by displaying this information in the top right hand corner for each gene product page of our BI. As a biocurator moves from the GP to its linked gene product page in the BI, they immediately see a summary of functional literature associated with the gene. Since each eGIFT iTerm is linked to corresponding GO terms, the eGIFT information is easily displayed as a list of GO terms found in the literature for that gene with links to the associated literature and additional links to the full eGIFT information page. Thus the BI page displays not only existing GO annotation but possible functional literature classified by GO terms. This enables that biocurator to rapidly assess which eGIFT predicted GO terms may already be annotated to the record and what additional functional information they may capture. Displaying this information on one page means that the biocurator can rapidly move from gene selection to annotation of literature, without intervening literature triage or searching.

Journal Database (JDB). Like many other biocuration projects we also track the literature that we annotate. Since we initially had difficulty in identifying literature that contained GO annotation and literature is only poorly associated with agricultural genes in any case, we sought to collect information about the papers that we identified for annotation. We developed the AgBase Journal Database (JDB), which we have subsequently integrated into our biocuration pipeline. A key feature for the JDB, which differentiates it from other JDBs for biocuration projects for which we are aware, is that we not only record articles which have GO annotation (which may also be parsed directly from the GO gaf) but also articles that do not have GO. While we intended this latter feature to measure our ability to identify functional literature, it also provides a source of manually curated true negatives for text-mining projects.

When a biocurator enters an annotation in the BI that contains a PubMed reference, this data is also recorded in the JDB. The JDB collects the PubMed ID, authors, title, journal and citation details and links out to the PubMed record. In addition there is a link in the BI where biocurators also record PubMed IDs that were manually curated and found to have no GO data or add references that have no PubMed ID. The JDB creates a link between the BI record and the

literature, which is also available to the public. Again since there are no MODs for the species that we annotate and no dedicated effort to link literature to genes and gene products, AgBase biocurators also routinely submit NCBI Gene Reference Into Function (GeneRIF) notes, most particularly when the literature they identify is not already linked to the Entrez Gene record. We also provide a JDB free text comment field that can, for example, be used to note information that may be relevant to ontologies other than the GO (most commonly tissue expression, cell type, post-translation modifications). The JDB not only provides links between genes products and literature but is also used to analyze which journals we frequently use (or would wish to use if access were provided). This data is forwarded to our institutional library for consideration.

Bulk Annotation Upload. In addition to providing an interface that allows biocurators to add annotations directly, we also provide a bulk annotation upload feature that enables us to add our sequence-based annotations from computational annotation pipelines, such as those generated by InterProScan analysis and mapping to GO. During the upload process annotations are quality checked to ensure that they meet current GO Consortium standards and that they are not duplicating existing GO annotations. Biocurators have the option of selecting either to discard duplicated annotations from the upload file or to use these annotations to override existing annotations in the BI with the newer annotations. Any annotation errors are flagged for manual review prior to loading into the BI.

Quality Checking and Error Reporting. As previously mentioned, our biocuration workflow does quality checking and error reporting at two stages – during the initial annotation entry (either via the BI or as part of the bulk upload process) and again during the regular AgBase updates. During the AgBase update process the quality checks are more extensive and only annotations that pass are forwarded to the public AgBase database. Annotations that are flagged as errors are passed to the associated Error Reporting interface where biocurators may view these based upon either the type of error or by biocurator; this enables biocurators to manually review and correct any annotation error prior to their public release.

The types of errors that we check for include errors in the GO Consortium produced GO annotation file checking script, as well as internal checks to ensure that our annotations map to existing records in AgBase and are not duplicated within our workflow. Examples of errors include: mandatory annotation fields not completed; unrecognized field formats; use of obsolete GO IDs; instances where an external accession is incorrectly linked to more than one record in the BI; and automatic checks to ensure that annotations transferred based on sequence or structure homology are still relevant.

Generating Annotation Reports. Another feature linked to our biocuration workflow is the ability for biocurators to quickly generate annotation reports and subsets of annotations contained in our BI. We do this by allowing them to set up Boolean searches based upon the gaf fields (including date and submitter). This allows biocurators to rapidly report how many GO annotations exist for a particular species, how many GO annotations were added during a specific time frame or to quickly find specified sets of annotations that they added. This simple interface reports both the number of GO annotations and gene product records returned by a search, as well as allowing the results to be downloaded as a gaf.

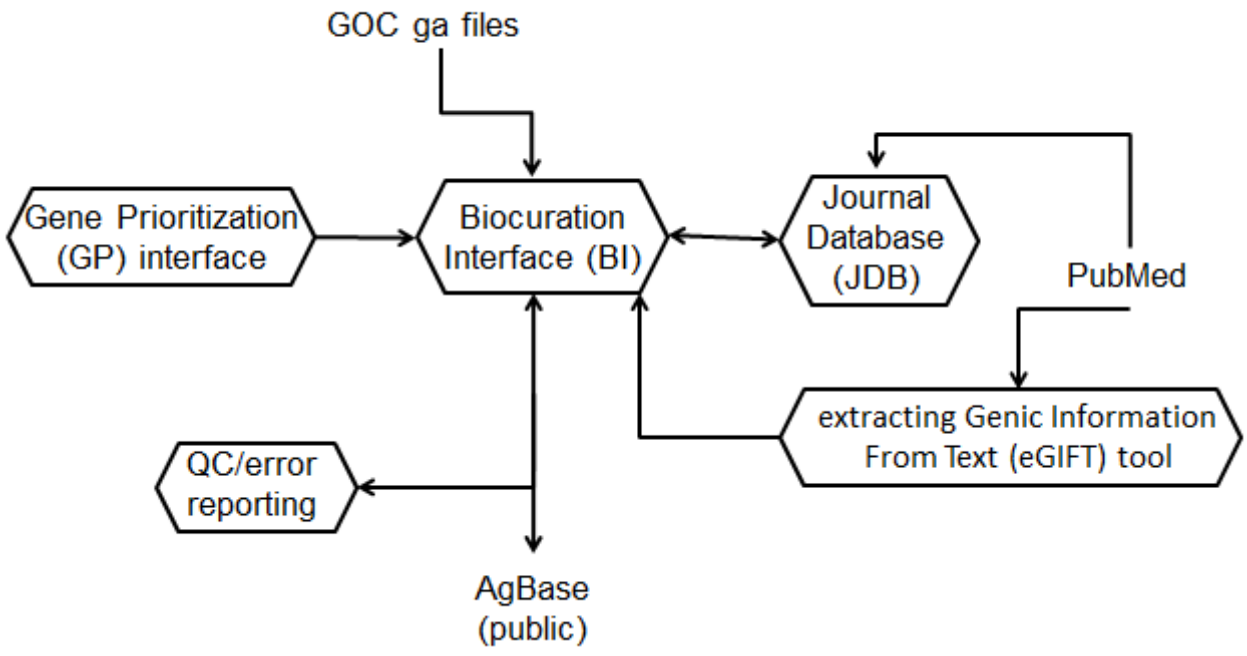
Funding. This work was supported by the US Department of Agriculture Agriculture and Food Research Initiative [grant number 2011-67015-30332], the US Department of Agriculture Cooperative State Research, Education and Extension Service [grant numbers 2007-35205-17941 and 2008-35205-18734] and the National Institutes of Health National Institute of General Medical Sciences [grant number 07111084].

References.

1. McCarthy, F.M., Gresham, C.R., Buza, T.J., *et al.* (2011) AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Res*, **39**, D497-506.
2. Tudor, C.O., Schmidt, C.J., Vijay-Shanker, K. (2010) eGIFT: mining gene information from the literature. *BMC Bioinformatics*, **11**, 418.

Figure 1. The AgBase biocuration pipeline. The AgBase biocuration pipeline draws from GO Consortium gene association files and from PubMed data and the output are the AgBase public gene association files. Briefly, genes to be annotated are prioritized as a ranked list in the Gene Prioritization (GP) interface, which are linked to records in the main Biocuration Interface (BI). The eGIFT tool enhances the ability for biocurators to identify and curate appropriate literature while the Journal Database (JDB) records reviewed literature. Biocuration must pass standard GO Consortium error and quality checks prior to public release.

Figure 1.



Curation at the Maize Genetics and Genomics Database

www.maizegdb.org

1. INTRODUCTION

a. Overview

History. In its first decade (1991-2003) as MaizeDB, a major role was the comprehensive capture from the literature of genes, their map locations, sequences, variations, gene functions, and agronomic traits/quantitative traits. Of some 2000 papers/year, about 500 were carefully annotated. These were largely pulled from an electronic subscription to Current Contents. Full text of relevant papers was photo-copied, and distributed to one of 3 full/part-time curators with different expertise (genetic mapping; gene functions and biochemistry; and quantitative traits), and who read papers to glean information for the database. Typically less than 500 papers were curated, either because there was no reference in the text to a maize gene or allele, or to trait inheritance. Data were stored in to relational database (Sybase as DBMS) developed in consultation with Stan Letovsky and Mary Berlyn both then at the *E Coli* Stock Center (New Haven, CT). Information [allele names, phenotypes, type of mutation, associated literature] about maize mutants was, and is still, directly entered, using forms, into the database by the Maize Genetics Cooperation Stock Center (MGCSC) or Coop (Urbana, IL).

Currently. The same range of experimental data are manually curated, but the focus of the database is now (1) the integration of the maize reference genome sequence with increasingly large electronic datasets of maize researchers, and (2) supporting agile access by breeders and maize researchers to these and other external data. The curation staff spends much time developing tutorials and helping researchers prepare large datasets for integration at MaizeGDB. Because of staff limitations, a limited amount of the functional genomics literature is carefully curated. See also Cannon et al 2012; Harper et al 2011; Schaeffer et al 2011.

b. Applications

The major users of the database are basic and applied researchers in maize genomics and breeding, and who require access to the current reference genome sequence, its functional and structural annotation, and related data stored elsewhere (Cannon et al 2012). Curator review and interaction with the community is a major part of the process (Schaeffer et al 2011). New tools include a metabolic pathways representation, based on the software that supports MetaCyc (Casi et al 2011).

2. CURATION See also Figure 1.

a. Triage

Currently, of the over 5000 articles/year for maize, we curate some 150:

(a) 60-70 papers recommended by our Editorial Board, which rotates each year and is expert in many aspects of maize biology and breeding. Curation is done by a single curator and has been the newest curator on the team.

(b) Other papers recommended by the community (<10-20/year).
(c) papers cited by contributors of maize gene reviews, a project to encourage community data curation (less than 40/year).
(d) As time permits, additional papers with newly sequenced maize genes (40-50/year)
(e) GO annotation to support a metabolic pathways database for maize, based on the BioCyc pathway tools of Peter Karp [this effort began summer 2011]. We plan to manually review only a few pathways, in consultation with our Working Group.

b. Links to the literature. See also Figure 2.
These are hardcoded using the PubMed ID, and DOI. We also provide links from each citation to several search tools at PubMed and GOOGLE (Google, Google Scholar, Usenet). These are set to search using the title of the article and are on all individual reference pages.

c. Data Capture
Information capture is typically by use of a form, with selectable values (for small controlled vocabularies), or requiring manual entry, with (1) lookup-capability that examines a synonyms table, and (2) nominal business rules/alerts. In addition, for literature, a tool based on TEXTPRESSO (Müller et al 2004) is used to load bibliographic citations based on their PUBMED_IDs, and supports curator review prior to entry into the database. Very large datasets from researchers may be managed by electronic loading into a copy of the database, where the curator is able to create tables and compare to existing data. We use an Oracle DBMS.

3. ENTITIES

A large number of entities are involved, with a focus on maize genome and function. Main entities are described here. More details are stored at the database website under documentation.

a, Locus (genes, probed sites, quantitative trait loci, etc.);
Variations of Locus (including cytogenetic inversions and translocations)
Genetic Markers/Probes of a Locus
Nomenclature is a major bottleneck. Often different names are used for the same object, or no gene name is specified. Many times a sequence accession can be used to define identity. MaizeGDB is the clearing house for maize gene nomenclature; however, we do not assign names, but advise researchers, both when asked and sometimes proactively. For specific problems in this area, we request input from the MGCSC (Stock Center) and/or a researcher who has dealt with this issue, either in maize or another species. For global problems that cannot be readily resolved, we request input from the Maize Nomenclature Committee, responsible for updating the guidelines used by the community. A MaizeGDB curator is a member of this committee. Other bottlenecks include discerning the inbred or germplasm used to derive a gene sequence; which of the computed gene-model name corresponds to a classical gene symbol; whether there is recombinational/genetic mapping to a chromosomal region. Much of this information is specified in footnotes, legends to figures and tables, or scattered in the text.

b. Gene Products
Bottlenecks are largely in nomenclature. We use the names provided in the literature, referring to UniProt/BRENDA for more widely accepted terms. Else we use the gene symbol name as a placeholder. Other bottlenecks include discerning the best evidence that the gene encodes a gene product. We are beginning to use GO; finding the best GO term takes time. Updating records to keep up with GO annotation changes, as it becomes more granular will also be a challenge.

c. Phenotypes and associated Germplasm/Stocks
Phenotypes include both qualitative and quantitative descriptors. Much of the phenotype entry is done by the MGCSC. At MaizeGDB we categorize phenotypes by a standardized trait vocabulary largely developed in 1995 at MaizeDB. We have made some mappings at upper levels to the Gramene trait ontology (Polacco et al 2006).

d. Terms [subcellular stages, tissues, developmental stages, relationship types, evidence codes, relations, chemicals, etc.].
These have developed over the past 2 decades. Maize-specific terms for tissues and developmental stages have been annotated to the Plant Ontology terms and accessions, and associations are provided to the PO group. GO is new for MaizeGDB, and is implemented largely to support a maize instance fo the BioCyc pathway tools.

4. RELATIONSHIPS

Two types of tables are used:

a. Relationships between members of the same entity are stored in a single table, “Relations”. An example would be “locus X regulates locus Y”.

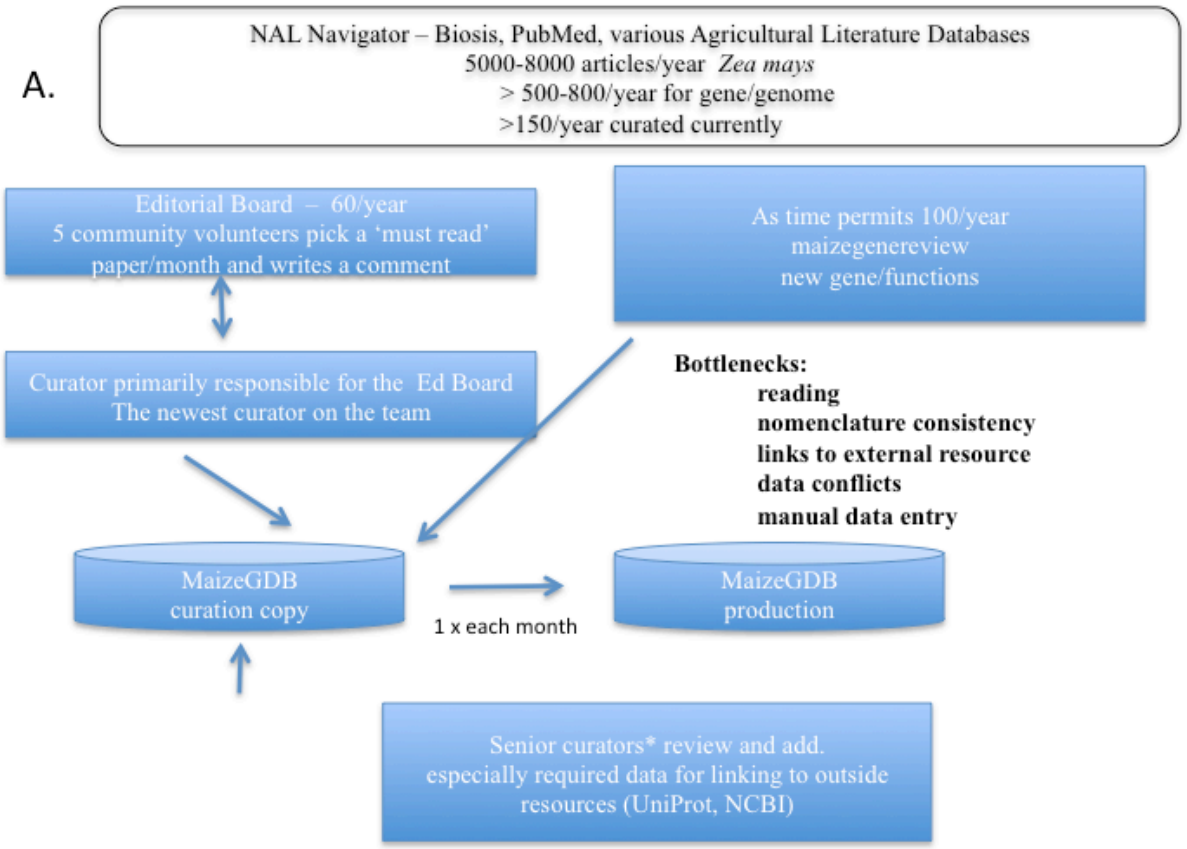
b. Relationships between members of two different entities are stored in tables specific to those two entities, for example, the table Locus_Gene_Product, used to represent relationship such as ‘Locus X’ encodes ‘Gene Product XYZ’

Both types of tables include the ‘relationship’, the ‘method’, and the information source (Person, Project and/or Literature). Controlled vocabularies are used for relations, and for methods.

5. REFERENCES

Cannon EK et al 2011 Int J Plant Genomics PMID 22253616
Caspi R et al 2012 Nucleic Acids Res 40:D742-753.
Harper LC et al 2011 Database (Oxford) 2011:bar022 PMID 21565781
Müller et al 2004 PLoS Biol 2:e309 PMID 15383839
Polacco ML 2006 Maydica 51:357-367.[*www.maydica.org/articles/51_357.pdf]*
Schaeffer ML et al 2011 Database(Oxford) 2011:bar022. PMID 21624896

A.



*senior curators include MaizeGDB staff (1.5 persons); the MGCSC (Maize Genetics Cooperation Stock Center) staff (2 persons); a retired former director of the project.

B.

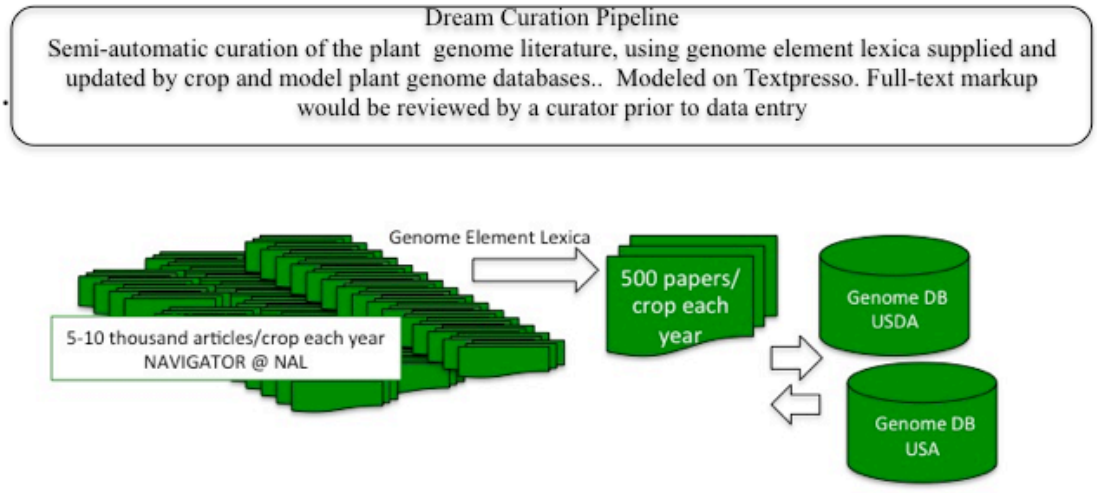


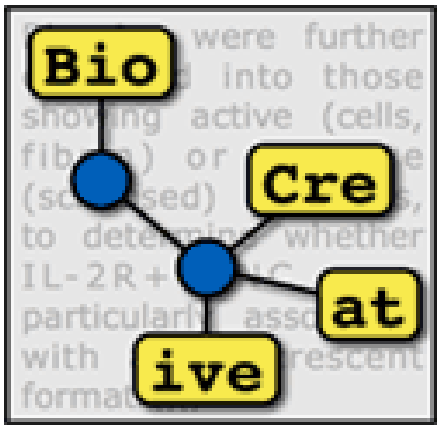
FIGURE 1 WORKFLOW at MaizeGDB for Literature Curation

FIGURE 2. Links to Literature

Hard-coded Links
PubMed ID to NIH
DOI to Journal Article -- http://dx.doi.org/10.1038/ng.551

External Resource Searching
PubMed, Google, Usenet
Example above is for Google Scholar

Track 3



An Overview of the BioCreative Workshop 2012 Track III: Interactive Text Mining Task

Cecilia N. Arighi[1,2*], Ben Carterette[2], K. Bretonnel Cohen[3], Martin Krallinger[4], John Wilbur[5] and Cathy Wu[1,2]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE

[2]Computer and Information Sciences Department, University of Delaware, Newark, DE

[3]Center for Computational Pharmacology, University of Colorado Denver School of Medicine, Aurora, CO

[4]Structural and Computational Biology Group, Spanish National Cancer Research Centre, Madrid, Spain

[5]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD

*Corresponding author: Tel: 302 831 3444, E-mail: arighi@dbi.udel.edu

Abstract

An important question is how to make use of text mining to enhance the biocuration workflow. A number of groups have developed tools for text mining from a computer science/linguistics perspective and there are many initiatives to curate some aspect of biology from the literature. In some cases the curation effort already makes use of a text mining tool, but there has not been a broad-based systematic effort to study which aspects of a text mining tool contribute to its usefulness for a curation task. Here we report on an effort to bring together a number of text mining tool developers and database curators to test the utility of tools and to set the stage for a more intense study of these issues in the coming year at the BioCreative IV Workshop.

Introduction

The BioCreative-2012 Workshop on Interactive Text Mining in the Biocuration Workflow aims to bring together the biocuration and text mining communities towards the development and evaluation of interactive text mining tools and systems to improve utility and usability in the biocuration workflow. Track III is an interactive text mining and user evaluation task. Like the BioCreative III interactive task [1], it is non-competitive, and the main goal is to engage users and collect specifications and metrics that will set the stage for the BioCreative IV challenge to be held in April 2013. Hosting the workshop as a satellite to the International Biocuration meeting provides a unique opportunity to engage curators in this activity.

Lessons learned from BioCreative III interactive task (IAT)

In the BioCreative III interactive task the goal was 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 included ranking genes by importance (based preferably on the amount of described experimental information regarding genes). There was also an optional task to assist the user in

finding the most relevant articles about a given gene. For this purpose a user advisory group (UAG) was assembled and played an important role in critiquing IAT systems, and in providing guidance for a future, more rigorous evaluation of IAT systems [1].

An important lesson learned from this activity was that providing specifications about a desired system is not sufficient; developers and users should team up early on and work together throughout the process of system development. Another relevant observation was that the users' adoption of automated tools into their curation process will depend heavily on performance and on the overall convenience of a tool.

Methods and Results

1-Interactive Task in BioCreative 2012

To explore some of the issues brought up by the UAG and described in the Introduction, we introduced Track III in the BioCreative workshop. On one hand, the track was open to any biocuration task, as opposed to BioCreative III, which focused on gene normalization/ranking; on the other hand, systems were asked to comply with a set of requirements to ensure the tools' performance and scope would be adequate for the proposed biocuration task (see Recruitment of Text Mining Teams). In addition, curators and developers were paired early in the process and, when possible, systems were tuned to the user's curation interest.

2-Recruitment of Text Mining Teams:

We openly invited text mining teams to participate in the interactive task by presenting systems that focused on any specific biocuration task, as opposed to Track I, which was tied to document prioritization. Registered teams (registered by November 15) were requested to submit a document describing their system and addressing the following questions:

1. Relevance and Impact: Is the system currently being used in a biocuration task/workflow?
2. Adaptability: Is it robust and adaptable to applications for other related biocuration tasks (i.e., can be utilized by multiple databases/resources)?
3. Interactivity: Does it provide an interactive web interface for biocurators' testing?
4. Performance: Can the developer benchmark the system and provide performance metrics prior to our evaluation?

In addition, teams should indicate the limitations of the system, provide details on the biocuration task, and suggest evaluation metrics.

Eight teams registered and provided the required system description by the agreed deadline (December 31, 2011), but only six provided benchmarking results and were ready for testing by March 5, 2012 (**Table 1**). The remaining two systems were invited to the workshop to participate in the demo and poster sessions, but not evaluated at the pre-workshop stage. The reported

metrics of the benchmarking provided some evidence that the system should be in good conditions to be evaluated during the workshop.

