DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

Original article

An overview of the BioCreative 2012 Workshop Track III: interactive text mining task

Cecilia N. Arighi[1,2,*], Ben Carterette[2], K. Bretonnel Cohen[3], Martin Krallinger[4], W. John Wilbur[5], Petra Fey[6], Robert Dodson[6], Laurel Cooper[7], Ceri E. Van Slyke[8], Wasila Dahdul[9], Paula Mabee[9], Donghui Li[10], Bethany Harris[5], Marc Gillespie[11], Silvia Jimenez[12], Phoebe Roberts[13], Lisa Matthews[14], Kevin Becker[15], Harold Drabkin[16], Susan Bello[16], Luana Licata[17], Andrew Chatr-aryamontri[18], Mary L. Schaeffer[19], Julie Park[20], Melissa Haendel[21], Kimberly Van Auken[22], Yuling Li[22], Juancarlos Chan[22], Hans-Michael Muller[22], Hong Cui[23], James P. Balhoff[24,25], Johnny Chi-Yang Wu[26], Zhiyong Lu[5], Chih-Hsuan Wei[5], Catalina O. Tudor[1,2], Kalpana Raja[27], Suresh Subramani[27], Jeyakumar Natarajan[27], Juan Miguel Cejuela[28], Pratibha Dubey[1] and Cathy Wu[1,2]

[1]Center for Bioinformatics and Computational Biology, [2]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19711, USA, [3]Center for Computational Pharmacology, University of Colorado Denver School of Medicine, Aurora, CO 80045, 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, [6]dictyBase, Center for Genetic Medicine, Northwestern University, Chicago, IL 60611, USA, [7]The Plant Ontology, Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331-2902, USA, [8]ZFIN, University of Oregon, Eugene, OR 97403-5291, USA, [9]Department of Biology, University of South Dakota, Vermillion, SD 57069, USA, [10]Department of Plant Biology, The Arabidopsis Information Resource (TAIR), Carnegie Institution for Science, Stanford, CA 94305, USA, [11]College of Pharmacy and Allied Health Professions, St. John's University, Queens, NY 11439, USA, [12]Merck Serono, Geneva CH-1211, Switzerland, [13]Pfizer, Inc., Boston, MA 02134, USA, [14]Department of Biochemistry, NYU School of Medicine, New York, NY 10016, USA, [15]National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA, [16]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME 04609, USA, [17]Department of Biology, University of Rome Tor Vergata, Rome 00133, Italy, [18]Institute for Research in Cancer and Immunology, Universite de Montreal, Quebec H3C 3J7, Canada, [19]USDA-ARS Plant Genetics Research Unit and Division of Plant Sciences, Department of Agronomy, University of Missouri, Columbia, MO 65211, USA, [20]SGD, Stanford University, Stanford, CA 94305, USA, [21]Oregon Health and Science University, Portland, OR 97239, USA, [22]WormBase, Textpresso, California Institute of Technology, Division of Biology 156-29, Pasadena, CA 91125, USA, [23]School of Information Resources and Library Science, University of Arizona, Tucson, AZ 85719, USA, [24]National Evolutionary Synthesis Center, Durham, NC 27705-4667, USA, [25]Department of Biology, University of North Carolina at Chapel Hill, NC 27599-3280, USA, [26]Institute of Information Science, Academia Sinica, Taipei, Taiwan 115, Republic of China, [27]Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India and [28]Department of Bioinformatics and Computational Biology, Technical University Munich, Garching/Munich 85748, Germany

***Corresponding author:** Tel: +1 302 831 3444; Fax: +1 302 831 4841; Email: arighi@dbi.udel.edu

Citation details: Arighi,C.N., Carterette,B., Cohen,K.B., *et al*. An overview of the BioCreative 2012 Workshop Track III: interactive text mining. *Database* (2012) Vol. 2012: article ID bas056; doi:10.1093/database/bas056

Submitted 10 July 2012; Revised 27 November 2012; Accepted 28 November 2012

In many databases, biocuration primarily involves literature curation, which usually involves retrieving relevant articles, extracting information that will translate into annotations and identifying new incoming literature. As the volume of biological literature increases, the use of text mining to assist in biocuration becomes increasingly relevant. 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. Some biocuration efforts already make use of a text mining tool, but there have not been many broad-based systematic efforts 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 text mining tool developers and database biocurators to test the utility and usability of tools. Six text mining systems presenting diverse biocuration tasks participated

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

in a formal evaluation, and appropriate biocurators were recruited for testing. The performance results from this evaluation indicate that some of the systems were able to improve efficiency of curation by speeding up the curation task significantly (~1.7- to 2.5-fold) over manual curation. In addition, some of the systems were able to improve annotation accuracy when compared with the performance on the manually curated set. In terms of inter-annotator agreement, the factors that contributed to significant differences for some of the systems included the expertise of the biocurator on the given curation task, the inherent difficulty of the curation and attention to annotation guidelines. After the task, annotators were asked to complete a survey to help identify strengths and weaknesses of the various systems. The analysis of this survey highlights how important task completion is to the biocurators' overall experience of a system, regardless of the system's high score on design, learnability and usability. In addition, strategies to refine the annotation guidelines and systems documentation, to adapt the tools to the needs and query types the end user might have and to evaluate performance in terms of efficiency, user interface, result export and traditional evaluation metrics have been analyzed during this task. This analysis will help to plan for a more intense study in BioCreative IV.

