DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas049, doi:10.1093/database/bas049

Editorial

BioCreative-2012 Virtual Issue

Cathy H. Wu[1,*], Cecilia N. Arighi[1], Kevin B. Cohen[2], Lynette Hirschman[3], Martin Krallinger[4], Zhiyong Lu[5], Carolyn Mattingly[6], Alfonso Valencia[4], Thomas C. Wiegers[6] and W. John Wilbur[5]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA, [2]Center for Computational Pharmacology, University of Colorado Denver School of Medicine, Aurora, CO 80045, USA, [3]The MITRE Corporation, Bedford, MA 01730, USA, [4]Structural and Computational Biology Group, Spanish National Cancer Research Centre, Madrid E-28029, Spain, [5]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20892, USA and [6]Department of Biology, North Carolina State University, Raleigh, NC 27695, USA

***Corresponding author:** Tel: +1 302 831 8869; Fax: +1 302 831 4841; Email: wuc@udel.edu

BioCreative: Critical Assessment of Information Extraction in Biology is an international community-wide effort for evaluating text mining and information extraction systems applied to the biological domain (http://www.biocreative.org/). The Challenge Evaluations and the accompanying BioCreative Workshops bring together the text mining and biology communities to drive the development of text mining systems that can be integrated into the biocuration workflow and the knowledge discovery process. To address the current barriers in using text mining in biology, BioCreative has further been conducting user requirement analysis, user-based evaluations and fostering standard development for text mining tool re-use and integration. This *DATABASE* virtual issue captures the major results from the *BioCreative-2012 Workshop on Interactive Text Mining in the Biocuration Workflow* and is the fifth special issue devoted to BioCreative.

Built on the success of the previous BioCreative Challenge Evaluations and Workshops (BioCreative I, II, II.5 and III) (1–4), the BioCreative-2012 Workshop was held in Washington DC on 4–5 April 2012, in conjunction with the Fifth International Biocuration Conference (5). Since its inception, BioCreative has benefited from close collaborations between the community of text mining developers and curators of biological databases including GOA (6), IntAct (7), MINT (8) and BioGRID (9). These interactions have provided literature corpora with standard annotations for the evaluation of automated systems and have allowed better understanding of the underlying annotation process as well as characterization of particular tasks where text mining systems could play a role in improving the manual literature curation process.

Challenge Evaluation tasks over the years have included ranking of relevant documents (document triage), extraction of genes and proteins (gene mention) and their linkage to database identifiers (gene normalization), as well as extraction of functional annotation in standard ontologies (e.g. Gene Ontology (10)) and extraction of entity relations (e.g. protein–protein interaction). Some text mining tasks (e.g. gene normalization) are of fundamental importance to different applications, thus have been the subjects of multiple Challenge Evaluations to improve the system performance. New tasks are also introduced to address new applications, tackling new entities, relationships and functional attributes (e.g. drug and disease).

The initial BioCreative challenges provided a valuable analysis of tool performance on component tasks of the biocuration workflow and promoted the implementation of text mining applications and web services. Nevertheless, those systems were not evaluated within the dynamic process of database literature curation, where the end users have to interact with the systems in order to complete certain annotation tasks.

To address the utility and usability of text mining tools beyond formal offline evaluation metrics, BioCreative III introduced an Interactive Task as a demonstration task focusing on gene-based document retrieval. A major goal was to improve text mining systems for computer-assisted biocuration—to support a human database curator rather than serving as a replacement. BioCreative-2012 took the next step in this process, bringing together the biocuration and text mining communities to develop and evaluate interactive text mining tools and systems and improve utility and usability in the biocuration workflow. We wished to take advantage of the many databases containing various types of biological information derived from scientific literature, to study and understand in more detail how

Database, Vol. 2012, Article ID bas049, doi:10.1093/database/bas049

human experts process natural language (free text) and extract information into a structured database.

The BioCreative-2012 Workshop consisted of three tracks: (I) a collaborative biocuration-text mining development task for document prioritization for curation; (II) a biocuration workflow survey and analysis task and (III) an interactive text mining and user evaluation task. In total, BioCreative-2012 attracted nearly 50 teams who registered for participation in the three tracks, with close to half of the groups completing their tasks. Twenty teams were selected by the Organizing Committee to participate at the workshop, contributing seven text mining systems in Track I, seven database workflows in Track II and six text mining systems in Track III. Nearly, 80 participants attended the workshop, with an almost even split between biocurators and text mining developers. Also attending was the 10-member User Advisory Group with representatives from many biocuration groups, particularly model organism databases and from the pharmaceutical industry.

The 'Track I Triage' task (11) invited text mining teams to develop tools or systems to assist curators in the selection and ranking of articles rich in information about chemicals and associated data based on the curation paradigm for the Comparative Toxicogenomics Database (CTD), which captures chemical–gene–disease relationships. The CTD project was chosen as a source for the task 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. CTD, with its own fully automated text mining pipeline, also has significant experience in text mining research and development (12). In addition to evaluating and ranking each system based on (off-line) recall and precision, participating groups were asked to provide a web interface, which was evaluated in terms of utility and usability for integration into the CTD curation process. The results of Track I showed that development of effective document prioritization tools, along with a user-friendly web interface, requires a high degree of systems development and integration, as well as close interactions with biocurators. Even with a short time frame from call-for-participation to system evaluation, several teams successfully created new systems based on the CTD functional specifications that may have long-term application for CTD and may be adapted for other curated databases (11).