The biocuration tasks proposed were widely heterogeneous, including information extraction (gene-disease relationships, protein-protein interactions), gene mention, gene normalization, ontology matching, and document retrieval (based on disease, chemical, and protein acetylation). Each system was assigned a coordinator to supervise and assist in the activity.

The list of systems and the accompanying documentation were posted on the BioCreative website [http://www.biocreative.org/tasks/bc-workshop-2012/track-iii-systems/] (Table 1) for curators to review and eventually sign up for testing.

Table 1 – Systems registered in Track III BioCreative 2012.

System	Tasks	Articles	Reported Benchmark		
			Precision (%)	Recall (%)	F-measure
TextPresso	Curation of subcellular localization using Gene ontology cellular component	Full-Text	66.7 (document level), 80 (+SVM, document level), 80.1 (sentence level)	95.2 (document level), 76 (+SVM, document level), 30 (sentence level)	NR
PCS (Charaparser)	Curation of Entity-Quality terms from phylogenetic literature using ontologies	NA	90 (Term-based EQ), 52 (Label-based EQ)	90 (Term-based EQ), 51 (Label-based EQ)	NR
PubTator	Document triage (relevant documents for curation) and bioconcept annotation (gene, disease, chemicals)	Abstract	86.73 (gene mention, abstract), 56.23 (gene normalization, full text), 85.42 (species recognition, abstract)	82.36 (gene mention, abstract), 39.72 (gene normalization, full text), 85.42 (species recognition, abstract)	84.49 (gene mention, abstract), 46.56 (gene normalization, full text), 85.42 (species recognition, abstract)
PPInterFinder	Mining of protein-protein interaction for human proteins (abstract and full legth articles):document classification and extraction of interacting proteins and keywords.	Abstract	81.28	71.27	75.94
eFIP	Mining Protein Interactions of Phosphorylated Proteins from the Literature. Document classification and information extraction of phosphorylated protein, protein binding partners and impact keyword	Abstract	95.52 (abstract level), 84.44 (sentence level), 86.11 (sentence level goldstandard)	96.96 (abstract level), 86.36 (sentence level), 60.78 (sentence level goldstandard)	NR
T-HOD	Document triage for disease-related genes (relevant documents for curation) and bioconcept annotation (gene, disease and relation)	Abstract	77.1	76.3	76.7
Tagtog*	Protein/gene mentions recognition via interactive learning and annotation framework	Abstract	NR	NR	NR
Acetylation*	Document retrieval and ranking based on relevance on protein acetylation	Abstract	NR	NR	NR

NR= not recorded, *Not fully evaluated at pre-workshop

3-Recruitment of Biocurators

We invited curators to participate in the evaluation of a system of their choice based on the list provided in Table 2 prior to the BioCreative workshop. The invitation was distributed via the International Society for Biocuration (ISB) mailing list, and the ISB meeting and BioCreative websites. As described above, the list of the systems was posted on the BioCreative website. A total of 21 curators registered and participated in this activity, curating the gold standard or evaluating the systems. Each system had in the end 2 curators for the evaluation. Table 2 shows the wide variety of databases represented by participating curators.

Table 2-Participating Databases/Institutions sorted based on selected categories.

Numbers in parentheses are the number of curators from each institution.

Category	Affiliation	Category	Affiliation
MOD*	Dictybase (2)	Phenotype	GAD (1)
	MGI (3)		Phenoscape (3)
	RGD (1)	Pathway	Reactome (2)
	SGD (1)	Ontology	Plant ontology (1)
	TAIR (1)	Pharma	Pfizer (1)
	Zfin (1)		Merck (1)
PPI*	MINT (1)	Literature	NML (1)
	BioGrid (1)		

*MOD stands for Model Organism Database, and PPI for protein-protein interaction.

4-Coordinators

Coordinators are members of the BioCreative workshop steering committee who assisted in supervising and facilitating the communication between curators and developers. Some of the roles of the coordinators included: i) matching and introducing curators to systems, ii) supervising the creation of the corpus to serve as a gold standard for use in the evaluation, iii) overviewing the activity, iv) ensuring participation of the teams at the workshop (registration), v) guiding curators on the steps needed to complete evaluation, and vi) collecting metrics.

Coordinators introduced curators to matching system developers and the latter provided the support needed to introduce the system and the curation task using various modalities such as via tutorials, documentation with examples, and live demonstration of the system.

5-Evaluation

5.1 Selection and Annotation of Gold Standard

The selection of suitable data collections was inspired by real curation tasks, as well as keeping in mind the biocuration workflows. The format of the annotated corpus varied depending on the system's output.

TextPresso-This system was set up specifically for the Dictybase curation group to curate GO subcellular location. The dataset consisted of a random selection of 30 full-text articles from 2011-2012 about *Dictyostelium discoideum* which were not yet annotated in the database and contained potential GO subcellular location annotation based on keyword search for "localiz" AND/OR "enrich" AND/OR "distribut ". Curators annotated 15 of these manually and 15 using TextPresso. In addition to the curator at Dictybase, a curator from the Plant Ontology evaluated the system as well. In this activity curators had to capture information such as:1) Paper Identifier,

2) Annotation Entity, 3) Paper Section, 4) Curatable Sentence, 5) Component Term in Sentence, 6) GO Term, 7) GO ID, 8) Evidence Code.

PCS-This system was set up for curation of phenotypes from evolutionary literature about fish. Curators of the Phenoscape and Zebrafish databases were involved in the evaluation. The dataset in this case was a set of systematic characters in NeXML format randomly selected from 50 articles about evolutionary literature in fish. In this activity curators had to capture the systemic characters in the form of entity and quality terms and IDs (from ontologies). In this activity curators did the evaluation totally manually, using Phenex, the curation system currently used by Phenoscape curators, and using the PCS system (consists of Phenex+Charaparser, the text mining tool).

PubTator- This system was not optimized for any particular curation group, but rather was intended to be of utility to a wide range of end-users. The curators involved were National Library of Medicine employees and Arabidopsis Information Resource (TAIR) curators. Rat Genome Database curators also evaluated the system, although their involvement is not discussed further at this time. In the case of NLM and TAIR, the datasets were selected from documents previously curated. The documents curated by NLM were sampled from the Gene Indexing Assistant Test Collection, while the documents curated by TAIR were sampled from all publications processed by TAIR in December 2011. The information captured included gene indexing in both cases, but with major differences of scope and target. The NLM curators worked at the mention level and normalized to NCBI Gene identifiers. The TAIR curators worked at the document level and normalized to TAIR's own nomenclature. Besides gene indexing, the TAIR curators also collected document triage information—abstracts were labeled as relevant for full curation, or not relevant for full curation.

PPInterFinder- This system was not optimized to any particular curation group, but since it involves PPIs, curators from BioGrid and MINT evaluated this system. In these cases the curation scenario was protein-centric, focusing on a biologically relevant group of proteins for which it is important to annotate protein interactions and phosphorylation events, namely, human kinases. Initially a random set of human kinases were chosen. For each kinase a Boolean gene/protein query was constructed using naming information contained in UniProt as well as information derived from two systems returning protein name aliases: PubMeMiner (http://hgserver2.amc.nl/cgi-bin/miner/genefinder.cgi) and FABLE (http://fable.chop.edu). Searches using the official gene symbols against these systems were used to enrich the Boolean query. These Boolean queries were used against a system particularly developed for the triage of protein interaction relevant abstracts called PIE (http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/index.html).

The top 30 documents for each protein (with a PIE score above 0.5) were selected. Out of these a random subset of abstracts was chosen as target documents for curation. Curators were assigned to curate 25 abstracts manually and 25 using the text mining tools. Overlaps between curators was arranged to allow a comparative analysis and agreement studies.

In this activity curators had to capture information such as: PubMed ID, interactants, interaction verb, relation, and evidence sentence.

eFIP- This system provides information about PPIs of phosphorylated proteins. Curators from the pathway database Reactome, the Saccharomyces genomic database (SGD), and Merck were involved in the evaluation of this tool. The selection of the dataset and curation procedure followed the one described for PPInterfinder. The information captured included: PubMed ID, interactants indicating phosphorylated interactant, interaction verb, relation, and evidence sentence.

T-HOD: This system collects lists of genes that have proven to be relevant to three kinds of cardiovascular diseases – hypertension, obesity and diabetes, with the last disease specified as Type 1 or Type 2. It can be used to affirm the association of genes with these diseases and provides evidence for further studies. Curators from Pfizer, MGI Phenotype, GAD and Reactome were involved in the evaluation of this tool. The dataset for this system consisted of 50 abstracts from 2011 that were randomly selected based on a PubMed search that included the disease name and the keywords "associated" and "gene". Curators were each assigned to curate 25 abstracts manually and 25 using the text mining tools. In addition, curators validated the literature with the system using a gene-centric approach. The information captured included gene name, Entrez gene ID, disease, relation, and evidence sentence.

Tagtog: This system was not fully evaluated in the pre-workshop. The team was invited to participate in the poster and demonstration sessions at the workshop.

Acetylation: This system was not fully evaluated in the pre-workshop, the team was invited to participate in poster and demonstration sessions at the workshop.

5.2 Evaluation task

For the evaluation, each biocurator needed to perform the following tasks:

- Install curator logger to track time-on-task and curator web-based activities. Biocurators installed a client-side web-browser add-on to allow tracking time and user activity during testing. Users were informed as to the nature of the data being collected and asked whether they wanted to opt out of data collection when the browser opened. The generated data was sent to one of the organizers automatically once a session was complete.

- Get Trained: use examples provided by the teams to become familiarized with the assigned system. Get information about the curation guidelines for the particular task. At this stage the teams and users had fluent interaction. In many cases, users reported system bugs that were addressed before the testing.

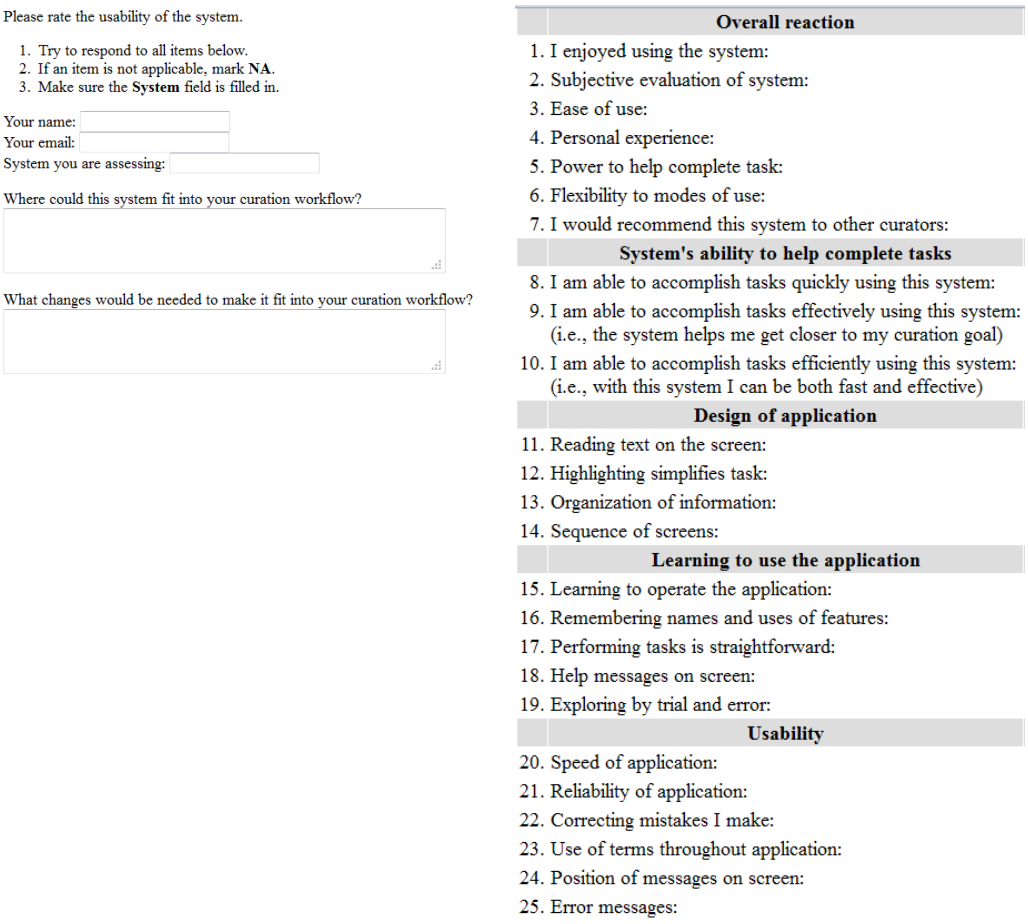
- Perform Evaluation: In all cases the testing involved manual curation of a set of documents and curation of another set using the selected system. Since the output format of the manual and system-assisted curation should be comparable, each system provided specifications for the required output in a spreadsheet. In one case curation was compared between an existing curation tool (Phenex) and curation tool plus text mining component (Phenex+Charaparser). This case reflects a more realistic scenario of a curation workflow and will provide a better sense on whether the text mining system assists in the curation.

- Complete Survey: After using a system for curation, users completed a survey in which they were asked additional questions about their experience with the system. Questions included subjective questions about overall reactions to the system, design, and if it helped completing the task (Figure 1). In addition, we asked two questions related to the biocuration workflow to address how the system would fit into the curator's workflow and the changes needed to make it happen.

After the evaluation we will report on performance and usability based on the following metrics:

Performance measure: we will compare: i) time on task for manual vs. system assisted curation using the data from the curator's logger (in addition we requested curators to time themselves independently); ii) compare annotated gold standard versus system output and report corresponding metrics (such as precision, recall, and f-measure), and iii) inter-annotator consistency. Results will be presented at the workshop.

Subjective measure: **Table 3** shows the preliminary results of the survey, expressed as average values for the set of questions in each category within the survey. In all questions, except question 11 in Design of application, the ranking corresponded to 1 being less positive/agreeable value to 7 more positive/agreeable value. So, in order to do a meaningful average value, the ranking in question 11 was reversed. The results show that there seems to be a consistency of the rating by the different curators for a given system, at least in relative terms when comparing among categories. The results also show that in many systems the main impact is in the Overall reaction of the user and the Ability to complete the task. Since the latter is expected to have a significant impact on the user general feeling about the system (not being able to complete the task, or make it hard to complete the task frustrates the user, regardless of the great design of the site), this may explain the lower rating on the Overall reaction. Most of the systems did well in the category Learning to use the system (13 out of 14 curators gave ratings higher than 4), although the help on screen was not rated by many curators. Further processing and postings of the results will be discussed at the workshop.



Please rate the usability of the system.

1. Try to respond to all items below.
2. If an item is not applicable, mark **NA**.
3. Make sure the **System** field is filled in.

Your name:
Your email:
System you are assessing:

Where could this system fit into your curation workflow?

What changes would be needed to make it fit into your curation workflow?

	Overall reaction
1.	I enjoyed using the system:
2.	Subjective evaluation of system:
3.	Ease of use:
4.	Personal experience:
5.	Power to help complete task:
6.	Flexibility to modes of use:
7.	I would recommend this system to other curators:
	System's ability to help complete tasks
8.	I am able to accomplish tasks quickly using this system:
9.	I am able to accomplish tasks effectively using this system: (i.e., the system helps me get closer to my curation goal)
10.	I am able to accomplish tasks efficiently using this system: (i.e., with this system I can be both fast and effective)
	Design of application
11.	Reading text on the screen:
12.	Highlighting simplifies task:
13.	Organization of information:
14.	Sequence of screens:
	Learning to use the application
15.	Learning to operate the application:
16.	Remembering names and uses of features:
17.	Performing tasks is straightforward:
18.	Help messages on screen:
19.	Exploring by trial and error:
	Usability
20.	Speed of application:
21.	Reliability of application:
22.	Correcting mistakes I make:
23.	Use of terms throughout application:
24.	Position of messages on screen:
25.	Error messages:

Figure 1-Snapshot of the user survey questions.

Table 3- User survey results. Results from 14 of the 17 curators that evaluated the systems. Results are shown as the average for the questions in the different sections of the survey. Scale 1 (worst) to 7 (best).

System:TextPresso	curator 1	curator 2
Overall reaction	2.8	3.8
Design of application	4.3	6.5
System's ability to help complete tasks	4.0	5.0
Learning to use the application	5.0	6.5
Usability	6.3	6.6

System:PCS	curator 1	curator 2
Overall reaction	3.5	2.3
Design of application	5.7	3.7
System's ability to help complete tasks	3.0	1.3
Learning to use the application	6.3	5.3
Usability	7.0	5.4

System:PubTator	curator 1	curator 2
Overall reaction	6.3	6.3
Design of application	5.8	4.8
System's ability to help complete tasks	6.0	6.3
Learning to use the application	5.6	6.0
Usability	6.2	6.0

System:PPInterFinder	curator 1	curator 2
Overall reaction	4.8	1.5
Design of application	5.3	3.0
System's ability to help complete tasks	1.0	1.0
Learning to use the application	7.0	3.0
Usability	5.8	2.5

System:eFIP	curator 1	curator 2
Overall reaction	4.2	5.7
Design of application	4.3	5.8
System's ability to help complete tasks	5.7	6.0
Learning to use the application	6.3	6.3
usability	6.2	6.4

System:T-HOD	curator 1	curator 2	curator 3	curator 4
Overall reaction	3.8	2.3	5.2	5.0
Design of application	4.8	4.0	5.3	4.0
System's ability to help complete tasks	3.0	2.3	6.0	3.0
Learning to use the application	4.8	4.5	5.8	5.5
Usability	4.2	2.8	6.2	4.5

6- Discussion

6.1 Some thoughts and limitations about the current approach:

Sentence vs. Document level validation: Many of the text mining systems required validation at the sentence level, whereas the curator decides at the abstract/document level; the systems would present multiple sentences for the same assertion that the curator needs to go through one by one, but in reality he/she reads the abstracts and annotates once.

What to compare?: In this "experiment" we compared manual versus system-assisted curation to have a common baseline, but we are aware that in reality this may not represent how the curator does the literature curation in their curation workflow, and therefore it will impact the performance of the curator in the manual task (time it takes to complete the task). We thought that this approach would still be informative and that with the curators' feedback we should be able to plan for a more realistic scenario for BioCreative IV. However, we have one case where curation using a current curation tool can be compared to the curation tool with text mining capabilities. We will report the outcome of this at the workshop.

Document- vs. gene-centric curation approach: In this task we proposed to have a document-centric approach for curation--given a set of documents, perform the task. But the way curation is approached depends on the database, and in many cases is gene-centric. The reason we chose a document-centric approach for Track III was because we wanted to expose the systems to a variety of examples, whereas the results from the gene-centric approach may be biased if only a few genes are tested.

Systems availability for proper testing: When planning for BioCreative IV it will be important for organizers to have access to functional systems much in advance to make sure that these are working properly. For this particular experiment, many of the systems were tuned according to the curation group that evaluated them, for example, TextPresso adapted the system for the curation of articles for DictyBase, and this included close coordination with the database to obtain the pdf articles about *Dictyostelium*, the import of gene vocabulary and other things. So even when TextPresso is fully functional and mature for Wormbase, development was needed to adapt it for DictyBase. In other cases, systems were simply newly developed and hadn't been subjected to extensive testing due to lack of time. To alleviate this we allowed users to report bugs or issues during the training period so they could be solved for the testing phase.

6.2 At the workshop

The results of this activity will be presented at the BioCreative workshop. In addition, based on the success of the demo session in BioCreative III, we extended this session to include a short usability evaluation by users. The Track III systems will demonstrate their system and curators attending the session will have the opportunity to try the system. We will collect their opinions via the user survey. We recruited a User Advisory Group (UAG) to assist in this endeavor. Each member will try two systems and all systems will be tested. Other curators present at the session

can select any system. Since the UAG will be advising on the BioCreative IV challenge planning, direct exposure to the activity is essential.

The current interactive task has been very challenging from multiple aspects: coordination, recruitment of curators that can properly evaluate the systems; selection of datasets; systems readiness; and data collection, format, and processing. However, it is a great experience for both systems and users for the systems to interact with potential users and learn about their real needs, as well as for the users to be exposed to tools that may assist them in their curation. In addition, this task provides the basis for discussion on how to design the BioCreative IV challenge.

Acknowledgements

The BioCreative 2012 workshop was supported under NSF grant DBI-0850319. We would like to thank the teams and the biocurators who participated in this activity.

References

1. Arighi, C., Roberts, P., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-aryamontri, A., Clematide, S., Gaudet, P., Giglio, M., Harrow, I., et al. (2011). BioCreative III interactive task: an overview. BMC Bioinformatics *12*, S4.

T-HOD: Text-mined Hypertension, Obesity, Diabetes Candidate Gene Database

Johnny Chi-Yang Wu[1], Hong-Jie Dai[1,3], Richard Tzong-Han Tsai[4], Wen-Harn Pan[2], Wen-Lian Hsu[1, 3*]

[1]Institute of Information Science, Acdemia Sinica, Taipei, Taiwan, R.O.C.
[2]Institute of Biomedical Sciences, Acdemia Sinica, Taipei, Taiwan, R.O.C.
[3]Dep. of Computer Science, National Tsing-Hua Univ., HsinChu, Taiwan, R.O.C.
[4]Dept. of Computer Science & Engineering, Yuan Ze Univ., Taoyuan, Taiwan, R.O.C.

[*]Corresponding author: Tel: +886-2-2788-3799 ext. 1804, E-mail: hsu@iis.sinica.edu.tw

Abstract

Text mined hypertension, obesity, diabetes candidate gene database (T-HOD) is a database developed to collect lists of genes that are associated with three kinds of cardiovascular diseases – hypertension, obesity and diabetes, with the last disease specified into Type 1 and Type 2. T-HOD employed the state-of-art text-mining technologies, including a gene mention recognition/gene normalization system and a disease-gene relation extraction system, which can be used to affirm the association of genes with the three diseases and provide more evidence for further studies. The primary inputs of T-HOD are the three kinds of diseases, and the output is a list of disease-related genes which can be ranked based on their number of appearance, protein-protein interactions and single nucleotide polymorphisms. Currently, 837, 835, and 821 candidate genes are recorded in T-HOD for hypertension, obesity and diabetes, respectively. We believe T-HOD can help life scientists in search for more disease candidate genes in a less time- and effort-consuming manner. T-HOD is available at http://bws.iis.sinica.edu.tw/THOD.

Introduction

In biomedical literature mining, an important task is to identify disease related genes as its bio-signature. Previous work such as the genetic association database (GAD)[1], which is built on the basis of manual processes, has high manpower costs owing to the difficulty of maintaining these databases with the large volume of ever-growing new literatures. In addition, the study of disease pathogenesis has become increasingly difficult because of the diverse factors involved in disease progression. In most cases, development of these diseases is modulated by the variations of multiple genes and their interactions with environmental factors. Therefore, elucidating the pathogenic mechanisms of diseases turns out to be a demanding task. To investigate the complex genetics behind diseases systematically, it is necessary to integrate the finding of both small-scale studies and high-throughput researches. However, there are only a few databases and review papers that compile HOD related genes from the literature.

In the field of diabetes, T1Dbase [2] integrates valuable information on candidate genes from several databases for type 1 diabetes, while T2D-Db [3] compiles from PubMed human, mouse and rat genes involved in the pathogenesis of type 2 diabetes. For obesity genetics, the review paper "The Human Obesity Gene Map: The 2005 Update" [4] lists candidate genes and/or potential loci up until the end of 2005. For hypertension genes, GAD lists hundreds of hypertension candidate genes along with genes for several other diseases. All of the above mentioned resources were compiled manually. However, due to limited human resources, such databases cannot always be kept up-to-date. In recent years, various groups have proposed using automated text mining approaches to reduce human effort in constructing and updating such databases [5-9]. SNPs3D [8] and PubMeth [5] are two such databases constructed using text-mining approaches coupled with manual review and annotation steps. SNPs3D compiles candidate genes and single nucleotides polymorphism (SNP) sites related to cancers, neurodegenerative diseases and metabolic syndromes. PubMeth contains information on DNA methylation for several cancers. These two databases extract gene names that have a high co-occurrence with the target diseases. However, using the co-occurrence-based approach alone tends to yield a huge number of false-positive relations because of the lack of syntactic and semantic analysis.