Introduction

Biological databases are an integral part of the tool set that researchers use on a daily basis for their work as they serve to collect and provide access to our expanding knowledge of biology. They enable a more systematic access to information that otherwise would be buried in unstructured text, facilitating programmatic analysis of biological datasets. Database biocuration is a key activity to provide high-quality information. It could be defined as the analysis, interpretation and integration of biological information, primarily to add value by annotating and interconnecting research data and results within a common biological framework (1). To achieve this, expert biocurators may need to read and extract relevant information from the biomedical literature. This literature curation presents a considerable bottleneck in the biocuration process both in terms of speed (efficiency) and cost (biocurator's time); however, text mining tools have the potential to speed up the curation process if they perform useful tasks with sufficient accuracy and speed (2). Hirschman *et al*. (2) conducted a survey, among a group of 30 biocurators representing 23 databases, which identified some biocurator priorities and showed that two-thirds of the biocuration teams had experimented with text mining and almost half were using text mining in some aspect of curation. Biocurators required tools that were simple to use, easy to install and straightforward to maintain by the intended end user. Rather than providing high-performance tools in terms of formal evaluation scores, biocurators were more concerned with practical aspects that can assist the biocuration process. Some of these aspects included the request of producing ranked results and confidence scores, linking of automatically extracted annotations to evidence passages in the text, providing visualization aids (such as highlighting different levels of annotations) and allowing flexible export of results in standard formats (2). With these needs in mind, BioCreative (Critical Assessment of Information Extraction in Biology, http://www.biocreative.org/) (3–6), whose aim is to promote the development of text mining and text processing tools that are useful to the communities of researchers and biocurators, introduced an interactive task (IAT) in BioCreative III (7). A critical aspect of BioCreative III was the active involvement of a representative group of end users to guide development and evaluation of useful tools and standards. The IAT, although demonstrative, fostered the interaction of developers and biocurators and inspired the development/improvement of interfaces that can be used in a biocuration workflow (8). The positive reception of this task by both the text mining and the biocuration communities prompted the organization of the BioCreative 2012 workshop, centered on interactive text mining in the biocuration workflow. In particular, the work presented in this article was an interactive text mining and user evaluation task. Like the BioCreative III IAT (7), it was non-competitive, and the goals were to engage users, provide the means to experiment with different approaches to formally assess interactive systems as well as to collect specifications and metrics that will set the stage for the BioCreative IV challenge to be held in October 2013. Hosting the workshop as a satellite to the International Biocuration meeting provided a unique opportunity to engage biocurators in this activity.

Lessons learned from BioCreative III IAT

In the BioCreative III IAT, the goal was to develop an interactive system to facilitate manual annotation of unique database identifiers for all genes appearing in an article. This task included ranking genes by importance (based preferably on the amount of described experimental information regarding genes) (7). There was also an optional task to assist the user in retrieving the most relevant articles for a given gene. To aid in carefully designing this task, a user advisory group (UAG; http://www.biocreative.org/about/biocreative-iii/UAG/) was assembled that played an active role in assessing IAT systems and in providing a

detailed guidance for a future, more rigorous evaluation of IAT systems (7).

Some important lessons learned from this activity include the following: (i) early team-up of developers with biocurators is important to work together throughout the process of system development; (ii) sufficient time is needed for system training; (iii) selection of a corpus that is relevant to the users domain of expertise (such as species-specific documents for model organism databases and pathway-centric documents for pathway databases) and (iv) encouragement of text mining developer participation in biocuration meetings to facilitate interaction with biocurators. As observed in the biocurators survey, a users' adoption of automated tools into their curation process will depend heavily on performance and on the overall convenience of a tool.

Built upon these observations, we designed the BioCreative 2012 workshop interactive track described here.

Materials and methods

This section provides an outline of the BioCreative 2012 IAT planning, starting with modifications from the previous BioCreative IAT, the recruitment of participants and coordinators, preparation of datasets and the evaluation. Figure 1 summarizes the workflow of the BioCreative 2012 IAT activity, divided into three main phases: preparation, training and evaluation and indicating the tasks performed by teams, biocurators and coordinators, along with some important dates. Some of the details are described as follows.

IAT in BioCreative 2012 workshop

Based on the considerations brought up by the UAG in BioCreative III, we introduced some modifications to the IAT in the BioCreative 2012 workshop, such as

(i) Teams presented documentation for their systems, curation guidelines when needed, a practice set for biocurators and benchmarking of the system previous to the evaluation. This was to ensure the tools' performance and scope would be adequate for the proposed biocuration task.
(ii) The systems could include any biocuration task as opposed to BioCreative III, which was limited to gene normalization/ranking. Biocurators with experience in the relevant biocuration tasks were recruited and paired with developers early in the process. This interaction allowed systems to be tuned to the user's curation interests to make results more relevant to them.
(iii) The period for a biocurator's training on and evaluation of a system was significantly extended (from 10 to 20 days) in comparison with BioCreative III.
(iv) The BioCreative 2012 workshop was hosted as a satellite to the International Biocuration meeting to encourage participation of text mining developers in the biocuration meeting as well as participation of biocurators in the BioCreative workshop.

Recruitment of participants

Text mining teams. We openly invited text mining teams to participate in the IAT by presenting systems that focused on any given biocuration task. Registered teams were requested to submit a document describing their system and addressing questions related to relevance and impact of the system, adaptability, interactivity and performance. In addition, teams were asked to indicate the limitations of the system, provide details on the biocuration task and suggest evaluation metrics. Each system was assigned a coordinator to supervise and assist in the activity (see 'Coordinators' section). The list of systems with brief description (Table 1) and the accompanying documentation were posted on the BioCreative website (http://www.biocreative.org/tasks/bc-workshop-2012/track-iii-systems/) for biocurators to select and sign up for testing.

Biocurators. We invited biocurators to participate in the BioCreative IAT by distributing the call for participation via the International Society for Biocuration (ISB) mailing list, and the ISB meeting and BioCreative websites. Biocurators had the option to participate at different levels, namely, by assisting in selecting and annotating datasets to create the gold standards, by participating in the pre-workshop evaluation of a system of their choice based on the list provided in Table 1, and/or by participating in the workshop. Around 40 biocurators participated in this activity, Table 2 shows the wide variety of databases/institutions they represented and the different participation level (dataset annotations and system evaluations).

Coordinators. Coordinators were members of the BioCreative 2012 workshop steering committee who assisted in supervising and facilitating the communication between biocurators and developers. Some of the roles of the coordinators included the following: (i) matching and introducing biocurators 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 biocurators on the steps needed to complete evaluation and (vi) collecting metrics.