The 'Track II Biocuration Workflow' task (13) invited curation teams to describe their curation process and workflow, starting from its criteria for selection of articles for curation to its culmination in database entries. Although biomedical text mining is an active research field, few text mining applications have been integrated into production biocuration workflows (14). To close this gap, the curation teams were asked to address a list of issues important to text mining developers and to identify possible insertion points for text mining and information extraction tools. The workflow analysis of seven participating databases identified commonalities and differences across the workflows, the common ontologies and controlled vocabularies used and the current and desired uses of text mining for biocuration. The workshop participants further identified text mining aids for gene indexing, document triage and ontology terms annotation as those most desired by the biocurators (13).

The 'Track III Interactive Text Mining' task (15) featured demonstration and evaluation of interactive text mining systems, some of which are currently being used in biocuration workflows. In addition to system evaluation (measured as precision and recall on application-specific curated data sets), a user study was conducted by selected expert biocurators prior to the workshop that included time-to-completion on curation tasks and post-study surveys. System demonstrations during the workshop provided direct interactions between biocurators and system developers, allowing end users to highlight both the strengths and the current limitations of each system and to provide feedback for improving the system based on user experience. Track III attracted the participation of a diverse range of systems representative of various biocuration scenarios covering diverse text mining tasks of importance for database curation. The user evaluation showed that a number of systems were able to improve efficiency of curation by reducing the time-to-completion over manual curation and/or improve annotation accuracy. The user survey of ~40 biocurators further highlighted the importance of the system's ability to assist them in completing the desired biocuration task, reflecting that the utility of the system has the most influence on biocurators' overall experience, particularly once design and usability concerns are largely satisfied (15).

BioCreative-2012 provides the basis for the BioCreative IV Challenge, which will culminate in the BioCreative IV Workshop to be held in Washington DC in 2013. The User Advisory Group continues to provide guidance on BioCreative IV planning from the biocurator and researcher perspectives with insights from BioCreative-2012. The CTD triage task is being further developed as one of the tasks for the BioCreative IV Challenge. The commonalities and database-specific aspects of literature-based curation as well as insertion points for text mining to simplify manual curation identified from the workflow analyses are being exploited to develop a new Challenge around Gene Ontology (GO) curation—to advance the state of the art in assisting this highly important, common and time-consuming data curation step that is largely lacking support from text mining at present, due to the complexity of the task itself and the absence of training data needed for text mining development. The goal is to further improve system

performance on a text mining task of fundamental importance to all databases that involve functional annotation. The user study and lessons learned from the diverse text mining systems that participated in the interactive track will lead to improved evaluation metrics and functional and standards requirements for an interactive task in BioCreative IV. To facilitate the development of text mining systems and pipelines that can be tailored for biocuration needs of various databases, BioCreative IV will also continue the discussion on system interoperability initiated at the BioCreative III Workshop. In particular, we will attempt to improve and formalize the development of common standards for data formatting and software modules to promote reusability of text mining tools.

This *DATABASE* virtual issue includes overview papers describing the three Tracks in BioCreative-2012 as well as papers describing selected participating systems demonstrating significant contributions to biocuration. The text mining systems were selected based on performance, scientific advancements, innovation and significant impact, including their utility and usability as evaluated by biocurators. The biocuration workflows were selected based on the depth and breadth of workflow coverage and the identification of clearly defined insertion points with functional requirements for text mining tools and approaches. As the fifth special issue devoted to BioCreative, the publication of this virtual issue will inspire further community engagement and discussion towards the ultimate goal of developing text mining systems for computer-assisted biocuration and knowledge discovery.

References

1. Hirschman,L., Yeh,A., Blaschke,C. *et al*. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6** (Suppl. 1), S1.
2. Krallinger,M., Morgan,A., Smith,L. *et al*. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.*, **9** (Suppl. 2), S1.
3. Leitner,F., Mardis,S.A., Krallinger,M. *et al*. (2010) An Overview of BioCreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 385–399.
4. Arighi,C.N., Lu,Z., Krallinger,M. *et al*. (2011) Overview of the BioCreative III Workshop. *BMC Bioinformatics*, **12** (Suppl. 8), S1.
5. Gaudet,P. and Mazumder,R. (2012) Biocuration Virtual Issue 2012. *Database*, **2012**, doi: 10.1093/database/bas011.
6. Camon,E., Magrane,M., Barrell,D. *et al*. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
7. Aranda,B., Achuthan,P., Alam-Faruque,Y. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
8. Ceol,A., Chatr Aryamontri,A., Licata,L. *et al*. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
9. Breitkreutz,B.J., Stark,C., Reguly,T. *et al*. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.
10. GeneOntologyConsortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
11. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2012) Collaborative biocuration-text mining development task for document prioritization for curation. *Database*, Vol. 2012: article ID bas.
12. Wiegers,T.C., Davis,A.P., Cohen,K.B. *et al*. (2009) Text mining and manual curation of chemical–gene–disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*, **10**, 326.
13. Lu,Z. and Hirschman,L. (2012) Biocuration Workflows and Text Mining: Overview of the BioCreative 2012 Workshop Track II. *Database*, Vol. 2012: article ID bas.
14. Hirschman,L., Burns,G.A., Krallinger,M. *et al*. (2012) Text mining for the biocuration workflow. *Database*, **2012**, doi: 10.1093/database/bas020.
15. Arighi,C.N., Carterette,B., Cohen,K.B. *et al*. (2012) An overview of the BioCreative 2012 Workshop Track III: Interactive Text Mining Task. *Database*, Vol. 2012: article ID bas.