Our database, T-HOD employed the state-of-art text-mining technologies we recently developed, including a gene mention recognition/normalization (GN) system [10-12] and a disease-gene relation extraction system [13]. Since gene names vary a great deal, different genes may contain the same name. Moreover, gene names may be ambiguous and easily confused with terms employed in other research fields. Our GN system was designed to alleviate the above problems, which was used to recognize gene terms and normalize them to their corresponding Entrez Gene IDs. In addition, we recently achieved promising results in extracting hypertension-related genes [13]. We extended and optimized the above systems to extract HOD genes in our T-HOD database.

T-HOD Interface and Implementation

As shown in Figure 1, the interface of T-HOD is divided into four regions. We will elucidate the function of each region in the following section, respectively.

Region 1: Control bar

Region 1 at the top of the frame contains a pull-down display menu. By clicking on the menu, users can select the disease of interest (Hypertension, Obesity, or Type 1/2 diabetes). Users can also decide whether to show specific gene information or use our advanced search function in this region.

Region 2: Candidate Gene list

After disease selection, Region 2 shows a list of curated candidate genes. Along each candidate gene, the list also displays the number of papers containing evidence sentences, as well as the number of SNPs and number of PPIs in separate columns. The list can be sorted by clicking on the column header, and it is accessible by hitting the “download” button at the bottom.

Region 3: Viewers

Region 3 provides several viewers, including sentence viewer, network viewer, advanced search option tabs, and statistics viewer. Users can switch between different viewers by clicking on the upper tags in this region.

Sentence Viewer: The sentence viewer provides curated evidence sentences for each selected candidate gene. If the candidate genes possess corresponding SNP information, the SNP evidence sentence would also be shown below the candidate gene evidence sentences. For each evidence sentence, the sentence viewer shows the source article’s PMID and year of publication with highlighted gene and disease terms. Display of the system can be adjusted by changing the font size of the texts. And in respect of valuable feedbacks, we constructed a user friendly interface for users to express their thoughts. In addition, for those who are interested in our database and plan to adopt its use in other studies, the information of T-HOD is attainable by hitting the “download” button below the gene list and supporting sentences, allowing them to acquire the disease-related genes and their supporting proof, respectively.

Network Viewer: Figure 2 shows the network viewer that presents a graphic-based gene-gene network for a selected candidate gene. For each selected candidate gene, the viewer integrates the corresponding PPI information recorded in the Human Protein Reference Database HPRD [14] to illustrate the gene-gene network. It allows users to discover the relations among extracted candidate genes. The blue node at the top of the window represents the gene that the user chose in Region 2. To cross examine the candidate genes, the user can double click on the nodes of other candidate genes shown in the same network. Accordingly, the network viewer will redraw the network graph based on the selected gene so that the user can navigate the database more smoothly.

Advanced Search: The advanced search option tab provides advanced search options that allow users to narrow down and specify the desired search results by the following items: publication date, Entrez Gene ID, gene name, and PubMed ID.

Statistics Viewer: The number of candidate genes and candidate SNP sites contained in T-HOD are summarized in the viewer. The statistics viewer also plots the number of candidate genes and the number of new candidate genes each year in bar charts as shown in Figure 3.

Region 4: Gene and SNP information

For each selected candidate gene, the information integrated from different resources is shown in Region 4. In this region, we integrate the following information from Entrez Gene and SNP database: the gene's official symbol, Entrez Gene ID, full name, synonyms and function summary. Users can also link to the corresponding database for further information.

Text mining-based Database Curation

Figure 4 shows the flowchart for constructing the T-HOD database. It is comprised of three stages: (1) Dataset Collection and Pre-processing, (2) Candidate Gene Extraction, and (3) Content Verification. We collected abstracts on HOD from PubMed, and used text mining techniques to extract HOD candidate genes and SNPs from them. The T-HOD curators verify the extracted list and curate the knowledge into the T-HOD database. In the following sub-sections, we will describe each stage in detail.

Stage 1: Dataset Collection and Preprocessing

In this stage, we collect HOD-related abstracts from PubMed and filter out those that are non-genetic. The filtered dataset are then pre-processed by several text mining components. After pre-processing, the genetic-related abstracts were split into sentences associated with section tags by using our section categorization component [15], such as "Results" and "Conclusion", which indicate their corresponding sections.

Stage 2: Candidate Genes Extraction

In Stage 2, we extract HOD-related candidate genes from the pre-processed dataset through the following steps. First, we employ a disease named entity recognition (NER) system to recognize disease terms in a sentence. Second, a GN system is used to recognize and normalize mentioned genes to their corresponding Entrez Gene IDs. Based on the results of the previous steps, if a disease term and a gene are present in the same sentence, they are extracted as a disease-gene (D-G) candidate pair. Finally, the D-G relation extraction system determines whether a relation indeed exists within this D-G pair.

Stage 3: Manual curation

While the employed text mining components have shown satisfactory scores (*cf.* Table 1), the text mined candidate genes are examined by all T-HOD curators in Stage 3 to further ensure the quality of the curated content. In this stage, newly extracted candidate genes and their corresponding evidence sentences and abstracts are presented to the T-HOD curators. T-HOD curators review each extracted candidate gene and remove the incorrect results. Currently, the curation process has only been done on abstracts before 2011. Because all annotated error cases are recorded to our SQL database, we can also use such data to modify our text mining components efficiently.

Proposed Task for Biocuration

For Track III–interactive text mining and user evaluation task of the BioCreative 2012 Workshop, we propose the following task for biocuration.

HOD Curation Task

When given a set of abstracts (compiled from those published in 2011) related to a specific disease, a bio-curator should:

1. Identify whether the abstracts contain disease-related gene information (curatable abstracts).
2. As for curatable abstracts, extract the following information: PMID of the abstract, gene terms and its corresponding gene ID from Entrez Gene, disease terms, relation assertion (positive or negative), and the evidence sentence containing the gene-disease pair.

Figure 5 shows the formal task descriptions provided to bio-curators. Note that since the abstract set used for the HOD curation task is publications from 2011, it is not yet verified by our T-HOD curators.

The bio-curator then compares their manually curated results with the text-mined results processed by T-HOD. For the convenience of bio-curators in analyzing the results, we developed an interface that directly provides the information of T-HOD in the desired output format with additional PubMed and Entrez Gene links. These results are also available for download. Furthermore, this interface is able of notifying the curators when an abstract is not found in our database, or it does not contain any relations of interest. An example of the interface output is shown in Figure 6. This interface is available at http://bws.iis.sinica.edu.tw:8080/THOD/request_sentence_list.

Results and Discussion

Evaluation of a candidate gene database is difficult. Different standards and perspectives can produce different results. In our previous work [13], the employed D-G extraction system has shown satisfactory area-under-curve (AUC) scores of 81.4% and 83% for hypertension and diabetes, respectively. In this work, we compared the performance of T-HOD with the contents of GAD. The bench marking results are shown in Table 1. Disease-related literatures that exist in both databases were chosen for evaluation. Performance for the identification of gene-disease relations in hypertension, obesity and type 2 diabetes documents all achieved a score around 75%. In contrast, relations of type 1 diabetes only obtained a score around 70%.

There are several possible reasons that may result in the difference between T-HOD and GAD results. In order for curating a gene-disease relation into our T-HOD, the identity of both the candidate gene and disease terms must be normalized. In the current implementation, gene terms are normalized by a collective entity disambiguation method [16] to its corresponding Entrez Gene ID, while disease terms are recognized through a list of vocabularies. Error in the normalization of genes and the imperfect list of disease terms utilized may lead to the loss of relations that are present within documents. In addition, the difficulty of extract D-G relation will increase when one or both the disease and gene are expressed with an anaphoric expression. Furthermore, T-HOD only recognizes D-G relations within the same sentence. Cross sentence relations are currently not available, but it is a topic worth studying in the future. Cooper and

Kershenbaum [17] has identified co-reference as one of the reasons for the decreasing recall in biomedical relation extraction task. Finally, determining negations within a sentence is also an important issue. The present T-HOD system can only deal with negation descriptions in basic phrase structures, which may be insufficient in distinguishing more complex negation narratives.

Funding

This work was supported by the National Core Facility Program for Biotechnology, Taiwan (Bioinformatics Consortium of Taiwan, NSC100-2319-B-010-002).

References

1. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nature Genetics* 2004, **36**(5):431-432.
2. Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, Burren OS, Cassen VM, Cavnor CC, Dolman GE, Flamez D *et al*: **T1DBase: integration and presentation of complex data for type 1 diabetes research**. *Nucleic Acids Res* 2007, **35**(Database issue):D742-746.
3. Agrawal S, Dimitrova N, Nathan P, Udayakumar K, Lakshmi SS, Sriram S, Manjusha N, Sengupta U: **T2D-Db: an integrated platform to study the molecular basis of Type 2 diabetes**. *BMC Genomics* 2008, **9**:320.
4. Rankinen T, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, Pérusse L, Bouchard C: **The Human Obesity Gene Map: The 2005 Update**. *Obseity* 2006, **14**:529-644.
5. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W: **PubMeth: a cancer methylation database combining text-mining and expert annotation**. *Nucleic Acids Res* 2008, **36**(Database issue):D842-846.
6. Hahn U, Wermter J, Blasczyk R, Horn PA: **Text mining: powering the database revolution**. *Nature* 2007, **448**(7150):130.
7. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining**. *BMC Complement Altern Med* 2008, **8**:58.
8. Yue P, Melamud E, Moult J: **SNPs3D: Candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:-.
9. Fang YC, Lai PT, Dai HJ, Hsu WL: **MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature**. *BMC Bioinformatics* 2011, **12**(1):471.
10. Dai H-J, Lai P-T, Tsai RT-H: **Multistage 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.
11. 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.
12. Dai H-J, Chang Y-C, Tsai RT-H, Hsu W-L: **Integration of gene normalization stages and co-reference resolution using a Markov logic network**. *Bioinformatics* 2011, **27**(18):2586-2594.
13. Tsai RT-H, Lai P-T, Dai H-J, Huang C-H, Bow Y-Y, Chang Y-C, Pan W-H, Hsu W-L: **HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features**. *BMC Bioinformatics* 2009, **10**(Suppl 15):S9.
14. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

15. Lin RTK, Dai H-J, Bow Y-Y, Chiu JL-T, Tsai RT-H: **Using conditional random fields for result identification in biomedical abstracts** *Integrated Computer-Aided Engineering* 2009, **16**(4):339-352.
16. Dai H-J, Tsai RT-H, Hsu* W-L: **Entity Disambiguation Using a Markov-Logic Network**. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP); Chiang Mai, Thailand.* 2011: 846-855.
17. Cooper JW, Kershenbaum A: **Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information**. *BMC Bioinformatics* 2005, **6**(1):143.

Table 1. Comparison of candidates genes in T-HOD with GAD.

	True Positive	False Positive	False Negative	Precision	Recall	F score	Number of Documents
Hypertension	165	42	49	0.797	0.771	0.784	150
Obesity	105	35	29	0.75	0.784	0.766	115
Type 1 Diabetes	60	24	27	0.714	0.69	0.702	73
Type 2 Diabetes	127	35	37	0.784	0.774	0.779	140
Overall	457	136	142	0.771	0.763	0.767	608

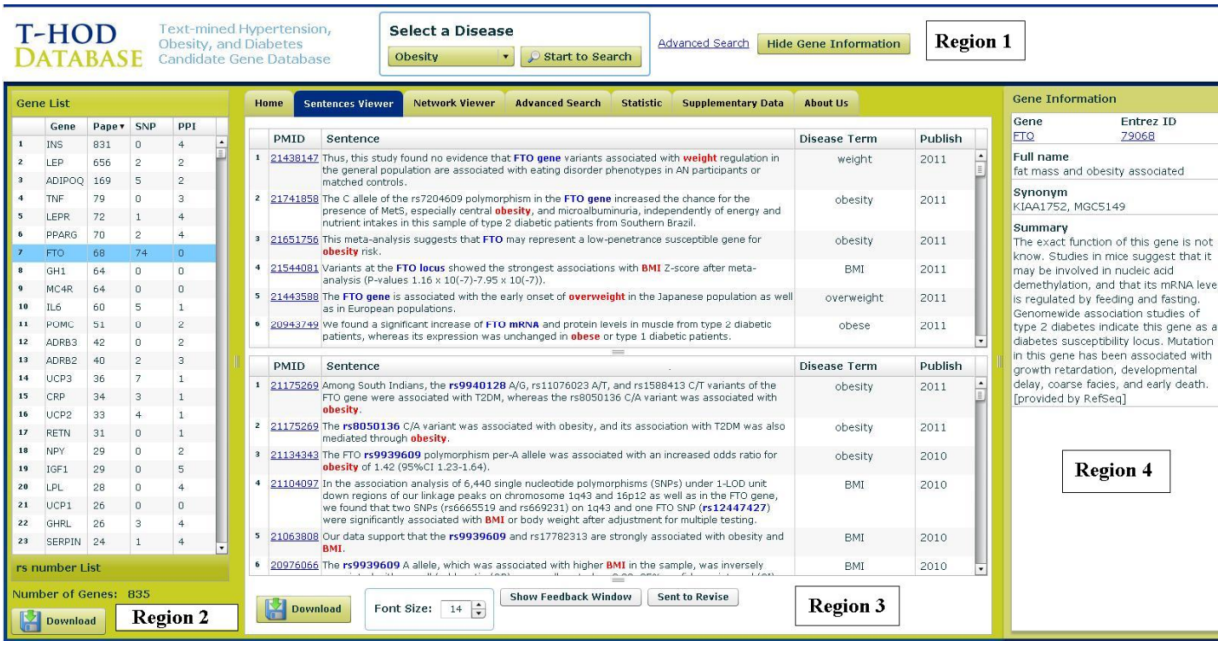


Figure 1. User interface of the T-HOD database. The user interface is divided into four regions for precise introduction.

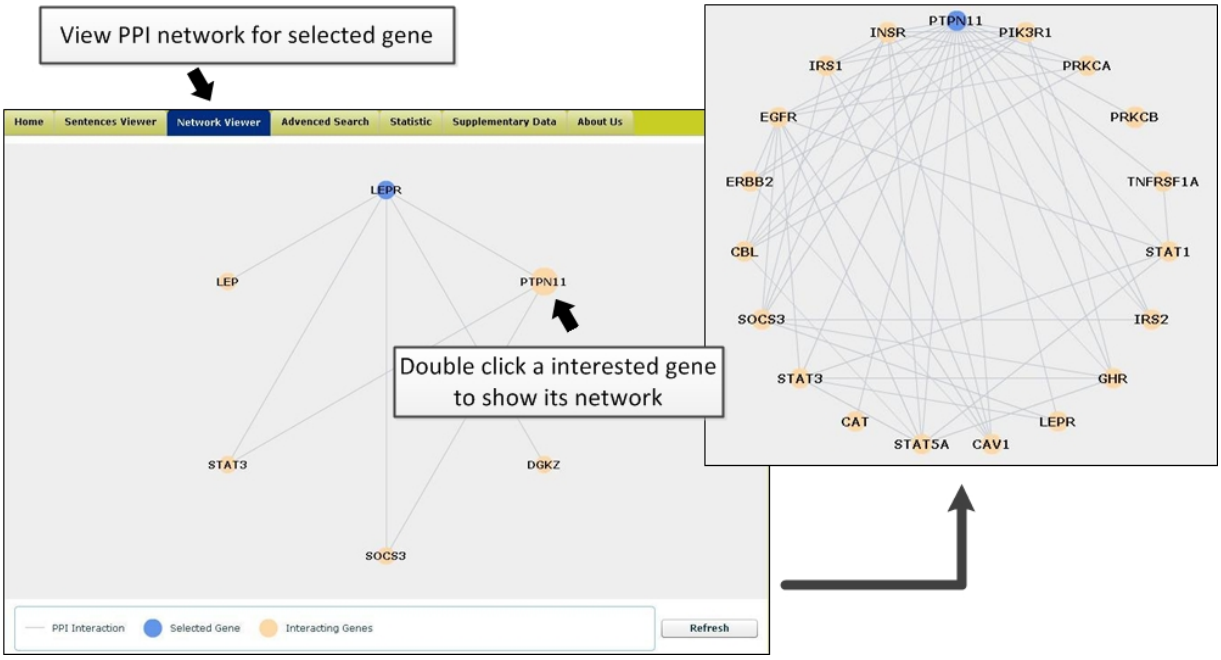


Figure 2. The network viewer of the T-HOD database.

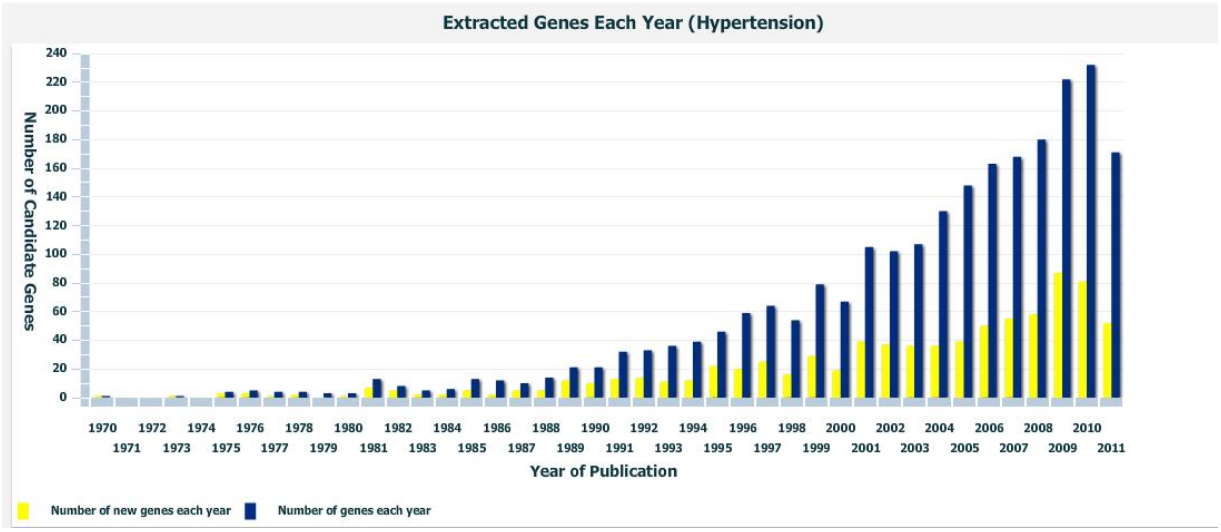


Figure 3. Statistics of the extracted hypertension candidate genes. The blue bars indicate the number of genes extracted each year, while the yellow bars specify the number of novel genes discovered each year.

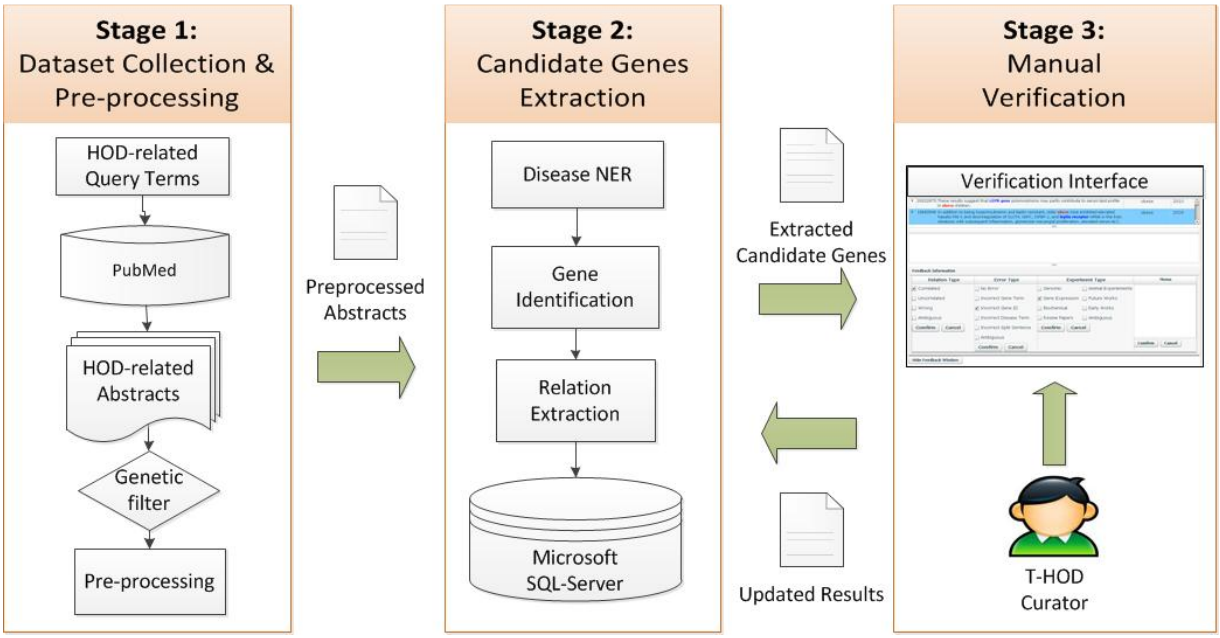


Figure 4. The flowchart of T-HOD database construction.

Manual Task: Curators will be given a list of PubMed abstracts for further processing, and should provide an output spreadsheet that contains the information of interest.

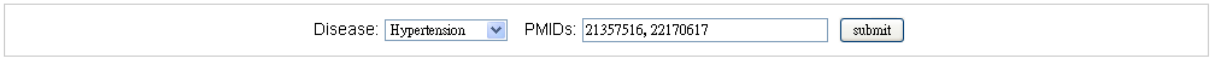
Using T-HOD: Curators will compare the information retrieved by T-HOD regarding the given set of abstracts with those that are extracted manually, analyze their differences and offer any suggestions for further improvement.

Input: Assigned set of specific disease-related abstracts.

Output: Output of the extracted information should be presented accordingly to the following format:

PMID | Gene ID | Gene Term | Disease term | Evidence sentence

Figure 5. Illustration of the proposed task for bio-curators.



PMID	Gene ID	Gene Term	Disease Term	Sentence
21357516	183	angiotensin II	hypertension	At 34 weeks of age Imai rats showed heavy proteinuria, hypoalbuminemia, **hypertension**, azotemia, glomerulosclerosis, tubulointerstitial inflammation, increased **angiotensin II** expressing cell population, up-regulations of AT1 receptor, AT2 receptor, NAD(P)H oxidase, and inflammatory mediators, activation of nuclear factor-kappa-B and reduction of Nrf2 activity and expression of its downstream gene products in the renal cortex.
22170617	183	angiotensin II	high blood pressure	In particular, we describe a new transgenic mouse model which demonstrates that intracellular **angiotensin II** is linked to **high blood pressure**.

Download

Figure 6. The interface for bio-curators.

Textpresso text mining: semi-automated curation of protein subcellular localization using the Gene Ontology's Cellular Component Ontology

Kimberly Van Auken[1], Yuling Li[1], Juancarlos Chan[1], Petra Fey[2], Robert J. Dodson[2], Arun Rangarajan[1], Rex L. Chisholm[2], Paul W. Sternberg[1,3], and Hans-Michael Muller[1,*]

[1]Division of Biology, California Institute of Technology, Pasadena, CA, [2]Center for Genetic Medicine, Northwestern University, Chicago, IL, [3]Howard Hughes Medical Institute, Pasadena, CA

*Corresponding author: Tel: (626) 395-8615, E-mail: mueller@caltech.edu

Abstract

Manual curation of experimental data from the biomedical literature is expensive and time-consuming; however, most biological knowledge bases still rely heavily on manual curation for data extraction and entry. We have developed and actively use a category-based information retrieval and extraction system for curating *C. elegans* proteins to the Gene Ontology's Cellular Component Ontology. The system's core components are the Textpresso full text database and index, a support vector machine (SVM) classifier for the presence of expression pattern data in a paper as well as a specifically designed semantic category search. All automatically extracted information is presented to a curator for validation in a user-friendly web-based interface. For this evaluation we will use articles about the organism *Dictyostelium discoideum* to test the precision and recall metrics for the corpus of an organism other than *C. elegans* and determine to what extent the system increases curation efficiency for dictyBase annotators.

System Description

The method is successful because authors describe subcellular localization in a sufficiently stereotypical manner. Stereotypical language can be used to empirically create new Textpresso categories specific for retrieval of sentences relevant to GO Cellular Component curation. Details about this method can be found in (1).