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

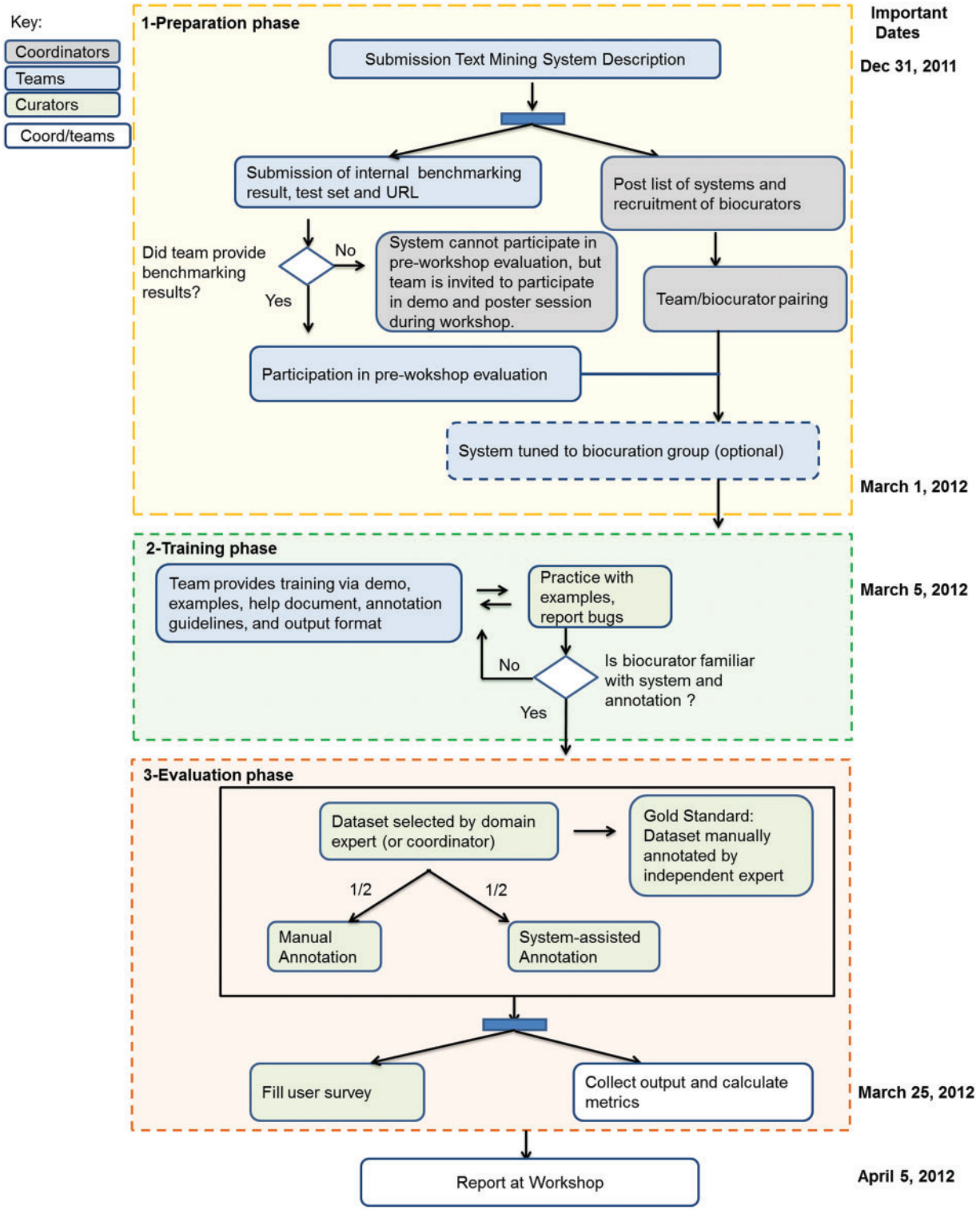

Figure 1. BioCreative 2012 workshop workflow. The chart shows the three main phases for this activity: (1) the preparation phase included the system and document preparation by teams, recruitment of biocurators to test each system and modification of the system for the assigned biocuration group; (2) the training phase actively involved both teams and biocurators, the former to provide the necessary support to use the system, the latter to learn about the curation task and the system functionalities, reporting system's bugs when necessary and (3) the evaluation phase included the selection of corpus and manual annotation by expert (to create gold standard), annotation of this corpus by biocurators, half manually and half system-assisted, along with time recording and filling of the user survey. The results were collected by teams and coordinators and presented at the workshop. Some important dates are indicated on the right side.

Table 1. Systems registered in BioCreative 2012 Track III

System	Biocuration tasks	Reported benchmark			
		Recall (%)	Precision (%)	F-measure	Corpus
Textpresso (14, 15)	Task: information retrieval (document classification), and information extraction (name entity recognition, ontology matching) Document: full-length articles	4-Category search			55 full-text articles from *Caenorhabditis elegans*, 43 of which contained information about subcellular localization
		Document level			
	Use: curation of subcellular localization using Gene Ontology cellular component URL: http://www.textpresso.org/dicty/	79.1	61.8	69.5	
		Sentence level			146 sentences (for precision) and 386 sentences (for recall) about subcellular localization derived from 60 articles 45 annotations for precision and 68 possible annotations for recall.
	Task for BioCreative: classification of articles as relevant or non-relevant for annotation of subcellular localization, identification of evidence sentences and derived GO annotation	30.3	80.1	44	
		GO annotation level			
		65.7	66.2	78.8	
PCS (CharaParser) (16)	Task: information extraction (name entity recognition, ontology matching) Document: textual description of phenotypic characters (which includes the organism's observable characteristics or traits) in NeXML format (33)	Term-based Entity-Quality (EQ)[a]			100 randomly selected phenotypic character descriptions curated by Phenoscape curators
	Use: curation of EQ terms from phylogenetic literature using ontologies URL: not web-based	90.0	90.0	90.0	
		Label-based EQ[a]			
	Task for BioCreative: description of a list of phenotypical characters as EQ statements, and use ontology terms (from TAO and PATO) to formalize these statements	51.0	52.0	51.5	
PubTator (18)	Task: information retrieval (document ranking) and information extraction (named entity recognition and normalization) Document: abstract-based	Gene mention module (abstract)			Reported in GeneTUKit (19)
		82.4	86.7	84.5	
	Use: document triage (relevant documents for curation) and bioconcept annotation (gene, disease, chemical) URL: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/	Gene normalization module (full text)			Reported in GenNorm (20)
		39.7	56.2	46.6	
	Task for BioCreative: classification of articles as relevant for TAIR curation, both TAIR and NLM identification gene mentions in abstracts and link them to Entrez gene	Species recognition module (abstract)			Reported in SR4GN (21)
		85.4	85.4	85.4	

(Continued)

Table 1. Continued.

System	Biocuration tasks	Reported benchmark			
		Recall (%)	Precision (%)	F-measure	Corpus
PPInterFinder (22)	Task: information retrieval (document classification) and information extraction (interaction pair and relationship) Document: text in Medline or XML format Use: mining of protein–protein interactions (PPIs) for human proteins URL: http://biomining-bu.in/ppinterfinder/ Task for BioCreative: classification of articles as relevant for human PPI, and identification of gene pairs in abstracts and link them to Entrez gene	Sentence level With PPI algorithm and gene mention/ normalization 71.3	 81.3	 75.9	693 sentences related to human PPIs from IntAct Database
eFIP (23)	Task: information retrieval (document classification) and information extraction (phospho-protein, interaction pair, relation and effect) Document: abstract-based Use: mining protein interactions of phosphorylated proteins from the literature. Extraction of phosphorylated proteins, protein-binding partners and impact keywords URL: http://proteininformationresource.org/pirwww/iprolink/eFIP.shtml Task for BioCreative: classification of articles as relevant for PPIs of phosphorylated proteins, and identification of phospho-protein, the interacting protein, the relation and effect	Abstract level 71.1 Sentence level 72.4	 86.5 67.9	 78.0 70.1	96 randomly selected PubMed abstracts containing sentences with trigger words for both phosphorylation and interaction mentions. This set contains 148 eFIP positive sentences.
T-HOD (27)	Task: information retrieval and information extraction (named entity recognition and relation) Document: abstract-based Use: document triage for disease-related genes (relevant documents for curation) and bioconcept annotation (gene, disease and relation). URL: http://bws.iis.sinica.edu.tw/THOD/ Task for BioCreative: classification of articles as relevant to a gene and a disease. Identification of genes related to disease in abstract	Document level 76.3	 77.1	 76.7	608 documents compiled from T1Dbase and T2Dbase for diabetes, genetic association database (GAD) for hypertension, and the gene list from 'The Human Obesity Gene Map: The 2005 Update' for obesity.
Tagtog[b]	Task: information extraction (named entity recognition) Document: abstract-based Use: Protein/gene mention recognition via interactive learning and annotation framework. URL: http://www.tagtog.net/ Task for BioCreative: given a set of documents annotate gene mentions	No metrics for this system			

System descriptions with task proposed in BioCreative and reported internal benchmark results. [a]Term-label EQs are EQ statements created strictly based on the original descriptions, independent of any ontologies, whereas the label-based EQs are the corresponding formal statements (using ontology terms). [b]This system only participated at the workshop.

Table 2. Participating databases/institutions in BioCreative Workshop 2012

Database/Institution type	Database/Institution	Gold standard annotation	Pre-workshop Evaluation	Workshop evaluation
Industry	AstraZeneca (1)			✓
	Merck Serono (1)		✓	✓
	Pfizer (1)		✓	✓
Literature	NLM (1)		✓	✓
Model Organism (MOD)/Gene Ontology Consortium (GOC)	AgBase (1)			✓
	dictyBase (2)	✓	✓	✓
	FlyBase (1)			✓
	MaizeDB (1)			✓
	MGI (3)		✓	✓
	SGD (1)		✓	✓
	TAIR (2)	✓	✓	✓
	WormBase (1)			✓
	XenBase (1)			✓
	ZFIN (1)		✓	
Ontology	Plant ontology (1)		✓	✓
	Protein ontology (2)	✓		
Pathway	Reactome (2)		✓	
Phenotype	GAD (1)		✓	
	Phenoscape (3)	✓	✓	✓
Protein–protein interaction	BioGrid (1)		✓	✓
	MINT (1)		✓	✓
	Others (approx. 11)			✓

Numbers in parentheses are the number of biocurators from each institution. Biocurators aided in dataset annotations and system evaluations

Datasets

The selection of suitable data collections for the evaluation was inspired by real curation tasks as well as keeping in mind the biocuration workflows. Each system had its own dataset that was selected by its coordinators and the domain experts that were involved in the annotation of the gold standard. In most cases, the dataset consisted of a collection of 50 PubMed abstracts randomly selected from a pool of possible relevant articles. A summary of the dataset selection and information captured is presented in Table 3. Note that the format of an annotated corpus varied depending on the system's output. This table also shows groups involved in the annotation of such corpora, and those who in the end evaluated the systems.

Evaluation

We planned two evaluations, a pre-workshop formal evaluation of the systems based on the selected corpus that included both systems' performance and subjective measures (explained later) and an informal evaluation consisting of the systems' testing at the workshop during the demonstration (demo) session. The latter included only the subjective measure representing mostly the user's first impression of a system.

Performance and usability of systems were calculated based on the following metrics:

As 'performance measures' we included comparison of time on task for system-assisted versus manual curation; and a precision/recall/F-measure of the automatic system versus the gold standard annotations (dataset independently manually curated by domain expert) and/or manual versus system-assisted annotations again rated by the gold standard.

We define these measures as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, FP and FN are true positives, false positives and false negatives, respectively.

For the 'subjective measure' we prepared a survey meant to record the subjective experience of the user with the system. The survey consisted of five main categories, namely, overall reaction, system's ability to help complete tasks, design of application, learning to use the application and usability, in addition to 'recommendation of the system' that was evaluated separately; these categories were based on those developed for the Questionnaire for User Interface Satisfaction (QUIS) developed by Chin *et al.* and shown to be a reliable guide to understanding user reactions (9). Each category contained questions to be rated

Table 3. Dataset preparation for systems in BioCreative Workshop 2012

System	Dataset selection for pre-workshop evaluation	Information captured	Biocurators involved in gold standard annotation	Biocurators involved in annotation in evaluation
Textpresso	30 full-length articles about *Dictyostelium discoideum* from 2011 to 2012 not yet annotated in dictyBase. This set contains 61 GO cellular component annotations in 124 sentences as annotated by senior dictyBase biocurator	Paper Identifier, annotation entity, paper section, curatable sentence, component term in sentence, GO term, GO ID and evidence code.	dictyBase senior curator	dictyBase and Plant Ontology[a]
PCS	50 textual descriptions of phenotypic characters in NeXML format randomly selected from 50 articles about fish or other vertebrates. Gold standard 50 character descriptions annotated by a senior Phenoscape biocurator	Entity term, entity ID, quality term, quality ID, quality negated, quality modifier, entity locator, count and more	Phenoscape senior curator	ZFIN and Phenoscape
PubTator	TAIR set: 50 abstracts (24 relevant) sampled from November 2011 for Arabidopsis already curated by TAIR NLM set: 50 abstracts sampled from Gene Indexing Assistant Test Collection (human)	Gene indexing: gene names and Entrez gene ID Document triage information: list of relevant PMIDs	Existing annotated corpus	TAIR and National Library of Medicine (NLM)
PPInterFinder	50 abstracts describing human kinases obtained by using a combination of tool/resources (such as UniProt, PubMeMiner, FABLE, and PIE).	PMID, protein interactant name 1, protein interactant name 2	NR	BioGrid and MINT
eFIP	PMID-centric: 50 abstracts randomly selected based on proteins involved in two pathways of interest to Reactome autophagy and HIV infection gene-centric: 10 first-ranked abstracts for 4 proteins involved in the adaptive immune system (Reactome: REACT_75774)	PMID, phosphorylated protein, phosphorylated site, interactant name, effect, evidence sentence	NR	Merck Serono, Reactome, and SGD[b]
T-HOD	PMID-centric: 50 abstracts from 2011 journals about obesity, diabetes or hypertension gene-centric: review relevancy of documents for four genes	PMID, EntrezGene ID, gene name, disease, gene–disease relation, evidence sentence	Protein Ontology senior curator	Pfizer, Reactome, GAD, and MGI