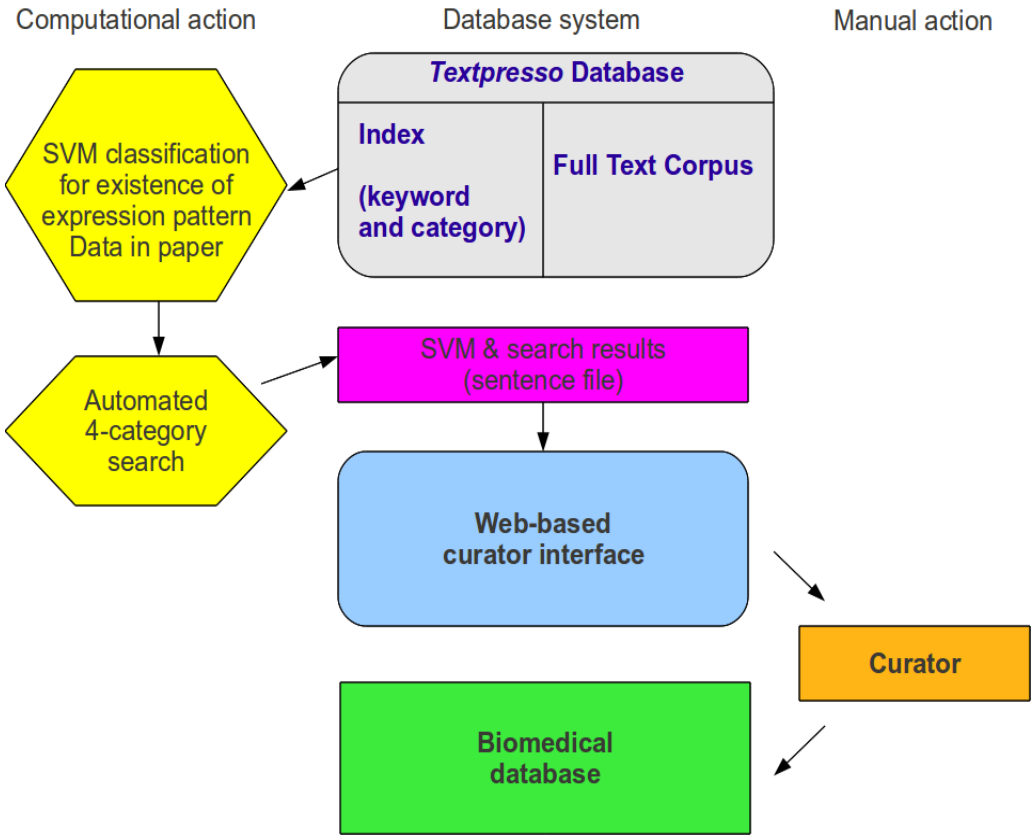


Figure 1: Schema of the curation system for GO Cellular Component Curation.

Figure 1 outlines the workflow of the curation system as it currently applies to the *C. elegans* corpus. The Textpresso database (2) is updated every night with the bibliography and full text of new papers pertaining to *C. elegans* research. They are subsequently classified by an SVM to find papers containing data related to expression patterns. This SVM was trained by a set of ~1000 *C. elegans* papers containing expression pattern data. Papers predicted by the SVM to contain this data type are then subjected to a specialized Textpresso category query. Using a training set of sentences that describe results of localization experiments in the published literature, we generated three new curation task-specific categories (cellular components, assay terms, and verbs) containing words and phrases associated with reports of experimentally determined subcellular localization. The Textpresso query searches the full text of the pre-filtered articles for sentences containing terms from each of the three new categories plus the name of a *C. elegans* protein. The results of this query are then stored in a file that contains the sentences that match the query, the paper ID and matched query items (protein, cellular component, assay term and verb). The curator can access the results stored in the file via a web-based curation interface, which is displayed in Figure 2. The curation form allows for inspection of the retrieved sentences and pre-populates two data fields with the protein name and component term, as extracted from the sentence. If a similar annotation using the extracted component term has been made in the past, the form suggests GO annotation terms based on what GO terms have previously been associated with the component term in the sentence. For each sentence displayed in the curation form, the curator can take several steps, from adding an annotation to the biomedical database to marking a falsely made extraction as "false positive", as

"not go-curatable" (e.g. the sentence describes localization in a mutant background), as a "scrambled sentence" (an artifact from erroneous pdf to text conversion) or as "already curated", a classification helpful for marking sentences that describe localization of commonly used organelle-specific markers. This latter classification allows curators to eliminate those markers for subsequent curation.

Relevance and Impact

The system is currently used by WormBase for GO cellular component curation. An automated pipeline sends the curator an e-mail if new sentences have been identified by SVM and the 4-category Textpresso query. Other groups using this approach include The Arabidopsis Information Resource (TAIR), which uses the Textpresso category searches, but not the SVM step, on a corpus that is updated every 6 months, and plans are underway to implement GO cellular component curation for FlyBase and dictyBase (see below). WormBase has also extended this general approach to mining macromolecular interactions as well as finding orthologs of human disease genes.

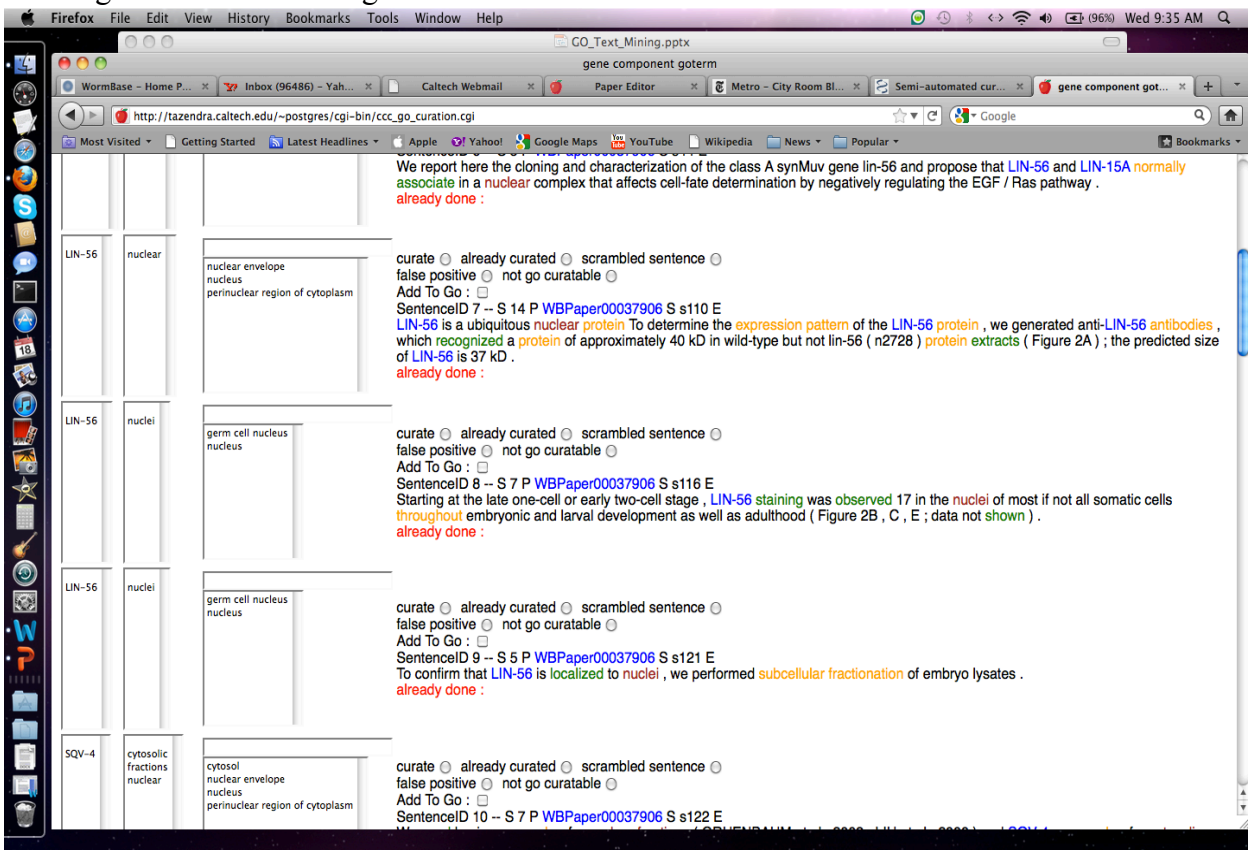


Figure 2: Curation interface to validate and extract GO Cellular Component Curation.

Adaptability

As the system is already used by other model organism databases and for other data types, it has proven its adaptability. The biggest challenge in adapting the curation workflow lies in the fact that the specific Textpresso categories may need to be modified to retrieve sentences of interest for (a) a new model organism or (b) a different data type. For example, plant-specific cellular component terms needed to be added to the Textpresso component category for effective searching of the Arabidopsis literature. New training sentences need to be retrieved manually and analyzed for word and phrase frequency so meaningful categories and corresponding lexica can be formed.

Interactivity

All curator activities are performed via a web-interface. Should further information be required to make a curatorial decision, the Textpresso search engine can be accessed via an interface to post additional keyword and/or category queries to the full text of all or specific papers. All other steps in the pipeline are automated via cronjobs, but could theoretically also be controlled via a simple web-interface.

Performance

Prior to the introduction of the SVM filtering step, we evaluated the *C. elegans* system on three levels; on the document level, the Textpresso-only system yields a recall of 95.2% and a precision of 66.7%. In this case, a true positive is a paper that contains a sentence describing subcellular localization of a protein, although the machine may or may not have picked the correct sentence for curation. When SVM is included in the process, recall drops to 76%, but precision increases to 80%. When evaluating the system on a sentence level (i.e., a true positive is a correctly identified sentence describing subcellular localization), the precision is 80.1%, but recall is only 30%. However, this recall is rescued to 66.2% when looking at the annotation level, i.e., how many of all possible annotations could be made from a set of papers. This is because the same information gets repeated across a paper or a corpus. There are no data yet on how the SVM step changes recall and precision on a sentence level.

Benchmark

For the BioCreative workshop, we chose to evaluate the system for the *Dictyostelium discoideum* corpus. One dictyBase curator provided a set of 30 documents possibly containing protein subcellular localization information in *Dictyostelium discoideum*. This curator will curate sentences and annotations from all 30 documents to generate a gold-standard annotation set. The 30 documents will then be split up into two subsets of 15 documents each. The first subset will be manually curated by biocurators who will record, in a spreadsheet, all sentences from the document relevant to Cellular Component annotation as well as the GO annotations they would make. Curators will measure the time it takes to capture the sentence information and make the appropriate annotations. For the second subset, we will evaluate the results of Textpresso searches provided in the same curation interface that WormBase curators use, except that it will be loaded with sentences that have been retrieved via the Textpresso search of this subset. For

these Textpresso searches we will compare the results of the 4-category search described above to that of a 5-category search that additionally includes a category containing words and variations for the terms 'Figure' and 'Table'. Such words, when present in a sentence, further refine search results to only those that describe experimental findings in a paper. The curator will validate or reject the computational output with the help of the interface and again, record the time it takes to do this. Together with recall and precision of the output of this interactive step we will be able to compare recall, precision and curation efficiency (in terms of time) with the purely manual curation of the first subset. In all cases the required output will be the paper IDs of all papers containing protein subcellular localization, the sentences of each paper from which an annotation can be made as well as the annotation itself.

Funding

This work was supported by the National Human Genome Research Institute at the National Institute of Health, grant # HG004090 and # HG02223.

References

1. Van Auken K, Jaffery J, Chan J, Müller HM, Sternberg PW. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*. 2009 Jul 21;10:228.

2. Muller HM, Kenny EE, Sternberg PW. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004 Nov;2(11):e309. Epub 2004 Sep 21.

PCS for Phylogenetic Systematic Literature Curation

Hong Cui[1,*], Jim Balhoff[2, 4], Wasila Dahdul[3], Hilmar Lapp[2, 5], Paula Mabee[3], Todd Vision[2, 4]
Zilong Chang[1]

[1]University of Arizona, [2]NESCent, [3]University of South Dakota, [4]University of North Carolina at Chapel Hill, [5]Duke University

*Corresponding author: Tel: 520-6214026, E-mail: hongcui@email.arizona.edu

Background

The Phenoscape project (http://www.phenoscape.org) seeks to extract systematic character data from the evolutionary literature. Systematic characters consist of two or more character states contrasting some aspect of phenotype, such as anatomy or behavior, of the taxa under study. The original purpose of these data was typically to recover phylogenetic relationships in the absence of, or together with, molecular data. However, such data have considerable value also for studying patterns of phenotypic evolution in a comparative context, *e.g.* on a given phylogeny, and linking data on natural phenotypic diversity with data on the roles of genes in the development of phenotypes in model organisms.

Phenex [1] is an interactive platform-independent desktop application designed to facilitate effective and consistent annotation of such data. It has been used in the Phenoscape project for several years and proven to be effective and user-friendly in supporting a manual annotation workflow. In this workflow [2], curators are required to parse the free text and search a set of ontologies to select the appropriate terms to annotate the characters using the Entity–Quality (EQ) model [3] (Table 1). If there are terms that are not covered by existing ontologies, the curators must wait for new terms to be added to them before they can complete the curation task. These time-consuming steps in the curation workflow prevent the efficient scaling of curation to increased phenotypic diversity.

CharaParser [4] is a text mining system that semi-automatically identifies candidate entity and quality terms in free-text character narratives and generates candidate EQ expressions for review by the human curator. The Phenoscape NLP work group (i.e., this team) is currently working to integrate CharaPaser into Phenex to produce a complete system with improved efficiency while retaining the rich and user-friendly features of the original Phenex system. For the time being, we shall call the complete system **Phenoscape Curation System (PCS)**.

The input to PCS is a set of articles from the phylogenetic systematic literature. Within each article is a semi-structured narrative of multiple systematic characters similar to those shown in Table 1. The output is a list of EQ statements that represent the original systematic characters. An EQ statement associates an entity term drawn from an organism-specific anatomy or process ontology such as the Teleost Anatomy Ontology (TAO: http://obofoundry.org/cgi-bin/detail.cgi?id=teleost_anatomy) and Gene Ontology Biological Processes (http://obofoundry.org/cgi-bin/detail.cgi?id=biological_process), with a quality term from the

PATO ontology (http://obofoundry.org/wiki/index.php/PATO:Main_Page). Table 1 shows three examples of original systematic character narratives (source: [5]) and their corresponding EQ statements created by human curators. Note that due to the limited coverage of existing ontologies, a precise E or Q term may not be found. In these cases, a broader term has to be used in order to link the narrative to an ontology (terms with * in Table 1 are examples of such cases, terms in "[]" are the more precise and more desirable terms that have yet to be added to the respective ontologies). Note also that there may be one or more EQ statements for each systematic character and some Es in the table are composed with multiple entity terms connected by *part_of* relation.

Table 1: Examples of Systematic Character Narratives and EQ Statements

Systematic Character	**EQ using terms**		**EQ using term IDs**	
	Entity	**Quality**	**Entity**	**Quality**
First dorsal-fin rays (1) deeply branched (2) unbranched	dorsal fin lepidotrichium	branched	TAO:0001418	PATO:0000402
	dorsal fin lepidotrichium dorsal fin lepidotrichium	Unbranched	TAO:0001418	PATO:0000414
Inner dentary tooth row (1) absent (2) present	dentary tooth row* [inner dentary tooth row]	count* [absent]	TAO:0001952	PATO:0000070
	dentary tooth row* [inner dentary tooth row]	count* [present]	TAO:0001952	PATO:0000070
Dentary (1) lower surface of dentary posterior to symphysis without any conspicuous notch (2) a notch along lower border of the dentary just posterior to the convoluted symphysis	ventral margin and (part_of some dentary) and (posterior_to some omandibular symphysis)	shape* [not notched]	BSPO:0000684 ⊓ (OBO_REL:part_of ∃ TAO:0000191) ⊓ (BSPO:0000099 ∃TAO:0001851)	PATO:0000052
	ventral margin and (part_of some dentary) and (posterior_to some mandibular symphysis)	notched	BSPO:0000684 ⊓ (OBO_REL:part_of ∃ TAO:0000191) ⊓ (BSPO:0000099 ∃ TAO:0001851)	PATO:0001495

PCS System Schematic Diagram

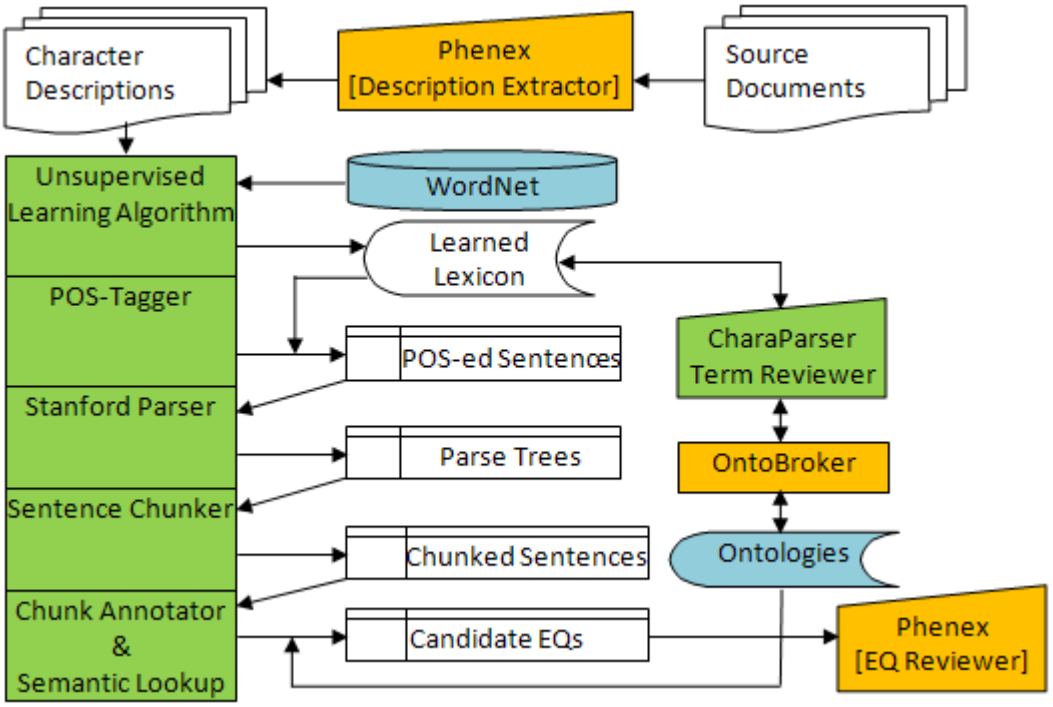


Figure 1. PCS System Schematic Diagram. Orange-colored components are part of Phenex, green-colored components are part of CharaParser, blue-colored parts are existing knowledge resources, and the remaining are the source or the intermediate results produced by the system. Trapezoidal components indicate modules that are interactive with users.

Figure 1 depicts the components of the PCS and related workflow, which involves the following steps:

1. Research assistants extract character matrices and narratives, together with other useful information (such as the identity of specimens examined), from source documents and record them in NeXML format (http://www.nexml.org) using Phenex.

2. CharaParser runs an unsupervised learning algorithm to collect entity terms and quality terms from character narratives.

3. Collected terms are reviewed by a human curator using CharaParser (Figure 2).

4. Entity and quality terms that are not in ontologies are submitted to target ontologies through an OntoBroker service that immediately provides provisional IDs.

5. Using the ontologies, including provisional terms, CharaParser executes a series of algorithms to produce candidate EQ statements in a csv table.

6. The csv table is fed to Phenex. Candidate EQ statements are reviewed by a human curator using the Phenex interface (Figure 3). Phenex supports ontology lookup by the curator if the term ID applied by the automated process is wrong (Figure 4).

7. User feedback from the EQ Review step is captured and used on the fly to correct errors generated by the automated process.

The OntoBroker (step 4) and user feedback (step 7) modules are currently not functional. PCS may be used in an interactive mode where an individual document is curated or a batch mode where hundreds of documents are curated.

PCS System Screen Shots

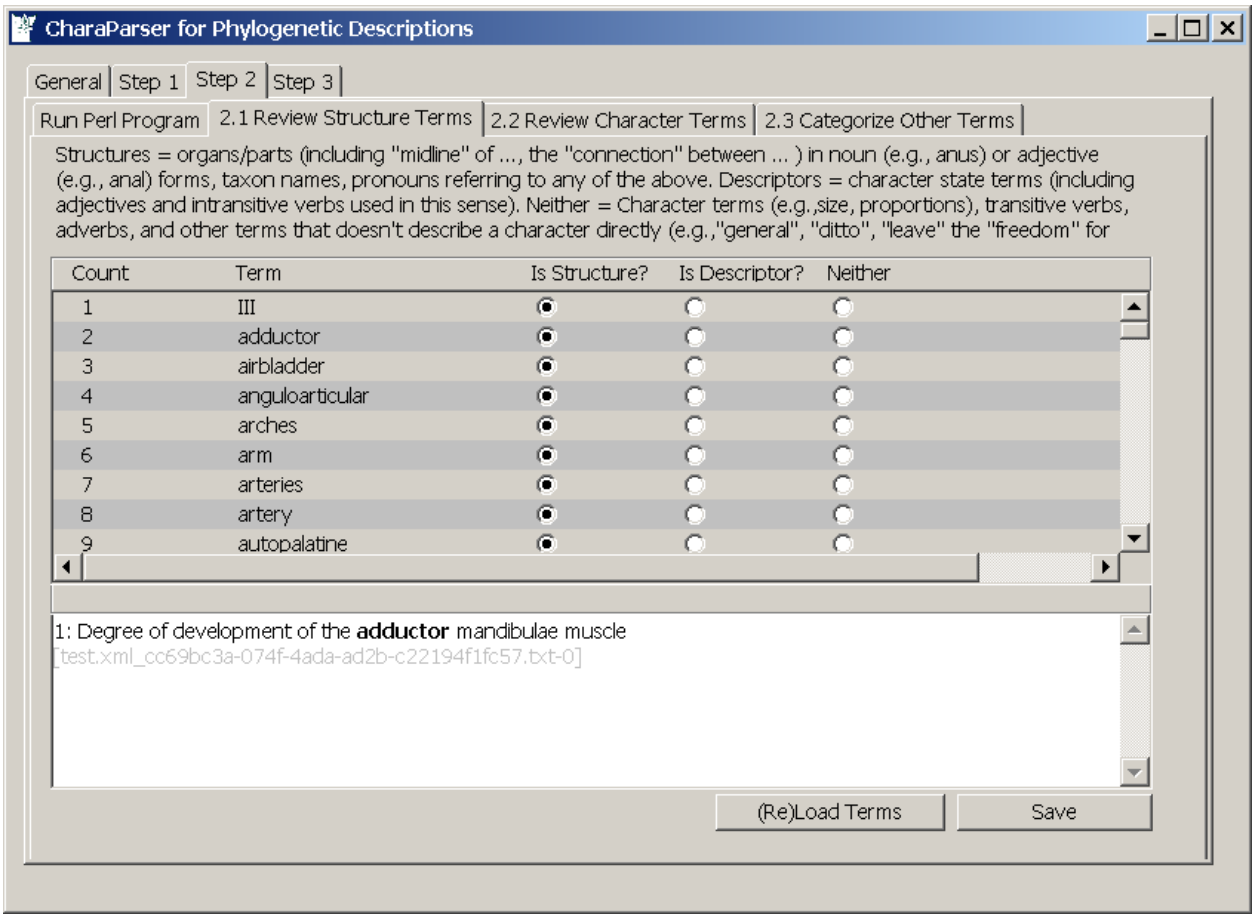


Figure 2. CharaParser Term Reviewer Module. Here entity (i.e. “structure”) terms identified by CharaParser are listed for the curator to review. The curator may assign a term to a different category (e.g., descriptor or neither) if that term is not an entity term. The original context in which a term appears is displayed in the lower box when the curator clicks anywhere on the row the term sits. On “2.2” and “2.3” tabs, the curator may review quality (i.e., “character”) terms or categorize additional entity/quality terms that the system failed to categorize.