NR:non-recorded. [a]Curator novice to GO annotation. [b]SGD curator participated in first evaluation which is not reported in performance results here.

based on a seven-point Likert scale (10). The nature and form of the questions was inspired in part by those developed for QUIS and in part by other work: the computer system usability questionnaire (11), the perceived usefulness and ease of use survey (12) and the USE questionnaire (13). Questions used by these surveys were modified by the authors to better address specific aspects of interfaces for biocuration. The survey remains available at http://ir.cis.udel.edu/biocreative/survey.html.

During the pre-workshop evaluation, biocurators curated half of the dataset manually following the format provided by each system and half using the assigned system, recording the corresponding time on task.

Results

This section describes the results of the BioCreative 2012 worskhop IAT, and it is organized as follows: we first describe the pre-workshop evaluation, including details on each participating system, followed by a summary of the results. Then we describe the activity at the workshop, and finally, we provide a general summary of the evaluation including results from the demo session.

Participating systems

Six of the seven teams registered and provided the required system description, benchmarking results and were ready for testing by the agreed deadline (Table 1). The biocuration tasks proposed by the registered systems were widely heterogeneous, including extracting gene–disease relationships or protein–protein interactions, finding the genes mentioned in an abstract and correlating the mentioned genes to systematic nomenclature, ontology matching and retrieving documents mentioning specific diseases or chemicals. The reported metrics (Table 1) provided evidence that the system performance was reasonable and the systems were in good condition for testing. An additional system participated only in the demo session during the workshop.

Pre-workshop evaluation

The task by the curators included training, annotation and filling of the user survey (Figure 1). For the training, each biocurator needed to perform a series of tasks that included getting familiar with the system and curation guidelines provided by developers. At this stage, frequent communication between biocurators and developers was encouraged, and various modalities were exploited. Some of the groups had teleconferences with biocurators or demonstrated their system in a webinar-like format, whereas others provided all documentation via e-mails or via the coordinator. During this time, users could also report system bugs that could be addressed before testing. The annotation (evaluation per se) involved in all cases manual curation of a set of documents and curation of another set using the selected system. The manual output was according to a format provided by the systems.

A summary of the setting and results for the individual systems is presented in this section. Tables 4–6 summarize the performance and subjective measures.

Textpresso (14). This system is designed to retrieve sentences describing subcellular localization of gene products from the full text of papers. To identify these sentences, papers are searched using Textpresso categories, which are 'bags of words' that encompass terms of a common semantic concept. The categories used for the subcellular localization search are as follows: 1-assay terms, 2-verbs, 3-cellular component terms, 4-gene product names, plus an additional category and 5 tables and figures. Matching sentences must contain at least one term from each of these categories (15). In BioCreative, Textpresso was applied for the curation of cellular localization in selected documents for *Dictyostelium discoideum* using the GO cellular component ontology. For evaluating the results of the Textpresso searches, biocurators examined sentences from two different searches, namely, Category 4, including Categories 1–4, and Category 5, which adds the fifth category. Textpresso results were evaluated at the level of sentences as well as GO annotations that could be made from those sentences. At the sentence level, precision is defined as percentage of sentences retrieved by Textpresso that were relevant (i.e. described subcellular localization) and recall as percentage of relevant sentences Textpresso retrieved from the test documents. At the GO annotation level, precision is defined as the percentage of GO annotations made from Textpresso sentences that match either the gold standard GO term or a parent term (i.e. a correct but less granular term) in the ontology. The performance of Textpresso on the evaluation dataset at the sentence level is comparable to that presented in the internal benchmarking for Category 4 (precision and recall 80.1% and 30.0%, respectively, compare sentence level in Benchmark Tables 1 and 4). Also, the performance is similar for both category searches (compare results for system alone in Table 4). Textpresso-based GO annotation results show that it provides high-precision annotations when compared with manual annotation (compare GO annotation level in system-assisted annotation and manual annotation in Table 4). In terms of curation efficiency, Textpresso increased curation efficiency, once biocurators were familiarized with the system, by decreasing curation time ~2.5-fold (Table 5). In all cases, recall is lower than precision, which is related to (i) technical issues of the system, (ii) missing category terms and (iii) a statement in a paper correctly describes localization but is missing a category term (i.e. the result is described using less than the four or five required Textpresso categories). The survey results

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

Table 4. System performance metrics in pre-workshop evaluation

System performance measure (%)	System output versus gold standard annotation	System-assisted annotations			Manual annotation		
Textpresso							
Sentence level							
Category 4[a]	System alone						
Recall	37.9						
Precision	77.5				Curator 1[b]	Curator 2[b]	
F-measure	50.9				55.1	26.9	
Category 5[a]	System alone				41.7	63.3	
Recall	39.7				47.5	37.8	
Precision	81.5						
F-measure	53.4						
GO annotation level							
Category 4[a]		Curator 1	Curator 2				
Recall		37.1	14.5		Curator 1[b]	Curator 2[b]	
Precision		78.3	77.8		86.8	39.5	
F-measure		50.3	24.4		42.8	41.2	
Category 5[a]		Curator 1	Curator 2		57.3	40.3	
Recall		32.2	11.3				
Precision		75.0	71.4				
F-measure		45.1	19.5				
PCS							
Term-based EQs[c]	System alone				Curator 1	Curator 2[d]	Curator 3
Recall	65.0				47.0	38.0	50.0
Precision	60.0				57.0	65.0	67.0
F-measure	62.4				51.5	48.0	57.3
Label-based EQs[c]	System alone				Curator 1	curator 2[d]	Curator 3
Recall	24.0				44.0	51.0	51.0
Precision	23.0				54.0	81.0	74.0
F-measure	23.5				48.5	62.6	60.4
		Phenex + Charaparser			Phenex		
Label-based EQs[c]		Curator 1	Curator 2[d]	Curator 3	Curator 1	Curator 2[d]	Curator 3
Recall		51.0	38.0	66.0	37.0	63.0	36.0
Precision		58.0	70.0	84.0	49.0	88.0	60.0
F-measure		54.3	49.3	73.9	42.2	73.4	45.0
PubTator							
NLM indexing mention-level	System alone	Curator 1			Curator 1		
Recall	80.1	98.6			91.0		
Precision	83.4	98.3			93.0		
F-measure	81.7	98.0			92.0		
TAIR indexing document level	System alone	Curator 2			Curator 2		
Recall	76.0	90.0			91.0		
Precision	73.9	77.1			75.0		
F-measure	74.9	83.0			82.0		
TAIR triage	System alone	Curator 2					
Recall	68.6	84.6					
Precision	80.5	100.0					
F-measure	74.1	92.0					