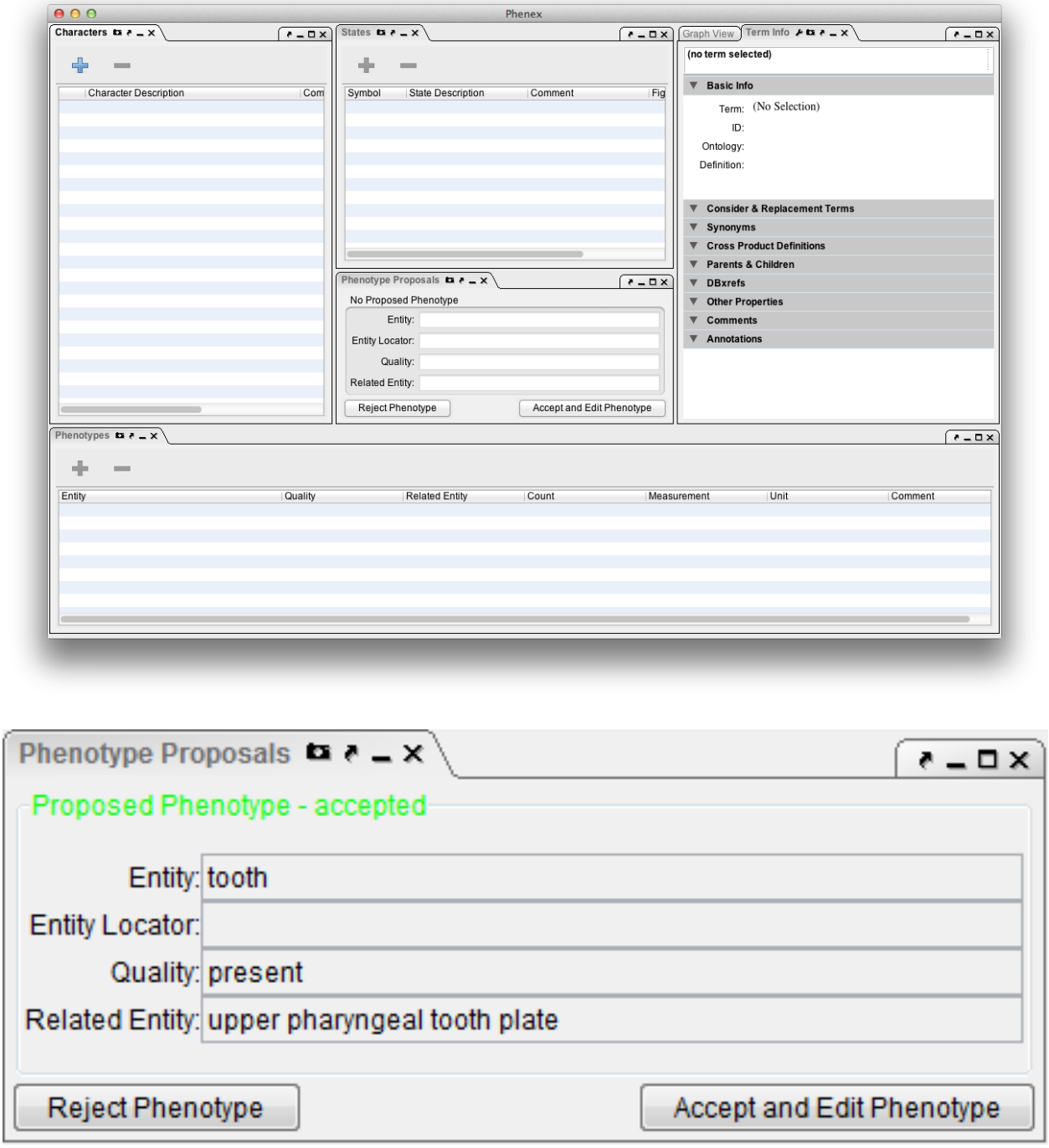


Figure 3. Phenex EQ Statement Reviewer Interface. The “Characters” and “States” panels will be loaded with original character and state descriptions from a source document. These provide the contextual information about a character/state and are used by the curator to evaluate the proposed EQ statement (shown in “phenotype proposals” section). The curator can reject the proposal or accept and edit it.

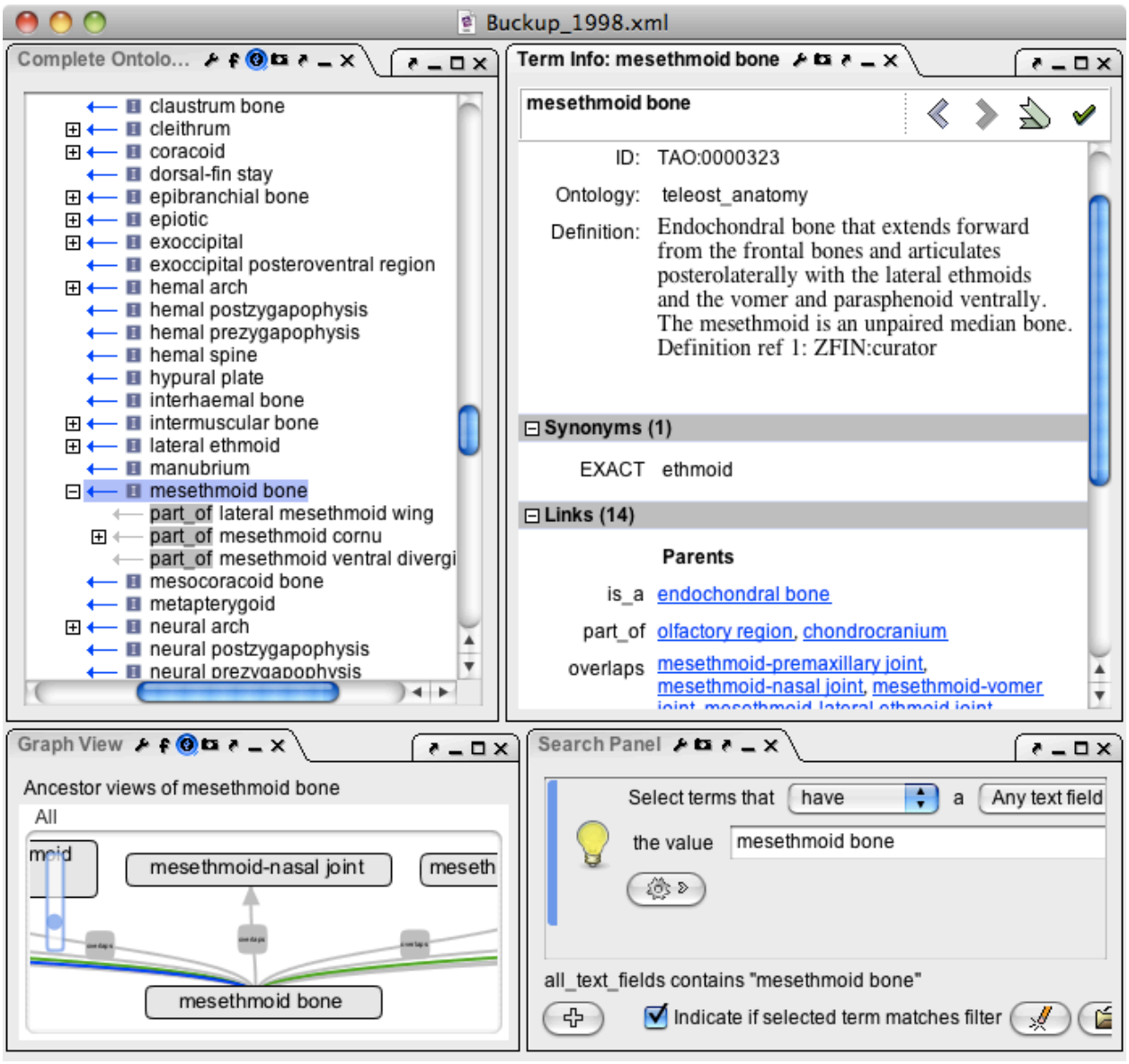


Figure 4. Phenex Ontology Look-up Interface. This interface allows the curator to search for a matching term in an ontology, for example, TAO. Besides the search panel, the interface can be configured to show the tree structure of a complete ontology, graph views of a term, and the detailed definition information of a term.

System Adaptability and Interactivity

The adaptability and interactivity of PCS is due to the complementary features of Phenex and CharaParser. Both tools are desktop applications, and we will be happy to install the system for any curator to use. The GUI of Phenex has been optimized for evolutionary biologists who are accustomed to working with lists of taxa, character narratives, and taxon-by-character matrices. It can be configured to load terms from any OBO ontology so that it can be applied to data curation for any taxonomic groups as long as the appropriate anatomy, taxonomy, and phenotype

ontologies exist. Similarly, CharaParser uses unsupervised learning methods to adapt a general-purpose syntactic parser for semantic markup of morphological narratives of any taxonomic groups. CharaParser has been evaluated and used to perform fine-grained semantic markup of morphological narratives of plants, ants, fish, and invertebrate fossils. Once completely integrated, PCS will have a look and feel consistent with the current Phenex GUI and will be useful for curating semi-structured systematic character narratives of any taxonomic groups.

Performance

The development corpus for CharaParser for phylogenetic description annotation is rather small, consisting of 3 phylogenetic publications written in two rather different styles. This was done because we needed to provide the larger pool of 50 publications for the BioCreative organizers to select "official" test data.

The development corpus provided 1,200 character descriptions. 100 of them were randomly selected for phenoscape curators to manually curate, resulting in our internal evaluation benchmark. We did not attempt to fine-tune CharaParser on this set of evaluation data or the development corpus, but used these data to adapt CharaParser from generating fine-grained annotations in XML to generating EQ statements. Tuning CharaParser's performance on EQ generation will be the task after we receive evaluation results based on the "official" test data generated from the larger corpus of 50 publications.

CharaParser mimics human curator's behavior in curating a description: first, key entity and quality terms are identified in the description, then, these terms are translated into formal labels/IDs used in ontologies.

In evaluating CharaParser, we evaluate CharaParser's performance in identifying entity and quality terms and its performance in translating term-based EQs to label-based EQs. The results are shown in Table 3. Note that E is not necessarily a single term, but may be a compound expression consisting of primary entity, entity locators (i.e., entities associated with the primary entity via part_of relation), and related entity or entities. A computer proposed EQ is considered matching an answer key when the EQ includes matching elements (either the proposed element contains the answer element or vice versa) for E and Q terms/labels.

Table 2: CharaParser's EQ generation performance

	Term-based EQ	Label-based EQ
Precision	90%	52%
Recall	90%	51%

The preliminary result suggests that our development focus for the next phrase is on improving the translation accuracy from extracted terms to terms in the ontology.

Proposed task for BioCreative Track III:

Input: A set of 50 systematic character narratives of fish or dinosaur selected by BioCreative organizer from the 50+ publications and a set of ontologies (PATO, TAO, BSPO: http://obofoundry.org/cgi-bin/detail.cgi?id=spatial).

Sample systematic character narratives are in Table 1 and more examples can be found at http://kb.phenoscape.org/taxon_annotations (click on "source" in the Results table).

All ontologies are accessible via OBO foundry at http://www.obofoundry.org/.

Output: EQ statements using terms and EQ statements using term IDs. These EQ statements reflect the semantic content of the input. Table 1 shows both types of output.

The task will be performed in two modes: using Phenex and using PCS. In each of the modes, the output EQ statements will be recorded in a table format like that of Table 1.

Task Mode: Using Phenex: curator create EQ statements using Phenex alone.

Task Mode: Using PSO: curator curate EQs proposed by CharaParser in Phenex.

Funding and Acknowledgement

This work was supported by the National Science Foundation [Grant No. DBI-1062404, DBI-1062542 and EF-0849982]. The authors thank Chris Mungall for his constructive comments on the earlier version of this paper.

References

1. Balhoff, J. P., W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. M. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision. 2010. Phenex: Ontological annotation of phenotypic diversity. *PLoS ONE* 5(5):e10500. doi:10.1371/journal.pone.0010500

2. Dahdul, W. M., J. P. Balhoff, J. Engeman, T. Grande, E. J. Hilton, C. R. Kothari, H. Lapp, J. G. Lundberg, P. E. Midford, T. J. Vision, M. Westerfield, and P. M. Mabee. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS ONE* 5(5):e10708. doi:10.1371/journal.pone.0010708

3. Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V., Haendel, M., Segerdell, E., Mungall, C. J., et al. (2007). Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol,* 22(7). doi:10.1016/j.tree.2007.03.013

4. Cui, H. In press. CharaParser for Fine-Grained Semantic Annotation of Organism Morphological Descriptions. *Journal of American Society of Information Science and Technology*.

5. Buckup, P. A. 1998. Relationships of the Characidiinae and phylogeny of characiform fishes (Teleostei: Ostariophysi). Pages 123-144 in Phylogeny *and Classification of Neotropical Fishes* (L. R. Malabarba, R. E. Reis, R. P. Vari, Z. M. S. Lucena, and C. A. S. Lucena, eds.). EDIPUCRS, Porto Alegre, Brazil.

PubTator: A PubMed-like interactive curation system for document triage and literature curation

Chih-Hsuan Wei[1,2], Hung-Yu Kao[2], Zhiyong Lu[1,*]

[1]National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, Bethesda, MD, 20894
[2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C
[*]Corresponding author: Tel: 301-594-7089, E-mail: Zhiyong.Lu@nih.gov

Abstract

Today's biomedical research has become heavily dependent on the access to biological knowledge encoded in expert curated biological databases. As the volume of biological literature grows rapidly, it becomes increasingly difficult for biocurators to keep up with the literature because manual curation is an expensive and time-consuming endeavor. Past research including our own has shown that (semi-)automated computer analysis can greatly improve the curation efficiency. In this work we propose PubTator, a Web-based interactive curation system for assisting literature curation. PubTator features a PubMed-like interface (many biocurators find it to be familiar) and is equipped with multiple state-of-the-art text mining algorithms for facilitating two specific annotation talks: document triage and gene indexing. A demo version is made publicly available at: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

1. Introduction

PubTator is a Web-based tool that allows curators to create, save, and export annotations. As shown in our past study [1], manual curation can greatly benefit from (semi-)automated computer analysis. Hence, PubTator is equipped with multiple advanced computer algorithms for assisting two specific curation tasks: a) document triage and b) bioconcept annotation (e.g. genes).

PubTator is developed based on a prototype system that was previously used at the NCBI for various manual curation projects such as annotating disease mentions in PubMed abstracts. In response to call for participation in BioCreative 2012, we significantly extended our previous system in developing PubTator. First, relevance ranking and concept highlighting were added to ease the task of document triage. Second, state-of-the-art named entity recognition tools (e.g. winning gene normalization systems [2,3] in BioCreative III) were integrated to pre-tag bioconcepts of interest, as a way to facilitate the task of gene/disease/chemical annotation. Third, PubTator was developed to have a look-and-feel similar to PubMed, thus minimizing the learning efforts required for new users. Furthermore, a standard PubMed search option is made available in PubTator, which would allow our users to make a hassle-free move of their saved PubMed queries (a common practice for curators doing document triage) into this new curation system. Finally, by taking advantage of pre-tagging bioconcepts, PubTator also allows its users to do semantic search besides the traditional keyword based search, a novel feature not available in PubMed.

2. System description

2.1 PubTator search page

For the convenience of many PubMed users, by default PubTator allows the same search syntax and returns identical search results as PubMed. This is achieved by using the Entrez Programming Utilities Web service API.

In addition to the traditional keyword search, an advanced semantic search is featured in PubTator, which enables our users to retrieve articles associated with specific semantic bioconcepts. In the current implementation, a user can choose from one the three semantic categories: gene/proteins, diseases, and chemicals. This is to specifically address a known problem in biomedical literature search: a bioconcept is often associated with multiple different names. When using the semantic search, a user can retrieve all the papers relevant to a concept without having to enumerate the entire set of possible aliases. For instance, searching for the breast cancer gene HER2 will also retrieve articles only mentioning its alternative names such as NEU or ERBB2 (e.g. See result #8 in Figure 2).

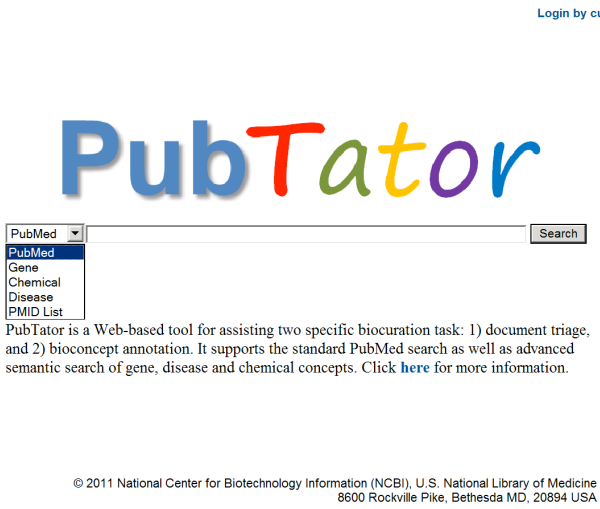


Figure 1: The PubTator homepage.

Finally, we also provide a search option of using a list of PubMed identifiers (PMIDs). This is desirable when one or more articles have been judged to be relevant and need to be curated.

The link in the upper right hand corner is for the user to sign in. Once signed in, it will show the name of the curator (See Figure 2).

2.2 PubTator results page

Following the tradition of PubMed, by default PubTator returns search results in the reverse chronological order. However, only 15 results are returned per page in PubTator vs. 20 in PubMed, making room for quickly displaying the abstract. As shown in Figure 2, a user can click the ABSTRACT link below the PMID to take a peek at the abstract without having to going to a separate abstract page.

As shown in Figure 2, relevance-based ranking is an alternative option in PubTator when a curation team provides PubTator with their curation guidelines and training data (e.g. CTD data in BioCreative III Track I). In such cases, we will pre-compute a relevant score for each candidate article by using machine-learning algorithms (e.g. SVM) [4]. Next, the computed scores will be normalized and subsequently used for ranking search results.

As a novel feature to help document triage, we also highlight key concepts in the title and abstract. Currently, four different concepts are pre-annotated and highlighted: Gene (purple), Chemicals (green), Diseases (orange), and Species (blue).

Figure 2: The PubTator results page.

To the right of the search results, we show two advanced search options. On the top panel, users can refine their search results by taxonomy. This feature is useful for those curation teams who work with a specific organism because by default we show results across all species. In the lower panel, one can choose to turn off one or more highlighted concepts if desired.

For the document triage task, a curator can select the relevant papers from the search results by simply checking the box next to its number. To further examine an article or perform the detailed annotation task, a curator then needs to go to the abstract page as described below.

2.3 PubTator abstract page

When an article title is clicked in the results page, PubTator returns its abstract page in response. Concepts are annotated in this page as follows: 1) a piece of text is color-highlighted and assigned

to a semantic category; and 2) a standard database identifier is searched and assigned to the selected text mention.

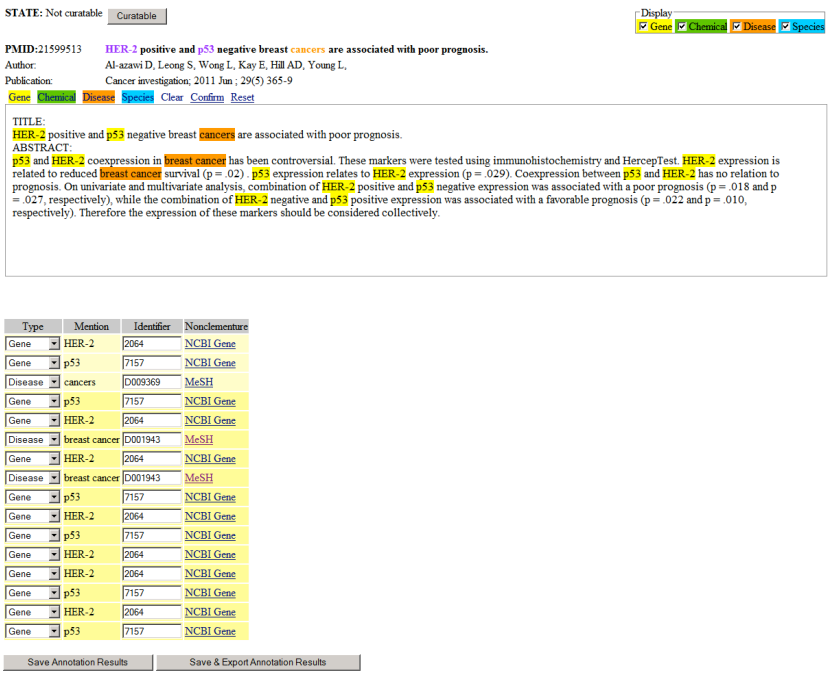


Figure 3: The PubTator abstract/annotation page.

As shown in Figure 3, at the very top of the page, the paper's current curation status is shown (curatable or not). Clicking on the button next to it will readily change its status, giving our user an option to perform document triage also in the abstract page. Immediately below is some publication metadata including the PMID, title, author(s), journal, and publication date.

Under the metadata, the title and abstract are displayed in a text box where a user can manipulate annotations in a number of different ways:

1. To create an annotation: selecting a piece of text and click one of the four semantic categories (e.g. gene).
2. To remove an annotation: selecting an existing annotation and click 'Clear'.
3. To reset annotations: by clicking 'Reset', system will return to the results that were last modified.
4. To commit annotations: by clicking 'Confirm', all highlighted text mentions will be added to the Table immediately below where the second step of concept annotation—assigning the concept id for the highlighted textual mention—is required.

To facilitate the concept annotation process, we pre-tag all concepts in the title and abstract using state-of-the-art text mining tools (See more in Section 5). However, if one or more concept categories are not needed for a specific task, those pre-computed concepts could be removed by clicking the x icon in front of the corresponding category. For the sample article shown in Figure 3, most machine generated annotations are correct; the only manual work is to correct an annotation in the title (change 'cancers' to 'breast cancers') and its MeSH ID accordingly in the

Table below where database identifiers are assigned to the corresponding selected text mentions. After accepting or correcting concept ids, the user can click to save or export all annotations (both text spans and concept ids) of the article. In either case, the time information for this annotation is also saved into the PubTator system.

3. Proposed tasks for BioCreative 2012 Track III

We propose two general tasks that can be achieved using PubTator. Once a user is committed, a customized version will be provided should they have any specific requirements (See more about our system adaptability in Section 4).

1. Document triage

This task will assess our system for assisting human curators to prioritize papers for more detailed curation. A curator with a specific need will decide a query (e.g. a chemical name in the case of CTD curation) and search it in PubTator. The curator will then examine the returned search results and mark relevant papers to be curated. The experience with PubTator can be compared with the system they are currently using or general-purpose systems like PubMed with respect to productivity and effectiveness.

Input: a concept (gene/disease/chemical) name/identifier OR any PubMed query

Output: a list of PMIDs that are selected for further annotation.

2. Bioconcept annotation

This task will assess our system for assisting manual annotation of various kinds of bioconcepts. A curator can use our system to create and export annotations with regard to specific concepts. For instance, annotating genes is a central task for many model organism databases. After entering a list of PMIDs, PubTator will return the corresponding articles with machine tagged pre-annotations. The curator can then accept/edit/remove them or create new annotations. The goal is to see if using PubTator can accelerate this labor-intensive manual process.

Input: a list of PMIDs

Output: PMIDs with corresponding annotations (database identifiers).

4. System adaptability and interactivity

In developing PubTator, we decided to develop it as a general curation tool rather than a specialized one in order to reach a broader community. However, once a team decides to adapt our system into their curation pipeline, PubTator is quite robust for customization. For instance, any team can use PubTator to perform the document triage task via a simple query. In this case, search results are returned in reverse time order by default. However, in the case where teams wish to have search results ranked by relevance, we can easily achieve this based on team provided training data. Indeed, this is the case we are doing for the CTD document triage task (See BioCreative 2012 Track I for details).

Similarly for bioconcept annotation, we can customize PubTator for different needs of model organism databases. For instance, in default setting NCBI Gene database is used in gene ID assignment. However, this can be changed to any other organism-specific gene nonclementure such as the Arabidopsis Genome Initiative locus identifiers.

Our system is Web-based and involves a great deal of human interaction. As described earlier, users are involved in many curation aspects ranging from selecting/deselecting articles in document triage to creating/editing/deleting pre-tagged markups in bioconcept annotation.

5. System evaluation

With regard to the underlying algorithms employed in the PubTator, some have already been extensively evaluated with exceptional performance (See Table 1 for details). We are currently evaluating the recall and precision of our algorithms in finding disease and chemical concepts. In addition, using the CTD data, a machine-learning based algorithm will be separately assessed in ranking curatable articles for document triage. All these benchmark experiments will be completed before the user-testing period in March 2012.

Module Name	Targeted Use	Precision	Recall	F-measure
GeneTUKit [2]	Gene Mention (abstract)	86.73%	82.36%	84.49%
GenNorm [3]	Gene Normalization (full text)	56.23%	39.72%	46.56%
SR4GN [5]	Species Recognition (abstract)	85.42%	85.42%	85.42%

Table 1: Reported benchmark performance of computational modules used in PubTator

Through participation in the BioCreative 2012 track III, we plan to collect interactive data and subsequently perform comparative analysis of curation effectiveness using our system vs. manual or another curation system. Furthermore, the user-curated data during the system testing could be used as the gold standard to report performance metrics such as precision and recall as requested by the track III organizers.