(Continued)

Table 4. Continued.

System performance measure (%)	System output versus gold standard annotation	System-assisted annotations				Manual annotation	
PPInterFinder							
PPI algorithm alone	System alone	Curator 1	Curator 2			Curator 1	Curator 2
Recall	NR	69.8	63.8			72.7	79.7
Precision		85.7	85.7			87.0	90.4
F-measure		76.9	73.2			79.2	84.7
PPI algorithm (gene mention/ gene normalization)	System alone	Curator 1	Curator 2				
Recall	NR	46.9	46.9				
Precision		85.7	85.7				
F-measure		60.6	60.6				
eFIP							
PMID-centric (sentence level)	System alone	Curator 1	Curator 2			Curator 1	Curator 2
Recall	NR	69.2	88.2			89.5	77.8
Precision		94.7	79.0			85.0	70.0
F-measure		80.0	83.3			87.2	73.7
Gene-centric (document level)	System alone	Curator 1	Curator 2			Curator 1	Curator 2
Recall	NR	78.6	85.7			100.0	77.8
Precision		91.7	85.7			83.3	77.8
F-measure		84.6	85.7			90.9	77.8
Document-ranking							
nDCG	93–100						
T-HOD							
PMID-centric (sentence level)	System alone	Curator 1	Curator 2	Curator 3	Curator 4		
Recall	70.0	56.0	22.0	24.0	42.0		
Precision	79.5	32.0	26.0	40.0	42.0		
F-measure	74.5	40.0	24.0	30.0	42.0		
Gene-centric (document level)	System alone	Curator 1	Curator 2	Curator 3	Curator 4		
Recall	54.3	56.0	30.0	26.0	42.0		
Precision	72.1	63.0	41.0	52.0	71.0		
F-measure	62.0	59.0	35.0	35.0	53.0		

[a]4-Category search use 'bag of words' for (1) assay terms, (2) verbs, (3) cellular component terms, and (4) gene product names, whereas 5-Category search also include words for Table and Figures. [b]Manual annotations don't necessarily correspond to either the 4- or 5-category search as curators do annotations for sentences that fit both criteria. [c]Term-label EQs are entity-quality statements created strictly based on the original descriptions, independent of any ontologies, whereas the label-based EQs are the corresponding formal statements (using ontology terms). [d]Curator ignore an unspecified number of CharaParser proposals to save time.

for the two biocurators involved were heterogeneous. It is relevant to mention that one of the curators was a novice to GO annotation, and results of the survey by this curator could reflect the experience with both system and curation task (Table 6).

Phenoscape Curation System (16). This system is designed for the curation of phenotypes from evolutionary literature on fishes and other vertebrates. Three biocurators did the evaluation using Phenex (17) (curation system used by Phenoscape biocurators), and using the Phenoscape curation system (PCS) system (consisting of Phenex plus CharaParser, the text mining tool). The curation task required curators to capture the phenotypical characters in the form of entity and quality terms (EQ) and identifiers (IDs) from a number of anatomic and phenotypic quality ontologies. Recall and precision on term-based EQs (i.e. EQs created strictly based on the original descriptions, independent of any ontologies) and label-based EQs (i.e. the result of translating and transforming the terms in

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

Table 5. Ratio of time for task completion: manual/system-assisted and curation time range

Time ratio manual/system					Time range (min)	
System	Curator 1	Curator 2	Curator 3	Curator 4	Manual	System
Textpresso	2.3	2.5[a]			375–692	150–297
PCS	1.0	0.8			135–210	165–210
Pubtator	1.8	1.7			83–135	49–79
PPInterFinder	0.9	NR			58	62
eFIP	2.4	2.5			88–120	35–50
T-HOD	0.9	1.3	1.2	4.0	110–140[b]	110–120[b]

NR, not recorded. [a]Only after getting familiar with the tool. [b]One curator was significantly faster 60 min manual to 15 min with T-HOD and is not shown.

Table 6. Overall rating for each system by category in pre-workshop evaluation

Subjective measure (overall median for each section)						
System	Overall evaluation	Task completion	System design	Learnability	Usability	Recommendation
Textpresso	4.0	4.5	6.0	6.0	6.0	3.5
PCS	3.0	3.0	4.5	6.0	7.0	3.0
PubTator	6.0	6.0	6.0	6.0	6.0	7.0
PPInterFinder	2.5	1.0	4.5	5.5	3.5	2.0
eFIP	5.5	6.0	6.0	6.0	6.0	5.0
THOD	4.0	3.0	4.5	5.0	5.0	3.0

Median value for questions linked for each of the categories. Likert scale from 1 to 7, from worst to best, respectively.

Table 7. Degree of correlation of top 10 questions to overall satisfaction measure

Question	Correlation
Q4: personal experience	0.719
Q10: task completion efficiency	0.622
Q8: task completion speed	0.569
Q5: power to complete tasks	0.568
Q9: task completion effectiveness	0.53
Q23: consistent use of terms	0.473
Q6: flexibility	0.443
Q25: helpful error messages	0.438
Q15: learning to perform tasks	0.431
Q3: ease of use	0.431

term-based EQs to their best-matched class labels in related ontologies) were calculated. The performance is significantly lower than the one reported in the benchmarking (compare results from system alone in Table 4 with those of Table 1). However, term-based performance of PCS has higher recall than biocurators' performance and similar precision, whereas label-based performance of PCS was about half of biocurators' performance (compare system alone versus manual curation in Table 4). Interestingly, inter-annotator agreement was low (precision among pair of annotators ranged from 31% to 77%, and recall 49% to 71%), which highlights the difficulty of phenotype curation. The comparison of performance on label-based EQs generated by biocurators using Phenex and PCS shows that the text mining tool improved curation accuracy for two of the three biocurators (compare Phenex and Phenex+Charaparser results in Table 4). Curation efficiency in terms of time on task was not improved by using the tool (Table 5). In this evaluation, PCS's failures relate to (i) the inherent difficulty of the phenotype curation task involved in translating term-based EQs to label-based EQs as there is no well-defined way to perform some of the translations; (ii) the incompleteness of ontology coverage (since 55% of the target EQ classes were not included in the ontologies, the maximum possible performance of CharaParser would be 45% precision/recall); and (iii) the failure in equipping CharaParser with all ontologies used by biocurators. The results from the three biocurator surveys were heterogeneous. A consistently low rating ($\leq$3) was given to all

Table 8. Overall rating for each system by category

Subjective measure (overall median for each section)						
System	Overall evaluation	Task completion	System design	Learnability	Usability	Recommendation
PubTator	6.0	5.5	6.0	6.0	6.5	7.0
eFIP	6.0	6.0	6.0	6.0	7.0	5.5
Tagtog[a]	5.0	5.0	5.0	5.0	6.0	4.5
Textpresso	4.0	5.0	5.0	5.0	6.0	4.5
PCS	4.0	3.0	6.0	6.0	6.0	4.0
PPInterFinder	4.0	2.5	5.0	5.0	5.0	3.0
T-HOD	4.0	3.0	4.0	5.0	5.0	3.0

Median for questions linked for each of the categories. Likert scale from 1 to 7, from worst to best, respectively. [a]This system was only reviewed at the workshop.

questions related to the system's ability to help complete tasks, whereas consistently high ratings (≥4) were given to the usability of the tool (Table 6). Feedback from biocurators indicated that the tool needs better recognition of entities and qualities from the free text to improve the recall of the system (currently 'to decrease the low recall of the system')

PubTator (18). PubTator is a web-based tool that allows biocurators to create, save and export annotations, with similar look and feel as PubMed. PubTator relies on three state-of-the-art modules: GeneTUKit for gene mention (19), GenNorm for normalization (20) and SRG4N (21) for species recognition. This system was set up for two specific biocuration tasks, namely, document triage (retrieve relevant papers to be curated) and bioconcept annotation. Biocurators from NLM and TAIR databases participated in the evaluation. The NLM biocurator worked at the mention level and normalized to NCBI Gene identifiers. The TAIR biocurator worked at the document level and normalized to TAIR's own nomenclature. Besides gene indexing, the TAIR biocurators also conducted document triage task—abstracts were labeled as relevant for full curation or otherwise. As shown in Table 5, PubTator-assisted curation significantly improved efficiency (~1.75-fold decrease in curation time) and also slightly increased accuracy compared with the baseline (compare manual curation with system-assisted in Table 4). According to the survey by the two biocurators involved, its users liked this system: median scores for each category are ≥6 and for all questions PubTator's ratings were ≥4 (Table 6).

PPInterFinder (22). This system was set up for extracting information about human protein–protein interactions. Biocurators from PPI databases evaluated this system, including biocurators from the MINT and BioGRID databases. The curation scenario was protein-centric, focusing on human kinases for which it is important to annotate protein interactions and phosphorylation events. PPInterFinder was evaluated at the sentence level where a true positive represents the number or proportion of interacting protein entities that were correctly identified by the system. The results show that there is a significant difference in the performance of this system when the processing stage of gene recognition is included in the IAT (compare PPI algorithm alone with PPI algorithm with gene mention in Table 4). It should also be noted that the performance differs (it is much lower) from the benchmark results reported in Table 1. The difference is due to protein names that could not be recognized or normalized, a problem already described in previous BioCreative efforts (BioCreative II and II.5 PPI task). The lower performance is also partly explained by the low inter-annotator agreement (36 agreements and 19 disagreements), which points that the annotation guidelines may not have been clear. In terms of curation efficiency, as recorded by time on task, there is information for only one biocurator and it shows the time for manual and system-assisted annotations were comparable (Table 5). According to the survey based on the two biocurators involved, this system was consistently rated very low in its ability to help complete tasks (Table 6). A few aspects of the design of the application were consistently rated on the high end. Biocurators indicated the need of more precise results and less false positives. Through the organization of this task, it was possible to exploit particular aspects of the biocurators feedback and evaluation in order to improve the PPInterFinder system. Aspects considered for this improvement include the reduction of the number of false-positive results by revising the dictionaries for relation keywords, by enhancing rules, patterns and the relation recognition algorithm.

eFIP (23). A main goal of eFIP is to suggest documents containing information that is relevant for biocuration of

Database, Vol. 2012, Article ID bas056, doi:10.1093/database/bas056

phosphorylated proteins related to protein–protein interactions. The eFIP system ranks abstracts based on the amount of relevant information they contain and presents evidence sentences and a summary table with the phospho-protein, the interacting partner and impact words (increase, decrease, block, etc.). eFIP integrates text mining tools such as eGRAB (24) for document retrieval and name disambiguation, RLIMS-P (25) for extraction of phosphorylation information, a PPI module to detect PPIs involving phosphorylated proteins and an impact module to detect temporal and causal relations between phosphorylation and interaction events in a sentence. In BioCreative 2012, two tasks were planned as follows: (i) a *PMID-centric* task, given a set of documents identify curatable documents (those with phosphorylated proteins related to PPI), with the corresponding evidence sentences and (ii) a *gene-centric* triage task, given a gene, validate the relevancy of the documents retrieved by eFIP (Do articles retrieved contain phosphorylation and related PPI for the query protein?). Biocurators from Reactome, SGD and Merck Serono participated in the evaluation. eFIP performance was evaluated for document retrieval, sentence-level information extraction and document ranking. Besides the documents in the given dataset, users were asked to validate the ranking by eFIP output for relevant genes. Both at the sentence and document levels, eFIP achieved higher precision (compare system-assisted with manual curation in Table 4), but lower recall than manual curation (in many cases one biocurator ignored redundant annotations). The inter-annotator agreement was significant as indicated by a Cohen kappa coefficient (26) of 0.77 (54 agreements/7 disagreements) and eFIP improved curation efficiency by decreasing curation time ~2.5-fold (Table 5). eFIP performance evaluation on document ranking as measured by nDCG (normalized discounted cumulative gain) based on the ranked lists of abstracts ranged between 93% and 100% (Table 4). In general, factors that contributed to a decrease of precision and recall in eFIP are mostly attributed to the PPI module: reporting interactions between entities other than proteins, failing to detect directionality on complex sentences and in a few cases inability to identify an interaction event. The survey by the three biocurators who participated in the evaluation indicates that users like the system (Table 6). It is relevant to mention that consistently high ratings ($\geq$5) were given for all questions in the category system's ability to complete the task. One of eFIP's strengths seems to be the ease of finding relevant articles in the literature as manual dataset selection for this activity has been very challenging for organizers due to the complex relations captured by this tool.