Acknowledgments The authors are grateful to Smith L, Comeau D and Dogan R for building the prototype annotation system. We also thank Wilbur WJ and Kim S for helpful discussion, Comeau D proofreading the manuscript, and the GeneTUKit authors making their software publicly available. Funding: NIH Intramural Research Program, National Library of Medicine.

References

1. Névéol A, Doğan RI, Lu Z (2010) Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics* **44**: 310-318.
2. Huang M, Liu J, Zhu X (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics* **27**: 1032-1033.
3. Wei C-H, Kao H-Y (2011) Cross-species gene normalization by species inference. BMC *Bioinformatics* **12**: S6.
4. Yeganova L, Comeau DC, Kim W, Wilbur WJ. Text Mining Techniques for Leveraging Positively Labeled Data; 2011. pp. 155-163.
5. Wei C-H, Kao H-Y, Lu Z (2011) SR4GN: a species recognition software tool for gene normalization. *Plos one*. Submitted

PPInterFinder – A Web Server for Mining Human Protein - Protein Interactions

Kalpana Raja[1], Suresh Subramani[1] and Jeyakumar Natarajan[1 *]

[1]Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamilnadu, India

*Corresponding author: Tel: +91 422 2428281, E-mail: n.jeyakumar@yahoo.co.in

Abstract

One of the most common and challenging problem in biomedical text mining is to mine protein-protein interactions (PPIs) from MEDLINE abstracts and full-text research articles because PPIs play a major role in understanding the various biological processes and the impact of proteins in diseases. We implemented, PPInterFinder – A web based text mining system to extract human protein-protein interactions from biomedical literature. PPInterFinder uses interacting keyword co-occurrences with protein names to extract information on PPIs from MEDLINE abstracts. In addition, our system has heuristics for negation recognition to exclude false PPIs and to improve accuracy. We evaluated the system performance on a dataset of 693 sentences related to human proteins, derived from IntAct database. The performance of the system is evident from its precision 81.28%, recall 71.27% and 75.94% F-score.

Introduction

Protein-protein interactions (PPI) are of central importance to understand the mechanisms of biological processes and diseases (1). The knowledge about PPI is rapidly growing with the results from high-throughput experimental technologies. Accordingly, a huge number of interaction data is being published in the literature (1, 2). A wide range of interaction databases such as IntAct, MINT, BIND and PIE have been developed by manually curating the protein interactions from various information sources. However, the rapid growth of biological publications in recent years made this time consuming task almost impractical for the PPI extraction. Consequently, many of the PPI data are still available only in the literature (3). Extraction of such information from biomedical literature has become an important topic in the field of biomedical natural language processing (BioNLP) (4). Several approaches ranging from co-occurrence principle to more sophisticated machine learning (ML) methods have been reported for extracting the PPI information (5-7).

Extraction of PPI from literatures broadly consists of two components, protein name recognition and PPI extraction, both of which are equally challenging. Though many approaches have been proposed for the extraction of PPI information from the biomedical literature, the problem still remains as an open challenge for the researchers to develop more accurate and robust automated methods to address the problem. Moreover, PPI systems specific to human genes/proteins are very few (8). In this paper, we present PPInterFinder, a novel PPI extraction web server for mining human PPIs from the biomedical literature. This dedicated human protein interaction web server is helpful to the users to find both known and potentially novel human PPIs from literature.

Material and Methods

Architecture and components

PPInterFinder is a freely available web server for mining human PPIs. The project is a combination of a Java libraries for relation keyword recognition, negation recognition, pattern recognition (PPI triplets), and extraction of PPI information. In addition, a Perl module implementing Perl/CGI scripts is used for uploading user inputs. The extracted PPI interactions are stored in MySQL relational database management system. The work flow of the system is as follows:

(i) Text preprocessing
(ii) Extraction of PPI information
(iii)Database storage

Figure 1 shows a general work flow of the system.

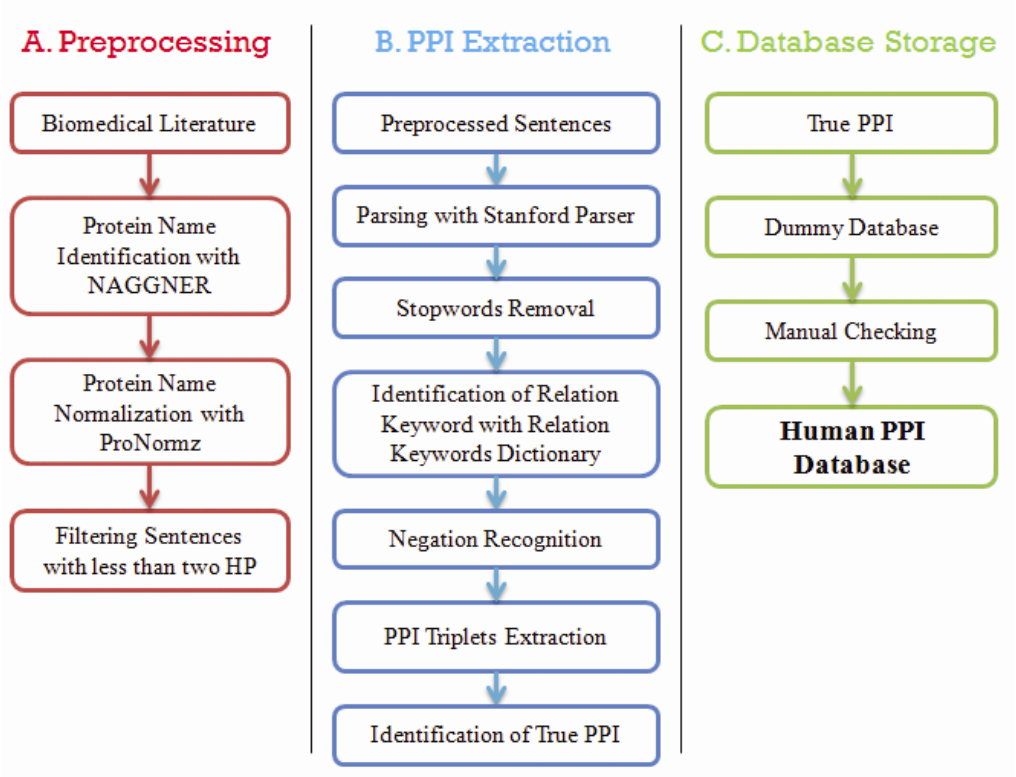


Fig 1. Work flow of PPInterFinder

Text preprocessing

The input text can be either a PubMed abstract in plain text format or in MEDLINE / XML format with unique PubMed ID. An initial preprocessing is carried out to match PubMed IDs with individual sentences in the abstract. Further processing includes (i) identification and normalization of protein names, and (ii) filtering out of input sentences with only one protein or no proteins names, The protein name recognition and normalization are carried out by our own tools namely, NAGGNER (9) and ProNormz (10) which are highly specific to human proteins.

Extraction of PPI information

Dictionaries

The success of PPI system relies on the successful identification of interaction keywords. To achieve this goal, we have developed a vast relation keyword dictionary, which consists of 374 relation keywords. The keywords are grouped into 93 subtypes by identifying the common root word for each subgroup (Supplemental Data 1). The relation keyword dictionary is created on

the basis of the keywords used in the previous works related to interaction extraction (11-14), and further augmented with relation keywords from other interaction databases such as IntAct, MIND and DIP.

In addition to the relation keywords dictionary, we further utilized a stop words dictionary, the task which is already well established in BioNLP. The stop word list used is the SMART search engine English stop word list available from (ftp://ftp.cs.cornell.edu/pub/smart/english.stop). The stop words dictionary improves the speed of the algorithm by reducing the number of words needed to be matched with the relation keyword dictionary during the pattern matching process.

Relation keyword recognition

The identification of the relation keyword is a vital step prior to the extraction of PPI information. Our text mining and natural language processing methods involving the identification of relation keywords is illustrated in Figure 2 and comprise of following steps.

(i) The input sentence is parsed using Stanford Parser (15) with grammer settings to englishPCFG module to generate the constituent tree of verb and noun phrases.
(ii) A group of words enclosed inside a pair of braces forms a single word bag. The algorithm separates such word bags with the header word VP (verb phrase) first followed by word bags with NP (noun phrase) as the header word based on the following set of rules:
    a. Each word bag is expected to contain either one VP or NP. Otherwise, split the word bag in such a way that every derived word bag contains only one VP or NP.
    b. Only the word bags having a verb or noun are selected for further processing.
    c. The algorithm then generates a keyword list by collecting the words from the selected word bags from b.
    d. Stop words are removed by pair-wise matching of each word of the keyword list with the stop words dictionary.
    e. Then, the algorithm performs a second pair-wise matching between the remaining words in the keyword list against the relation keyword dictionary.
    f. The final matching word is declared as the relation keyword.

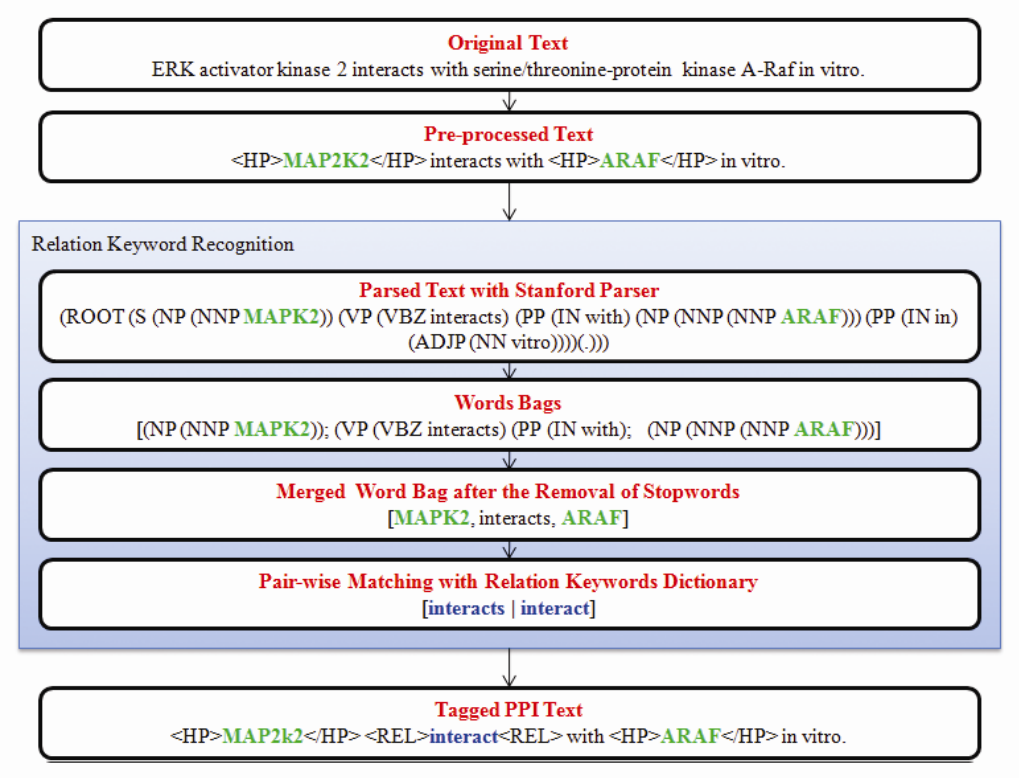


Fig 2. Algorithm to extract relation keyword

Negation keyword recognition

The success of every automated PPI extraction from biomedical literature invariably depends on the proper recognition of negation keywords (16). In the present study, we consider the recognition and tagging of four negation keywords *'no', 'not',* and *'neither/nor'* as mostly these negation keywords associated with false PPI information. These negation keywords are normally occur as an adverb (e.g. 'not'), or a determiner (e.g. 'no') or a coordinating conjunction (e.g. 'neither/nor'). T he algorithm locates the presence of any negation keyword in the parsed sentence through pattern matching, similar to relation keyword recognition.

Pattern recognition to extract PPI triplet sentences

In biomedical text, the relationship between two entities (protein – protein), commonly named as PPI triplets, can be expressed in different abstract forms (11, 17). In the present study, we used the following three types of abstract forms depending on t he position of the relation keyword with the two proteins.

PPI Form1:	PROTEIN1 – token* - RELATION – token* - PROTEIN2
Examples:	PROTEIN1 interacts with PROTEIN2
	PROTEIN1 has weak association with PROTEIN2
PPI Form2:	RELATION – token* - PROTEIN1 –token* - PROTEIN2
Example:	interaction between PROTEIN1 and PROTEIN2
PPI Form3:	PROTEIN1– token* - PROTEIN2 –token* - RELATION
Example:	PROTEIN1 and PROTEIN2 complex

Form1 is the most common form with relation keyword in between the pair of proteins (protein – relation – protein). In Form1, the relation keywords are commonly a verb, verb with additional tokens/words or even a noun. Form2 and Form3 are comparatively rare with relation keywords at the corners (relation – protein – protein / protein – protein – relation). In such cases the relation keyword is mostly a noun.

The number of words (tokens*) in between the relation keyword and the proteins vary widely and the current study includes all the sentences satisfying the above three abstract forms regardless of the number of tokens* in between the relation keyword and the protein. This is illustrated in following two examples. In example 1, the number of tokens between the relation keyword and the protein is 1 and in example 2 the number of tokens between the relation keyword and the protein is 5.

Example 1:
PubMed ID: 11854419: Point mutation in fus6 -T236 disrupts PROTEIN>CSN1S1</PROTEIN> <RELATION>interact</RELATION> with <PROTEIN>CSN2</PROTEIN> and CSN4 in a yeast-two-hybrid assay.

Example 2:
PubMed ID: 12529446: <PROTEIN>ARHGEF1</PROTEIN> p specifically <RELATION>interact</RELATION> with the GDP-bound form of <PROTEIN>F3</PROTEIN> p (Cdc42D118A or Cdc42T17N).

PPI extraction

Finally, we incorporated four rules for extracting PPIs from a PPI triplet sentence related to the three abstract forms discussed above. Our rule sets are capable of processing simple sentences with at least two proteins (Rule 1), simple sentences with at least two proteins and a negation keyword (Rule 2), complex sentences having more than two proteins (Rule 3), and complex sentences having more than two proteins and a negation keyword (Rule 4). Fig 3 shows the input and the extracted output of PPInterFinder.

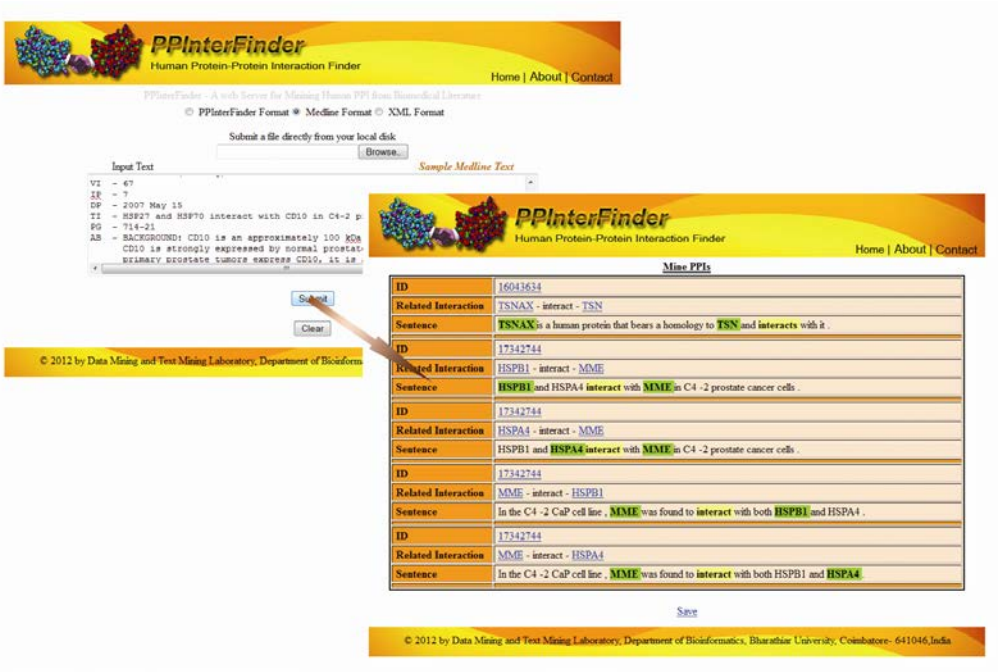


Fig 3. Screenshot of PPInterFinder web server

Rule 1: Sentences with two proteins and a relation keyword

The procedure is the most simple to extract two proteins and the relation keyword identified during the relation keyword recognition phase. Example 3 i llustrates the extraction of PPI information between the two proteins MAP2K2 and ARAF.

Example 3:
PubMed ID: 11909642: <PROTEIN> MAP2K2 </PROTEIN> <RELATION> interact </RELATION> with </PROTEIN> ARAF <PROTEIN>in vitro.

Rule 2: Sentences with two proteins, a relation keyword and a negation keyword

The approach is very similar to Rule 1, except the role of negation keyword in filtering the false PPI information from the biomedical literature, as shown in Example 4.

Example 4:
PubMed ID: 16899217: There was <NEGATION> no </NEGATION> detectable <RELATION> interact </RELAION> between <PROTEIN> PSMC6 </PROTEIN> and <PROTEIN> PSMC5 </PROTEIN>.

Rule 3: Sentences with more than two proteins and a relation keyword

We use an algorithm for Rule 3 as illustrated in Fig 4. The complexity of the algorithm depends on the number of proteins present in the input sentence.

(i) The word position is assigned to each word in the sentence, starting from 0.
(ii) A hash table is generated to hold the recognized proteins, relation keyword and their corresponding word position.
(iii) Next, the relation keyword in the hash table is identified.
(iv) All possible PPI triplets are generated by combining the relation keyword with each of the preceding and succeeding proteins.
(v) Finally, all the true PPIs are declared.

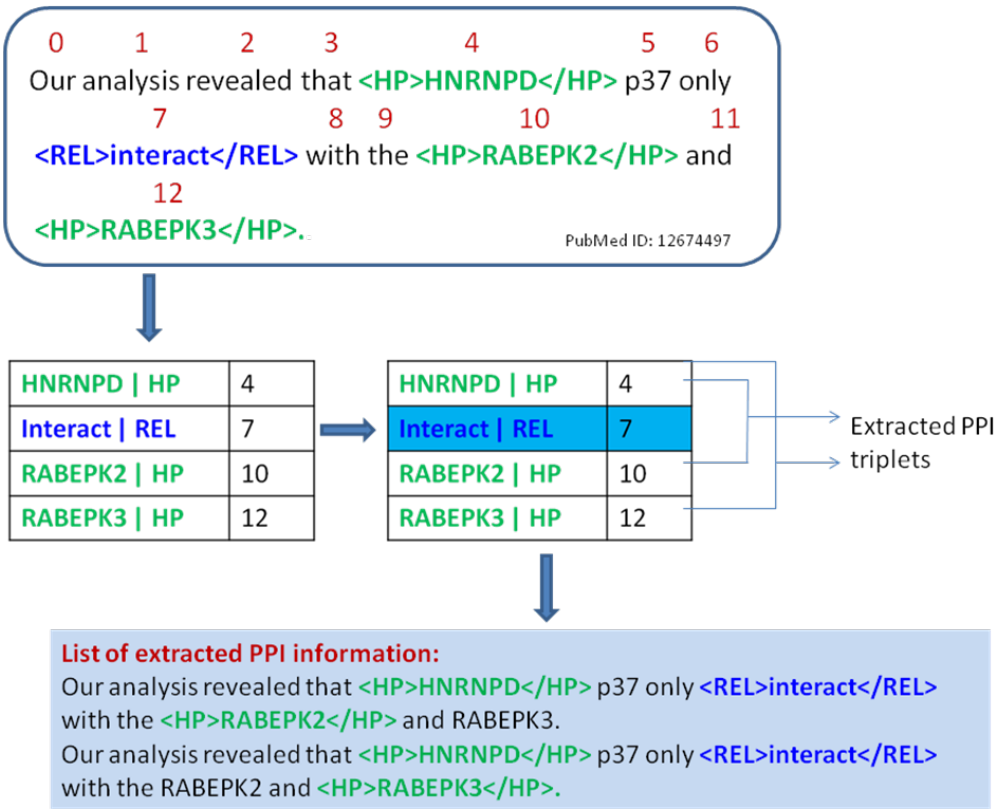


Fig 4. Flowchart for Rule 3 algorithm

Rule 4: Sentences with more than two proteins, a relation keyword and a negation keyword

The algorithm is very similar to Rule 3 and in addition it checks the presence of negation keyword to filter the false PPI information. Rule 4 implements two separate rules based on the presence of negation keyword *'not'* or *'no'* (Rule 4a) and *'neither/nor'* (Rule 4b).

Rule 4a: Negation keyword as 'not' or 'no'

The algorithm is a hybrid of both Rule 2 and 3. The proteins following the negation keyword are considered to be false PPIs and subsequently eliminated.

Rule 4b: Negation keyword as 'neither/nor'

The algorithm looks for a specific order of PPI triplets with the negation keyword to confirm the true PPIs as illustrated in Fig 5.

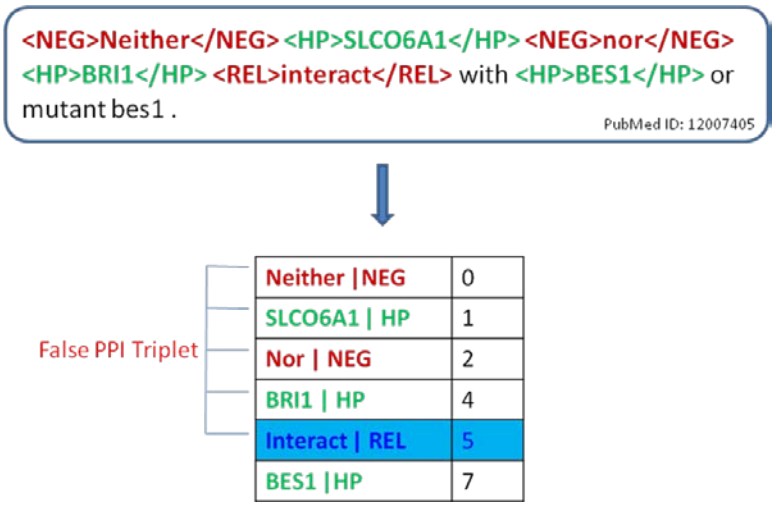


Fig 5. Flowchart for Rule 4b algorithm

Database storage and validation

All the extracted true PPIs are stored in a local database implemented in MySQL relational database management system. The database stores the details on PubMed ID, two proteins, relation keyword with the corresponding root keyword and the sentence from which the interaction is extracted. Presently, we use two databases, a dummy database and a main database to validate and store the true PPI information. The extracted PPI information is stored initially in the dummy database and later checked manually with the corresponding PubMed article and then updated to the main backend database. The idea behind the maintenance of two databases is to ensure the accuracy of PPI entries in the main database. We will be in the process of developing an automated system to validate the entries of the dummy database before being updated to the main database.

Results and Discussion

Dataset and Evaluation

Two biologists with knowledge on human protein collected a dataset of 693 sentences related to human PPIs from IntAct Database (18). Most of the sentences in the dataset include two or more proteins while few contain one or more negation keywords. The biologists manually tagged the 693 sentences based on the co-occurrence of protein names with the relation keyword. We used the same dataset in PPInterFinder to extract the PPI information in order to evaluate the performance of the system. We compared the extracted PPIs from PPInterFinder with the manual tagged sentences. The performance is measured via the commonly used criteria precision, recall,

and the F-measure. The definition of these performance measures is given by Equation (1) to (3) respectively.

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (3)$$

where TP refers to the number or proportion of protein entities that were correctly identified by the system; FN refers to the number or proportion of protein entities that the system failed to identify; and FP refers to the number or proportion of protein entities that were incorrectly identified by the system.

Performance of PPI Extraction algorithm

We evaluated the system performance at two stages, initially during text preprocessing phase and finally after relation keyword recognition and PPI extraction phase. A manual analysis on the output from two stages is carried out by biologists in order to ensure the performance and accuracy PPInterFinder. We find 61 sentences out of 693 were filtered out after the text preprocessing phase for not having a minimum of two HGNC approved human protein names. Manual checking with the help of the biologists confirmed that out of 61 filtered sentences 48 sentences have no human protein names (Supplementary Data 2). For example, the protein *NALP1* is not a human protein and our system correctly identifies the same and excludes that sentence (Example 5). However, the human proteins in the 13 remaining sentences are unidentified as these sentences not use HGNC approved protein names though they are human proteins. For instance, *Raf* is a human protein and not identified during the preprocessing phases as it is the non HGNC approved name (Example 6).

Example 5:
PubMed ID: 17418785: We attempted to map the region of
<OtherProtein>NALP1</OtherProtein> required for binding Bcl-XL.

Example 6:

PubMed ID: 16301319: <OtherProtein>Erbin</OtherProtein> bound to <HumanProtein>SHOC2</HumanProtein> and inhibited the interaction of <OtherProtein>Sur-8</OtherProtein> with <OtherProtein>Ras</OtherProtein> and Raf.

The next crucial phase after text preprocessing is the recognition of relation keywords. 276 sentences were filtered out during this stage for not having a relation keyword or having more than one relation keyword. The challenge involved in PPI extraction from sentences having more than one relation keyword is complex and not included in our current study. So, we excluded those sentences during this stage. Manual analysis on these 276 sentences filtered out during the relation keyword recognition stage confirmed that all excluded sentences are appropriate one. Consequently, only 356 sentences were processed for PPI extraction. PPInterFinder generated 470 PPI extractions from these 356 sentences (Supplementary Data 2). Cross referencing of the output with the manual annotations by the biologists revealed that 382 are true PPI extractions while 88 termed to be incorrect due to wrong PPI pairing. Further, 154 sentences remain unprocessed by the PPI extraction algorithm as they are not confined with the three abstract forms for PPI extraction. Overall, we obtained the performance of PPInterFinder on human PPI dataset with 81.28% precision, 71.27% recall and 75.94% F-measure (Table 1). However, these results not include the 13 sentences excluded during pre-processing stage due to non-recognition human proteins, which are not in the HGNC database.

Human PPI Dataset (693 Sentences)	
True Positive	382
False Positive	88
False Negative	154
Precision	81.28
Recall	71.27
F-score	**75.94**

Table 1: Evaluation of PPI extraction algorithm on human PPI Dataset
Precision, Recall and F-measures on human PPI Dataset

Conclusion

We described a w eb based human PPI extraction system PPInterFinder to mine the PPI information from a given biomedical literature. The server extracts PPI information based on the co-occurrence of the relation keyword with proteins using heuristic approaches. PPIInterFinder can be useful for biologists to search proteins/genes of their interest and to find some novel PPIs from published biomedical literature. In present form, the system is available for human PPI information extraction on s ingle sentences with two or more proteins and one interacting

keyword. We will be expanding the system further to extract the information across sentences and on s entences having multiple interacting keywords using hybrid machine learning approaches. Further, we will incorporate an automated module to cross check the extracted PPI entries with PubMed database, to ensure the accuracy of the extracted information before updating to the PPI database.

Acknowledgements

This work was supported by Department of Information Technology, Government of India. [DIT/R&D/BIO/15(22)]

Supplementary Data

Supplementary Data 1: List of relation keywords [http://www.biomining-bu.in/ppinterfinder/]

Supplementary Data 2: Evaluation of human PPI dataset
[http://www.biomining-bu.in/ppinterfinder/]

References

1. Kann,M.G. (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform.,* **8,** 333–346.
2. Huang,M., Ding,S., Wang,H. and Zhu,X. (2008) Mining physical protein-protein interactions from the literature. *Genome Biology*, **9**(Suppl 2):S12.
3. Cusick,M.E. et al. (2009) Literature-curated protein interaction datasets. *Nat. Methods,* **6**, 39–46.
4. Miwa,M. et al. (2010) Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.,* **8**, 131–146.
5. Kabiljo,R. et al. (2009) A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics,* **10**, 233.
6. Giles,C. and Wren,J. (2008) Large-scale directional relationship extraction and resolution. *BMC Bioinformatic*s, **9,** S11.
7. Björne,J. et al. (2010) Complex event extraction at PubMed scale. *Bioinformatics,* **26,** i382–i390.
8. He, M., Wang, Y. and Li, W. (2009) PPI Finder: A Mining Tool for Human Protein-Protein Interactions. *PLoS ONE*, **4**(2): e4554.
9. Kalpana, R., Suresh, S. and Jeyakumar, N. (2012) NAGGNER – A hybrid named entity tagger for tagging human proteins/genes. *In the proceedings of the 10th Asia Pacific Bioinformatics Conference, Melbourne, Australia.*
10. Suresh, S., Kalpana, R. and Jeyakumar, N. (2011) ProNormz – An automated web server for human proteins and protein kinases normalization. *In the proceedings of the 2nd International Conference on Bioinformatics and Systems Biology(INCOBS), Chidambaram, India.*
11. Bui,Q.C., Katrenko,S. and Sloot,P.M.A (2011) A hybrid approach to extract protein–protein interactions. *Bioinformatics,* **27**(2): 259-265.
12. Temkin,J.M. and Gilder,M.R. (2003) Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics.* **19**(16): 2046–2053.
13. Huang et.al. (2004) Discovering patterns to extract PPI from full texts. *Bioinformatics.* **20**(18): 3604-3612.
14. Ono,T. et al. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics,* **17,** 155–161.
15. Klein,D. and Manning,C.D. (2003) Accurate Unlexicalized Parsing. *In the Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.*
16. Chowdhary.R., Zhang.J. and Liu.J.S. (2009) Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics,* **25**(12), 1536-1542.
17. Rinaldi,F. et al. (2010) OntoGene in BioCreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinformatics,* **7,** 472–480.
18. IntAct Dataset, ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining/

Mining Protein Interactions of Phosphorylated Proteins from the Literature using eFIP

Catalina O Tudor[1*], Cecilia N Arighi[1,2], Qinghua Wang[1,2], Cathy H Wu[1,2], K Vijay-Shanker[1]

[1] Department of Computer and Information Sciences
[2] Center for Bioinformatics and Computational Biology
University of Delaware, Newark DE, USA

*Corresponding author: Tel: 302 831 8496, E-mail: tudor@cis.udel.edu

Abstract

One important aspect of proteins is the regulation of their molecular function and/or subcellular localization by post-translational modifications. In particular, phosphorylation is one of the most studied modifications as it has a significant impact in the regulation of many cellular processes. Phosphorylation of a protein may lead to activation or down-regulation of their activity, alternative subcellular location, and interaction with different binding partners. Extracting this information is critical to the interpretation of protein networks and prediction of the functional outcomes, and this is the main motivation for capturing this type of information in the Protein Ontology (PRO). To assist PRO curators, we have developed eFIP, a web-based tool, to find relevant abstracts mentioning phosphorylation of a given protein (including site and kinase), coupled with mentions of protein-protein interactions (PPIs), and evidence for the impact of phosphorylation on the interaction. eFIP combines information provided by applications such as eGRAB, RLIMS-P, and an in-house protein interaction module, and displays the results in a highlighted and tabular format for a quick inspection and validation. Validated results can be exported including evidence sentences. In our internal evaluation of 90 abstracts using eFIP interface for validation, eFIP yielded 95.52% and 96.96%, precision and recall, respectively, at the abstract level, and 84.44% and 86.36%, precision and recall, respectively, at the sentence level. A second evaluation of eFIP was conducted against a gold standard corpus consisting of 151 sentences, and this yielded 86.11% precision, 60.78% recall, and 83.44% accuracy.
URL: http://biotm.cis.udel.edu/eFIP/.

Introduction

One important aspect of proteins is the regulation of their molecular function by post-translational modifications. In particular, phosphorylation of specific intracellular proteins/enzymes by protein kinases and dephosphorylation by phosphatases provide information of both activation and deactivation of critical cellular pathways, including regulatory mechanisms of metabolism, cell division, cell growth and differentiation. Often, protein phosphorylation has some functional impact. Proteins can be phosphorylated on di fferent residues, leading to activation or down-regulation of their activity, alternative subcellular location, and distinct binding partners. One such example is protein Smad2, whose phosphorylation state determines its interaction partners, its subcellular location, and its cofactor activity.

However, protein interaction data involving phosphorylated proteins is not yet well represented in the public databases. Extracting this information is critical to the interpretation of protein networks and prediction of the functional outcomes, and this is the main motivation for capturing this type of information in the Protein Ontology (PRO). PRO is a structured representation of protein forms and protein complexes, and includes relationships to properties of those entities (4). To assist curators in finding relevant abstracts for curation (with mentions of phosphorylation of a given protein (including site and kinase), coupled with mentions of protein-protein interactions (PPI) and possible impact of phosphorylation), we have developed eFIP (Extracting Functional Impact of Phosphorylation) (1). This tool currently focuses on identifying phosphorylated forms of proteins and the participation of these phosphorylated forms in protein-protein interactions (PPIs as defined by De Las Rivas and Fontanillo (PMID 20589078)). The types of PPIs captured by this tool include binary interactions, protein-protein complex and protein-a class of proteins, as these are all objects and relationships of interest for PRO curation.

The System's Modules

The pipeline of the system involves a document retrieval module, called eGRAB (5), an information extraction tool for identifying mentions of phosphorylation, called RLIMS-P (2,3), and a PPI module developed in-house. The goal is to identify the participation of phosphorylated forms of a protein in protein-protein interactions. Possible inputs are a protein name, a protein identifier, or a list of PMIDs. Although the protein identifiers are species-specific, we do not limit the abstracts to the ones mentioning the protein with the corresponding species. Instead, we use the identifier to retrieve all possible names of the protein to be used in the document retrieval module. We concentrate on computing results for proteins coming from vertebrates. Currently, eFIP considers only the abstract of a paper, since one of the components, RLIMS-P, works only with abstracts. We plan to extend our approach to the Results section of papers.

Extractor of Gene-Related ABstracts (eGRAB)

eGRAB is used to gather the literature for a given gene/protein. eGRAB starts by gathering all possible names and synonyms of a gene/protein from knowledge bases of genes/proteins (e.g., EntrezGene and UniProtKB), searches PubMed using these names, and returns a set of disambiguated abstracts to serve as the gene/protein's literature. This technique filters potentially irrelevant documents that mention the names in some other context, by creating language models for all the senses, and assigning the closest sense to an ambiguous name. eGRAB is currently being used in other systems. The approach and its evaluation are provided in (5).

Rule-based LIterature Mining System for Protein Phosphorylation (RLIMS-P)

RLIMS-P is a system designed for extracting protein phosphorylation information from abstracts. It extracts the three objects involved in this process: the protein kinase, the phosphorylated protein (substrate), and the phosphorylation site (residue/position being phosphorylated). RLIMS-P utilizes extraction rules that cover a wide range of patterns, including some specialized terms used only with phosphorylation. Additionally, RLIMS-P employs techniques to combine information found in different sentences, because rarely are the three objects (kinase, substrate, and site) found in the same sentence. RLIMS-P has been benchmarked and the results are presented in (2). A detailed description of the system can be found in (3).

The Protein-Protein Interaction (PPI) Module
In eFIP protein-protein interactions include the following types: binary interactions, protein-protein complex and protein-a class of proteins, as these are all objects and relationships of interest for PRO curation. After surveying many PPI tools that have been described in the literature, none of these wereavailable for download to use in eFIP's pipeline. Additionally, the PPI tool would have to be easily adaptable for our needs. One example of additional features we wanted to be able to incorporate is the ability to detect interactions involving only one partner, in the cases in which the other partner is implicit, or the ability to detect anaphora resolution when one of the partners or both are described by pronouns "it" or "they".

Hence, the PPI module is an in-house implementation designed to detect mentions of PPI in text. This tool extracts text fragments, or text evidence, that explicitly describe a type of PPI (such as binding and dissociation), as well as the interacting partners. The primary engine of this tool is an extensive set of rules specialized to detect patterns of PPI mentions. The interacting partners identified are further sent to a gene mention tool to confirm whether they are genuine protein mentions. Consider the sample phrase "several proapoptotic proteins commonly become associated with 14-3-3." "14-3-3" is a protein, whereas "several proapoptotic protein" prompts the need to further identify the actual proteins (Bad and FOXO3a) that interact with 14-3-3.

The extraction of phosphorylation impact on protein interaction
Our main goal is to find protein interaction information about a particular protein when it is in its phosphorylated state. For this, we select abstracts that contain both phosphorylation and protein interaction mentions involving the same protein. We define impact as any direct relation between protein phosphorylation and protein interaction. Therefore, The relation could be positive (phosphorylation of A increases binding to B), negative (when phosphorylated A dissociates from B) or neutral (phosphorylated A binds B).

For example, consider the following sentence: "Phosphorylated Bad binds to the cytosolic 14-3-3". In this example, we can tell that the phosphorylation happens before the binding, as one of the interactants is reported to be "phosphorylated Bad". However, we cannot tell if the phosphorylation has any impact on the binding itself, i.e., if 14-3-3 binds to Bad regardless of its form, phosphorylated or non-phosphorylated. In contrast, the next sentence shown here not only mentions the phosphorylation happening before the interaction, but also describes how the interaction is dependent on t he phosphorylation: "Bad phosphorylation induced by survival factors leads to its preferential binding to 14-3-3 and suppression of the death-inducing function of Bad."

Results and User Interaction
The input in eFIP is a protein name (or identifier), or a list of PMIDs. The output is a ranked list of PMIDs, each accompanied by a summary of the information found within.

If a protein name or identifier is provided, eFIP outputs all relevant articles where the corresponding protein is phosphorylated and implicated in a protein-protein interaction. eFIP ranks these papers, taking into consideration the confidence assigned to each of the steps involved: the detection of the phosphorylation mention, the detection of the partners of the interaction, and the detection of the impact of phosphorylation on t he interaction. If a list of

PMIDs is provided as input, then multiple phospho-proteins might be involved in protein-protein interactions. eFIP lists relevant PMIDs for one phospho-protein at a time, first considering the phospho-protein that has more mentions of phosphorylation and PPI in the documents provided.

An example output is shown in Figure 1 for protein BAD. There are 1,331 documents linked to protein BAD as determined by eGRAB. Alternatively, we can provide a list of PMIDs, e.g., 8929531, 10949026, 11526496, 11583580, 12351720, 12438947, 15896972, 16139821. The user can click on the PMID or "view evidence", and this will display the abstract with the relevant information (phospho-protein, phospho-site, interactant, impact words) marked in text. Often, the information can be found in one sentence alone, but there are situations when the information could span multiple sentences.

The user interacts with the system both at the input and output stages. If the input is a protein name, the system interacts with the user by providing all the proteins that match the specified name, and it then asks the user to select the correct protein from the list. Once the results are displayed, the user is able to tell the system which abstracts (see Figure 1) and which sentences within these abstracts (see Figure 2) are correctly identified with a phosphorylation-PPI relation. Corrections can be provided for both the phosphorylation mention and the PPI mention. Moreover, evidence can be added for sentences that eFIP missed to identify.

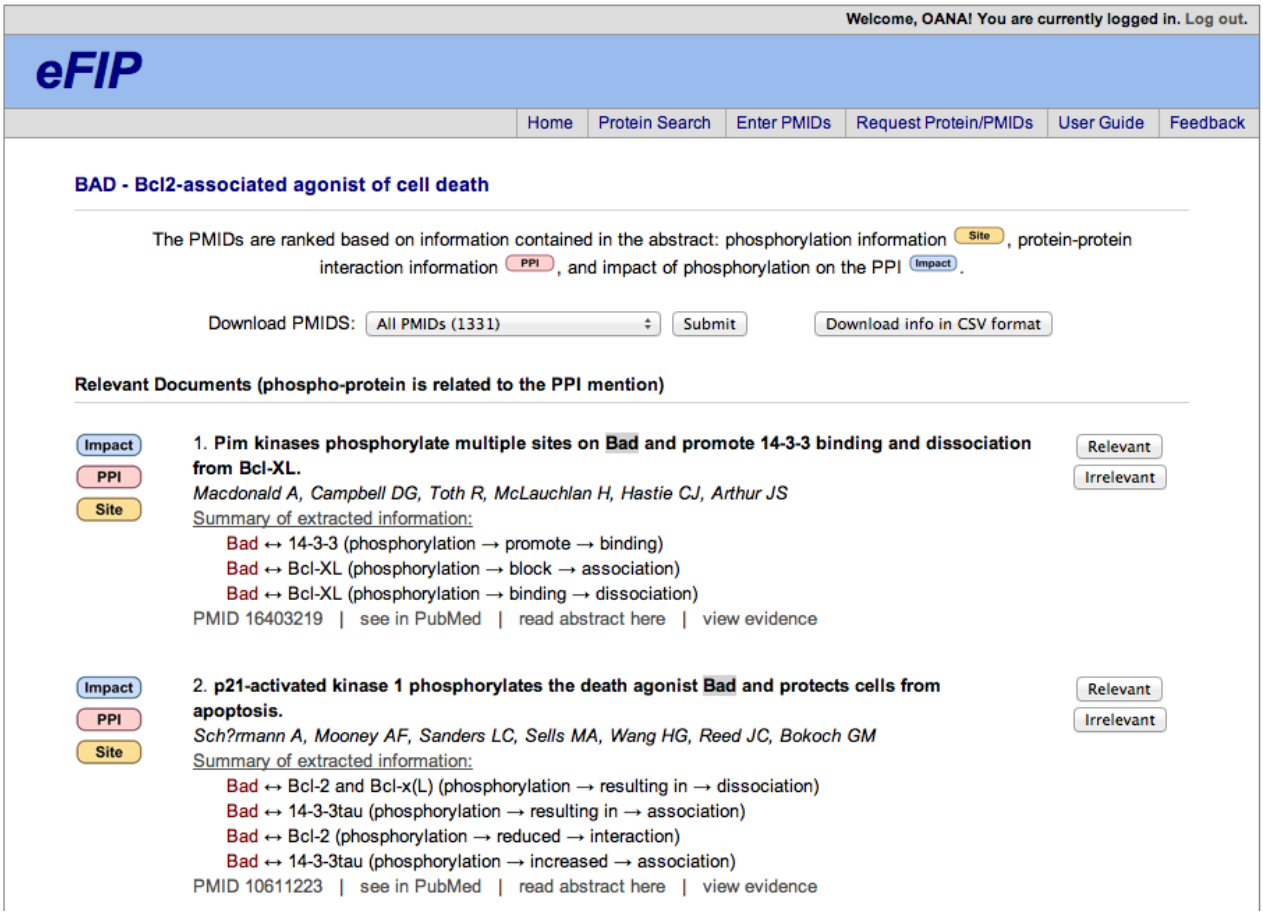


Figure 1. Example output of ranked PMIDs for protein BAD.

The user can choose to save the information in a .csv file. This information can be saved for an individual PMID or a group of PMIDs, and will include the latest corrections provided by the user. The interface shown in Figures 1 a nd 2 w ill also be used to evaluate eFIP at the BioCreative Biocuration Task. Biocurators will be asked to annotate abstracts containing phosphorylation and PPI mentions, both by using eFIP (to assess its usefulness) and by manually recording data in a spreadsheet (for a gold standard used to assess the performance of eFIP).

Figure 2. Example output of information extracted from PMID 11583580.

Evaluation and Discussion

We conducted two evaluations, one using eFIP's annotation interface for 90 randomly selected abstracts, and the second using a gold standard set provided for 151 randomly selected sentences containing trigger words for phosphorylation and PPI.

First Evaluation. We used the annotation interface as shown in Figures 1 and 2. Two annotators were asked to mark the relevance of 90 randomly selected abstracts, as well as the relevance of the evidence displayed in eFIP for each of these abstracts. Of the 90 abstracts, 64 were marked as true positives, 3 as false positives, 21 as true negatives, and 2 as false negatives. Thus, we report a precision of 95.52% and recall of 96.96% at the abstract level. Of the 135 evidence sentences shown to the annotators, 114 were marked as true positives and 21 as false positives. Thus, we report a precision of 84.44% at the sentence level. Additionally, there were 18 cases indicated by the annotators to be false negatives, yielding a recall of 86.36% at the sentence level.

Second Evaluation. We also wanted to see how well eFIP performs when compared with a gold standard set created by the annotators without any influence from eFIP. For this, we randomly selected a set of 151 sentences containing trigger words for phosphorylation and PPI, and asked the same two annotators to mark their relevance. Using this set of annotated examples, eFIP yielded 86.11% precision, 60.78% recall, and 83.44% accuracy.

Discussion. Some cases missed by eFIP are due to shortcomings in the detection of phosphorylation. For example, no phospho-protein is found in the following sentence: "Such phosphorylation attenuated its association with SPHK1a". It is not obvious if "its" refers to the phospho-protein or some other protein mentioned in a previous sentence. Other cases are due to shortcomings in the detection of proteins involved in the PPI. For example, consider the following sentence "Phosphorylated BAD was sequestered in the cytosol bound to 14-3-3". The PPI tool could not detect that it is "BAD" that is "bound to 14-3-3". A low recall in the second evaluation is attributed to our strict rules for detecting PPI mentions, as well as shortcomings in detecting phosphorylation mentions when the sentence contains a complex grammatical structure. In the future, we plan to bring additional rules for detecting PPI mentions, as well as use a sentence simplifier to ensure the matching of the patterns used in the detection of phosphorylation.

References

1. Arighi, C.N., Siu, A.Y., Tudor, C.O., Nchoutmboube, J.A., Wu, C.H., and Shanker, V.K. (2011) eFIP: A Tool for Mining Functional Impact of Phosphorylation from Literature. *Bioinformatics for Comparative Proteomics,* Methods in Molecular Biology, vol. 694, 63-75.

2. Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K., and Wu, C.H. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21, 2759-2765.

3. Narayanaswamy, M., Ravikumar, K.E., and Vijay-Shanker, K. (2005) Beyond the clause: extraction of phosphorylation information from Medline abstracts. *Bioinformatics* 21 Suppl 1, i319-i327.

4. Natale, D.A., Arighi, C.N., Barker, W.C., Bult, C.J., Caudy, M., et al. (2011) The Pro Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Research* 39, D539-45.

5. Tudor, C.O., Schmidt, C.J., Vijay-Shanker, K. (2010) eGIFT: Mining Gene Information from the Literature. *BMC Bioinfomatics*. vol. 11, 418.

Searching of Information about Protein Acetylation System

Cheng Sun*, Mengni Zhang, Yangwei Wu, Jiansheng Ren, Yunshen Bo, Lei Han, Donghong Ji

Department of Computer, Wuhan University, Wuhan, China

*Corresponding author: Cheng Sun Tel: +86 15927378425, E-mail: gensun.cc@gmail.com

Abstract

We construct an information retrieval system for acetylation. We collect items from Pubmed to make the corpus for the system. We rank the items by their relevance to the key word searched and re-rank the items considering their relevance to acetylation. The final ranking results are the comprehensive results based on the relevance to the key word searched and the relevance to acetylation.

1、Introduction

1.1 Motivation

In BC Workshop 12, a system for a specific biocuration task based on text mining or natural language technology is demanded. To a system of this kind, we are interested in acetylation. In our previous work and research, we did some studies about this subject but never developed a system for it.

1.2 Background

Heavy jobs have been done to acetylation. A great many academic papers have been published. Most of them are stored in database. However, there is no database, which specifically targets acetylation. In this situation, it is troublesome for people in this territory to locate the literature they need.

1.3 Our advantages

First, in our previous work, something relevant to natural language processing in the biological domain has been done. Second, we accumulated technology for general problems in natural language processing.

1.4 Feasibility

In our discussion, we found out there were some troubles in retrieving resources in the biological domain. With feedback from the people researching in the biological domain, we selected the final aim--acetylation. To prevent the barriers beyond us, we simplified the system into a level which is within our ability.

1.5 Characters

In the general database, various kinds of academic literatures are stored together. There is no database which is made specifically for acetylation. It is troublesome and time-consuming to retrieve the relevant literature. The system we will make is to cope with this trouble.

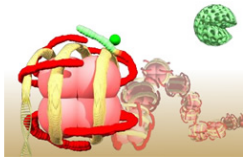
1.6 Corpus for the system

We search Pubmed with several key words which are relevant to acetylation and collect the search results from them. The detail is described in appendix one.

2、Interface

The interface is as follows. After one word is input and searched, the items the searching results will be ranked according their relevance to the key word. Each item of the results includes the title, authors, abstract and PubMedID. The full text is not available due to copyrights. The following figure shows the general interface, which can be accessed by http://nlp.whu.edu.cn:8080/track3/

Figure 1 General interface



Searching of Information about Protein Acetylation

Search

Copyright © wuhan university NLP lab

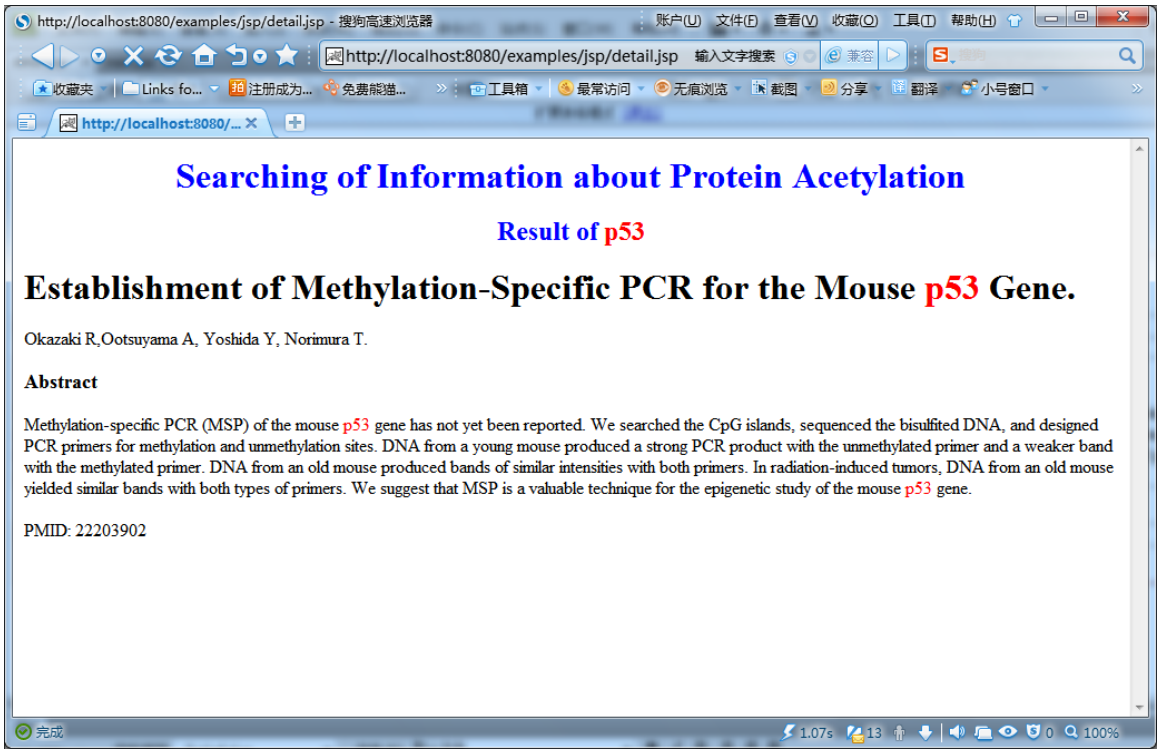
There is another interface which can be accessed by http://nlp.whu.edu.cn:8080/track3/preRerank.jsp

This interface is similar to the above one. But, the search results are different. The results are the results before reranking. We first rank the results without considering the features in biology. The re-ranking is made considering the features in biology especially the features relevant to acetylation,which is described in detail in section 3.

Fiture 2 The results for the key word p53

The following is one result item for the key word p53.

Figure 3 one result item



3、Working description

There are two parts in the program including the part for documents pre-processing and documents-retrieving.

3.1Pre-processing

First, the abstract of each paper is read and the document object is constructed. After that, the index, segmentation will be done. Following that, the score based on the relevance to acetylation will be given. Set the score as **A**, which will be recorded in the index file.

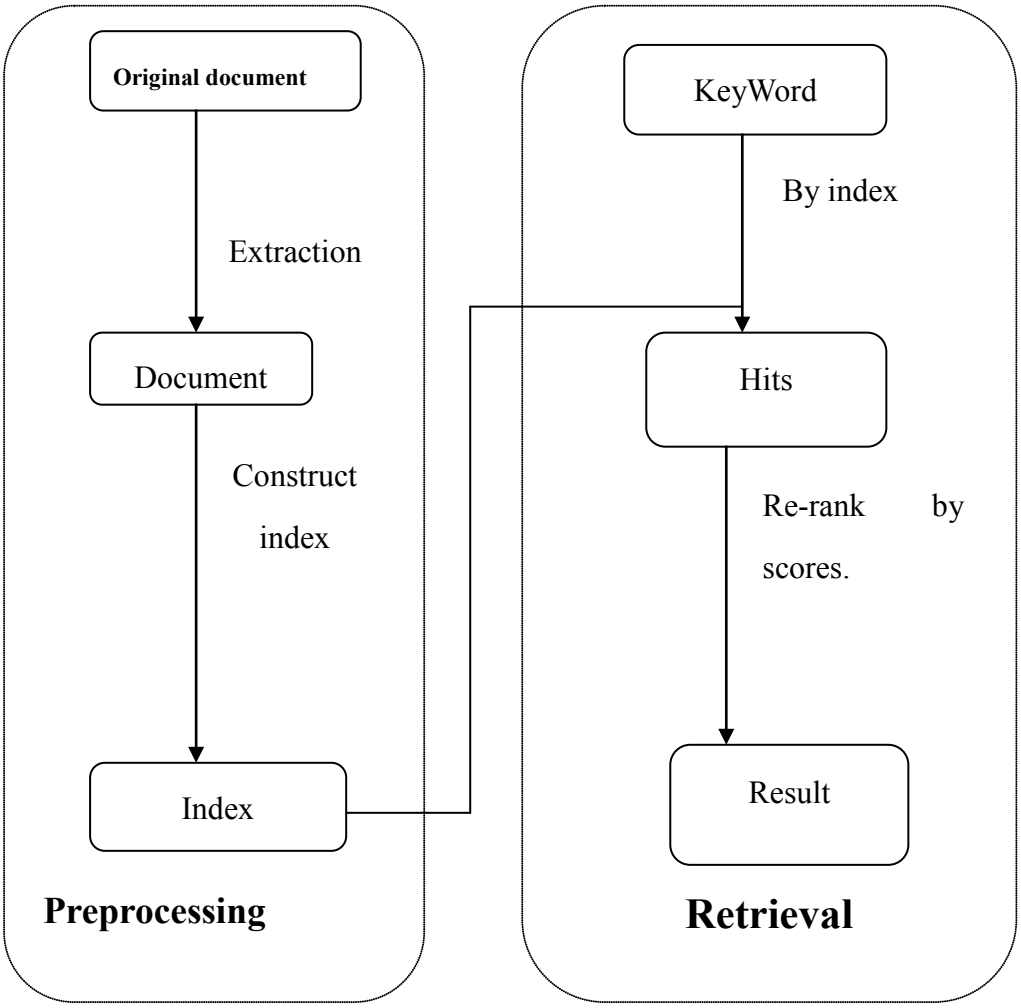
3.2 Retrieving

In this process, there are two steps. The first is search. The second is ranking.

Search: After the user inputs a keyword, the search for the key words based on the index will be done. The results will be stored in hit.

Ranking: Give the score to each doc according to the general search. The score is based on TF-IDF calculation. Set this score as **B**. The final score of the doc will be given as the combination of A and B .**S=A*x+B*y**（x，y is the weight of basic relevance and ranking respectively. After that, the output will be given in the descending order of the final scores.

Figure 4 Overall descriptions

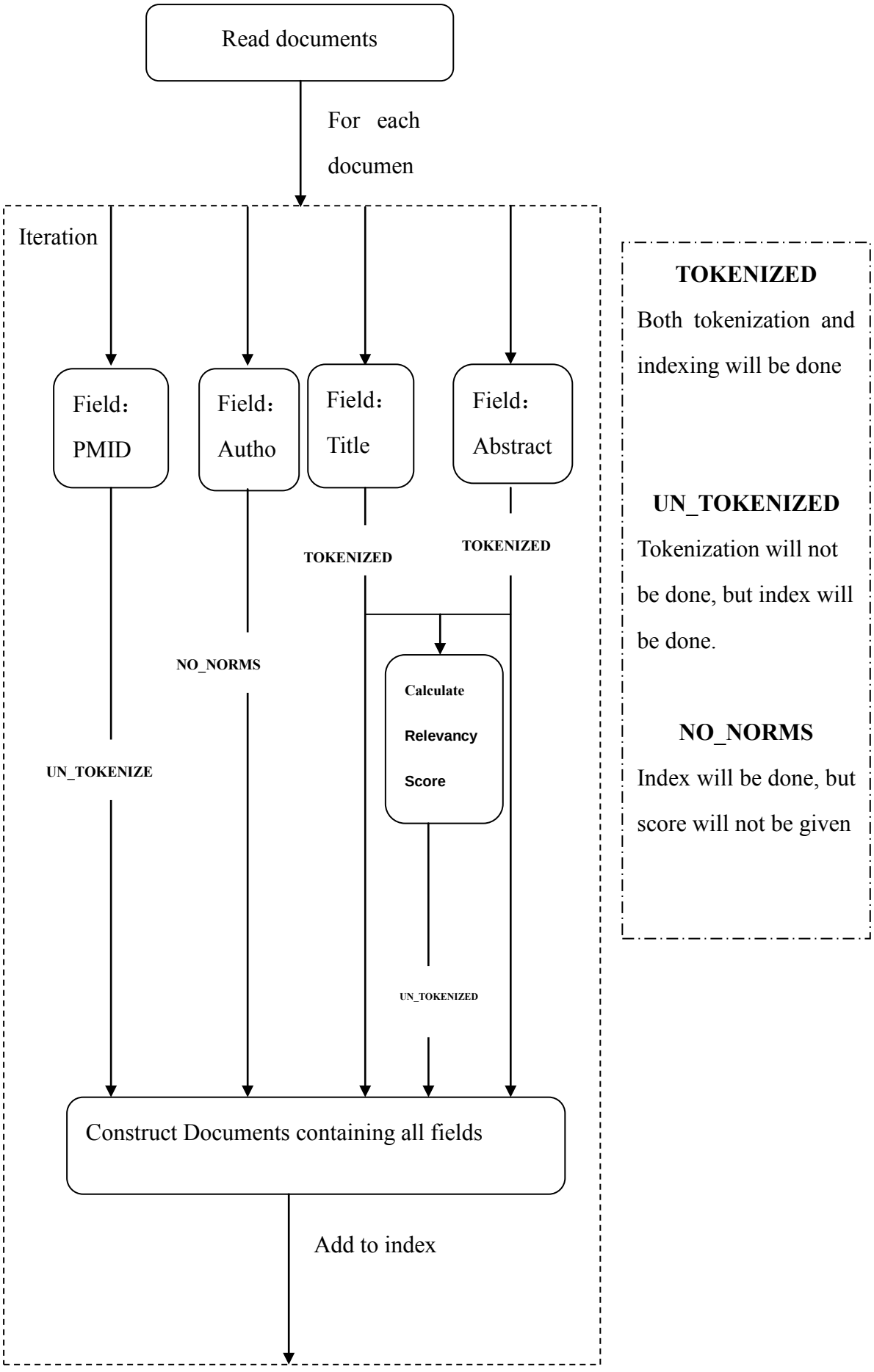


Specific Procedure：

Document pre-processing

First, the program fetches information about the documents and extracts PMID, Title, Abstract, Author, RelevancyScore, and construct a field to each of them, which represents the properties of the document. After that, the program will construct index with words.

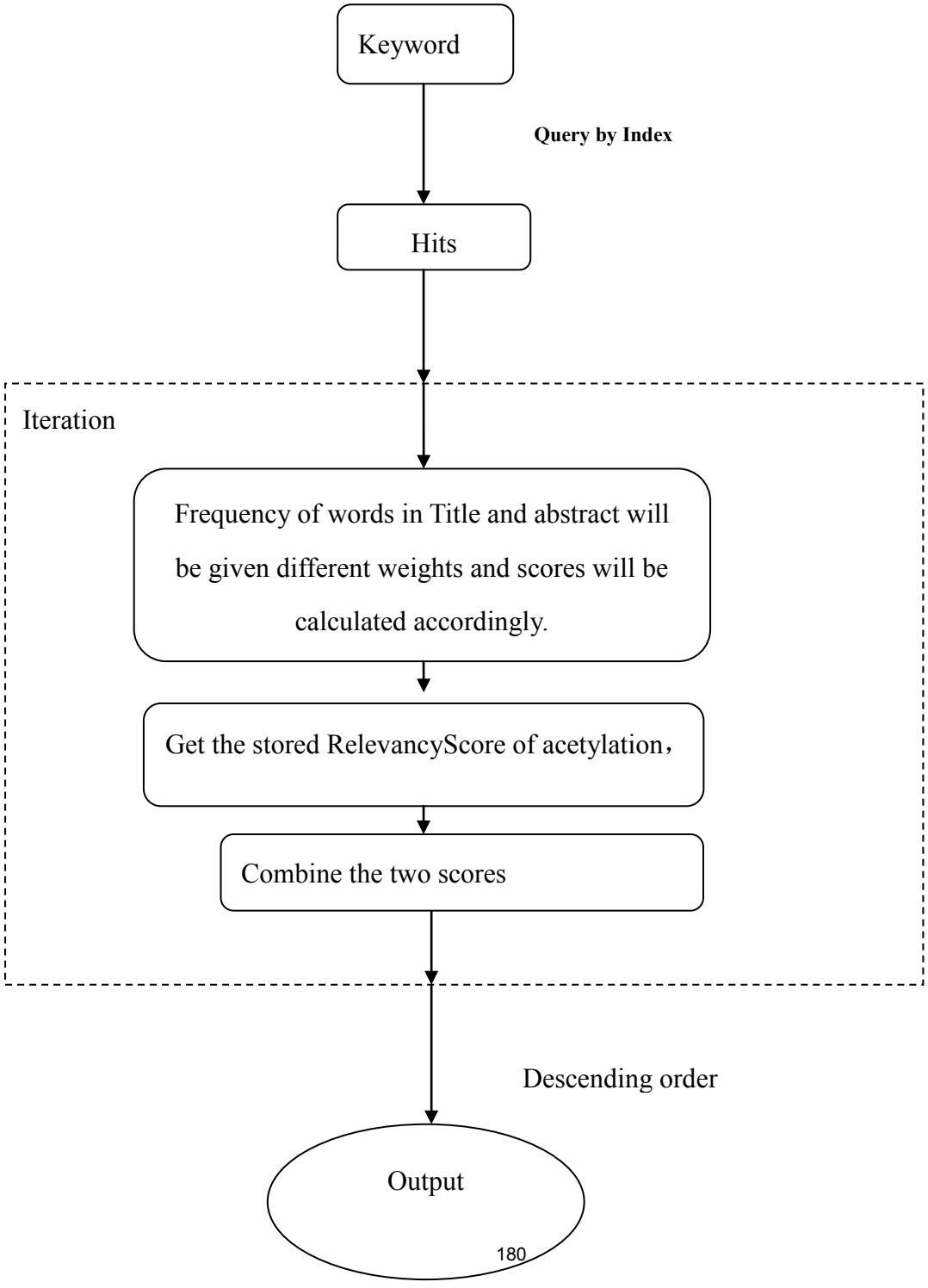
Figure 5 Specific procedure



Document Indexing

Based on indexing, the key words the user inputs will then be matched. In ranking, the weights of title and abstract are different. The score of the document will be calculated as the score of the document related to the keyword. After that, the weighted mean of RelevanceScore will be calculated to generate the overall scores of the documents. The output will listed in descending order.

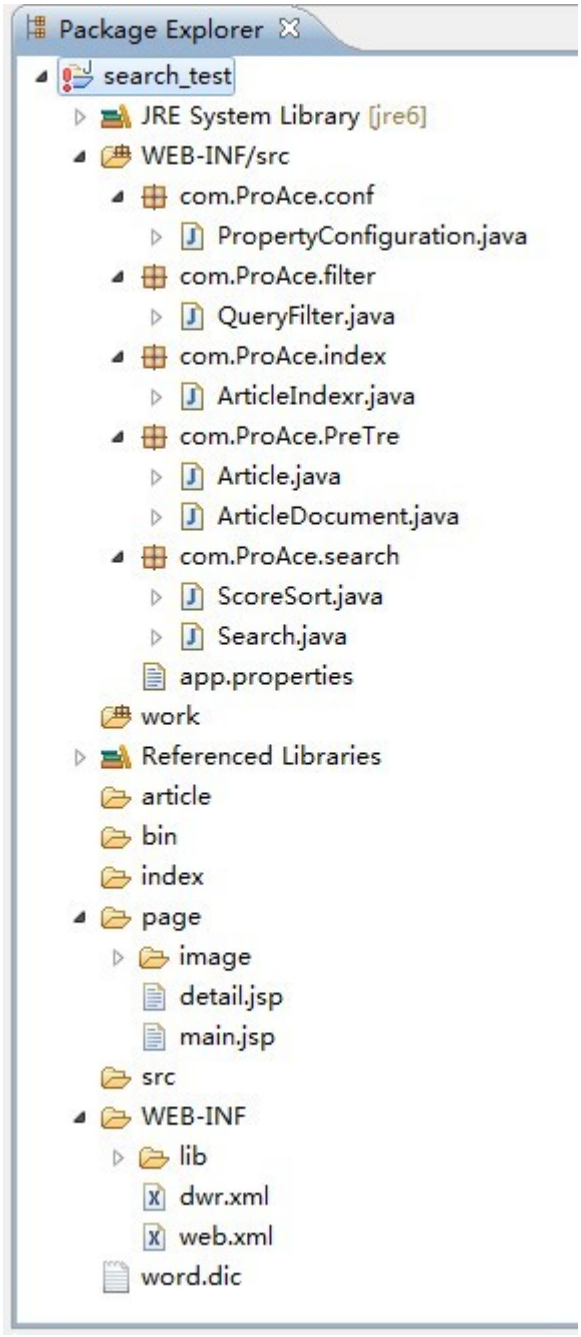
Figure 6 Document indexing



4、Program description

The structure of the project is as follows

Figure 7 Project structure



word.dic is used to store the key words and the relevance score.

The folder WEB-INF store configuration information and reference libs the whole project.

The folder page store s jsp pages.

WEB-INF/src stores java codes.

com.ProAce is the prefix of packages ,including all the codes of this project.

Filter is used to store the retrieving results and score it for the second time.

Index is the code for indexing.

PreTre contains the codes for pre-processing.

Search contains the codes relevant to searching.

app.properties stored some information in configuration, which will be read out by PropertyConfiguration.

Appendices

Appendix one Corpus for the system

The items collected for the system are as follows.

Key words used to collect items relevant to acetylation

Acetylation	198
acetytransterase	395
ARD1	84
ATP-citrate lyase	199
deacetylation	399
GNAT	172
HAT	198
HDAC	199
Histone	199
NAT	393
nsN-alpha-terminal Acetylation	303
NuA4	114
P53	394
PACF	14
regulation of gene expression	200
Rtt109	68
Rtt109_Vps75	11
SAGA	369
Sas2	60
SIRT1	392
SIRT2	160

SIRT3	158
SIRT4	40
SIRT5	41
SIRT6	62
SIRT7	32
TFIIB	198
TFIIE	198
TFIIF	199

The total items are 4840 with duplicated items being removed

Appendix two Ranking algorithm

Scoring algorithm for the documents of acetylation for a single word

For the relevance of key words to index and acetylation, we consider issues in two aspects. First, the relevance of each document and acetylation should be got from the database. Second, the relevance of the key word and the documents, which are relevant to acetylation, should be got.

To the relevance of a document and acetylation, we can get accurate value in pre-processing. The algorithm is as follows.

Scoring for the key word

We looked up to a great many resources relevant to acetylation and searched for the key words which may be relevant to acetylation. After that, select n key words which are close to acetylation. The number n will be set according to the specific experiment performance. According to the relevance between the document and acetylation, we set a rational value for each of the key word manually.

For acetylation, after we made the value table for the key words which are closely related to acetylation, after we studied some resources. This step is crucial. And it is also subjective. The values are due to adjustment based on the experimental results.

Table 1 Scores assignment for key words

No.	Key word	Score (ten as the bigger number)
1	GNAT	9.5
2	NAT	8.5
3	nsN-alpha-terminal Acetylation	8.5
4	HAT and HDAC	9.0
5	Histone acetytransterase ,HAT	9.0
6	acetylation	8.5
7	deacetylation	8.0
8	Hat1	7.8
9	PACF	6.8
10	Sas2	5.0
11	SIRT1~~SIRT7	6.5
12	ARD1	5.0
13	SAGA	7.6
14	NuA4	7.2
15	Rtt109	7.8
16	Rtt109_Vps75	8.2
17	P53	6.8

Suppose there are n key words $K_i\ (1 \le i \le n)$， the score of them is $V_i\ (1 \le i \le n)$。

We employed the scoring rule in BM25. We give scores to the document based on the score of key words to represent the relevance of a document and acetylation. The scoring algorithm is as follows.

$$SA(D_j)=\sum_{i=1}^{N}[V_i\cdot\frac{f(K_i,D_j)\cdot(k_1+1)}{f(K_i,D_j)+k_1\cdot(1-b_1+b_1\cdot\frac{|D_j|}{avgdl})}]$$

- $SA(D_j)$ represents the score of document D_j with respect to acetylation.

- $f(K_i,D_j)$ represents the frequency of existence of the ith key word in Document D_j • $|D_j|$ represents the length of document D_j（ie. The number of words ）

- $avgdl$ represent the average length of all the documents in the database

Another two parameters k_1 and b_1 are used to adjust the precision. The default values of them are $k_1=2, b_1=0.75$.

According to this, we will get the relevance of each document and acetylation. For the consistence of measurement, the relevance is normalized.

$$A(D_j)=\frac{SA(D_j)-\min\limits_{1\le k\le N}\{SA(D_k)\}}{\max\limits_{1\le k\le N}\{SA(D_k)\}-\min\limits_{1\le k\le N}\{SA(D_k)\}}$$

To the relevance of the words in the input of the user and the document, we employed BM25 algorithm and got the following scoring algorithm.

$$SB(q,D_j)=IDF(q)\cdot\frac{f(q,D_j)\cdot(k_2+1)}{f(q,D_j)+k_2\cdot(1-b_2+b_2\cdot\frac{|D_j|}{avgdl})} \qquad (*)$$

$$IDF(q)=\begin{cases}\log(\frac{N-n(q)+0.5}{n(q)+0.5}) & n(q)<0.5N\\ 0 & otherwise\end{cases}$$

- $SB(q, D_j)$ the score of the relevance of the query word and Document D_j.

- $f(q, D_j)$ the frequency of relevance of the query word in Document D_j

$|D_j|$ represent the length of Document D_j i.e. the number of words in D_j

- *avgdl* represents the average length of the documents in the entire database.

- k_2 and b_2 are used to adjust precision.

- $IDF(q)$ is inverse document frequency,

- N is the total number of documents in the database.

- $n(q)$ represents the number of documents, in which, the key word q exists.

We divide the whole document into three parts--Title abstract, main body. We gave scores to each of them and calculated the weighted sum.

For example: suppose the title of document D_j is D_{j1}, the abstract of document D_j is D_{j2}, the main body of document D_j is D_{j3}. The score will be given in the following way.

$$SB(q, D_{j1}) = \begin{cases} (k_3 + 1) \cdot IDF(q) & q \text{ exists in } D_{j1} \\ 0 & otherwise \end{cases}$$

$$SB(q, D_{j2}) = IDF(q) \cdot \frac{f(q, D_{j2}) \cdot (k_3 + 1)}{f(q, D_{j2}) + k_3 \cdot (1 - b_3 + b_3 \cdot \frac{|D_{j2}|}{avgdl_2})}$$

$$SB(q, D_{j3}) = IDF(q) \cdot \frac{f(q, D_{j3}) \cdot (k_3 + 1)}{f(q, D_{j3}) + k_3 \cdot (1 - b_3 + b_3 \cdot \frac{|D_{j3}|}{avgdl_3})}$$

k_3 and b_3 are used to adjust precision t he default value of them are $k_3 = 2, b_3 = 0.75$

Because the measurements of relevance are different the scoring process in Title, Abstract and main body, we need to normalize them. Set he normalized score of Title, Abstract and main body as $SB'(q,D_{j1})$、$SB'(q,D_{j2})$ $SB'(q,D_{j3})$ ；

The weighted average score of them are calculated as follows.

$$B(q,D_j)=\alpha\cdot SB'(q,D_{j1})+\beta\cdot SB'(q,D_{j2})+\gamma\cdot SB'(q,D_{j3})$$

This score is relevance score of the document and the query words.

$\alpha>\beta>\gamma>0$， they are the weights of the three different parts $\alpha+\beta+\gamma=1$。The default value is $\alpha=\frac{1}{2},\beta=\frac{1}{3},\gamma=\frac{1}{6}$.

Combining the two steps above, we get the relevance of each document and acetylation and the relevance of the key word and each document. According to two relevancies, we can get the weighted sum, and get the final score of relevance between the key word and the document with respect to acetylation.

$$Score(q,D_j)=x\cdot A(D_j)+y\cdot B(q,D_j)$$

In it $x,y>0$ and both x and y represent the weights