T-HOD (27). 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. T-HOD relies on state-of-the-art text mining tools for gene identification (28) and for disease recognition, and disease–gene relation extraction (29). For BioCreative 2012, two tasks were planned as follows: (i) a PMID-centric task, given a set of documents identify sentences with gene–disease relations and (ii) a gene-centric task, given a gene, validate the relevancy of the output from T-HOD (Do articles retrieved contain gene–disease relation for the given gene?). Biocurators from Pfizer, Reactome, MGI and GAD participated in the evaluation. For the calculation of performance metrics in the PMID-centric approach, information regarding a gene–disease relation mentioned in an abstract including the gene term, gene ID and the sentence describing the relation all have to be exact with the gold standard in order to be a true positive. In the gene-centric approach, only the gene terms have to be correct to be considered as a true positive. The precision of T-HOD at the document level in the evaluation was similar to the one reported as benchmark results, but the recall was lower (compare Table 1 with system alone document level in Table 4). Interestingly, the performance of the system alone when compared with the gold standard was significantly higher than the biocurator's set (compare system alone with system-assisted annotation in Table 4). The inter-annotator overall agreement was moderately low 57.47%, which may explain in part the performance results. The performance of the system in PMID-centric evaluation was lower than the gene-centric task mainly due to the fact that both the entity recognition (for gene and disease) and the relation extraction have to be correct in order to achieve good performance (Table 4). Of these tasks, the gene term recognition and normalization were the most difficult. In addition, there were some cross-sentence gene–disease relations in the gold standard, which is not yet supported by the system. In terms of curation efficiency, only one of the biocurators reported a significant increase 4-fold (Table 5). This biocurator only went through the positive examples suggested by the system and did not check for any false negatives. However, this biocurator seems much faster than others given that the manual curation also took a significantly shorter time than for the other biocurators (60 min versus 110–140 min for other three biocurators, Table 5). According to the survey of the four biocurators involved in the evaluation, the system's ability to help complete tasks was the category with lowest median, whereas learning the application and usability were the ones with highest (Table 6). Some of the suggestions by the users included expanding to non-disease-centric queries, expanding scope to other diseases, improving some aspects of the interface (e.g. display window does not auto-adjust into the proper size of the browser; users are unable to return to reconsider

their last action) and providing more documentation. T-HOD is working on an enhanced version based on these suggestions.

BioCreative 2012 workshop. At the workshop, which took place in Washington DC on April 4–5, 2012, each participating team presented the results of the pre-workshop evaluation. In addition, based on the success of the demo session in BioCreative III, we extended this session to include a usability evaluation by users. The teams demonstrated their system and biocurators attending the session had the opportunity to try systems. Each session was 30 min long. We collected opinions via the same user survey described in the 'Materials and Methods' section. We recruited new members for the UAG to assist in this endeavor. Each member was assigned two systems to ensure all systems were tested. Other biocurators present at the session could also test by selecting the system of their choice. At the end of the testing, the user had to fill the same user survey to the best possible extent and provide their first impression about the systems. We collected 22 survey responses in this activity. The results are included in the analysis shown in the next section. Although not reported in the metrics for this task, two additional groups demonstrated their systems: ToxiCat (30) and ODIN (8), both of which participated in the Triage challenge (Track I) of the workshop. Survey responses are available in Supplementary Table 1.

Overall analysis of the evaluation results

Note that we are aware of the limitations of this analysis both in terms of limited numbers of biocurators per system and the widely different nature of the tasks systems perform. For example, we cannot directly compare the performance metrics across systems. However, we can derive some useful observations and these are described next. The performance results from the pre-workshop evaluation indicate that a set of systems were able to improve efficiency of curation by speeding the curation task significantly (1.7- to 2.5-fold faster than performing it manually, Table 5). Acquiring familiarity with system output and curation tools were shown to be key for maximizing efficiency at least in one case (in Textpresso system one biocurator was novice to curation task, once familiarized the efficiency improved). Some of the systems were able to improve annotation accuracy when compared with manual performance (e.g. PubTator, eFIP and PCS, compare system-assisted with manual annotations in Table 4). In terms of inter-annotator agreement, the factors that contributed to significant differences for some of the systems were the expertise of the biocurator on the curation task (as happened in GO annotation), inherent difficulty of the curation task as was the case of the annotation of phenotypes and not following provided annotation guidelines (e.g. cases where an annotator is asked to mark all sentences but he/she chooses to pick a representative one). The results also show that many of the systems rely on the combination of many different text mining modules and how the performance of each one impacts significantly the performance of the entire system (addition of gene mention/normalization algorithm in PPInterFinder decreases performance of the system significantly, compare Table 4).

We hypothesize that questions that correlate highly to overall measures of satisfaction reflect greater importance to biocurators in general; if a system receives high scores for one question while scoring low on overall satisfaction, that question may not be particularly important to the biocurator's experience. Based on the questionnaire in http://ir.cis.udel.edu/biocreative/survey.html, Table 7 ranks the top 10 survey questions by the degree of correlation (Kendall's tau rank correlation (31), since the Likert scale is discrete) in their responses to three overall measures of satisfaction: Questions 2 (the biocurator's subjective evaluation of the system), 7 (whether the biocurator would recommend the system) and Question 1 regarding whether the biocurator enjoyed using the system. This ranking is computed as follows: we first compute the median rating for each system on each question and then rank systems for each question by that median. We compute Kendall's tau correlation between this ranking and the ranking of systems by the median rating for Questions 2, 7 and 1 and then take the average of the three resulting tau correlations. Note that these three questions are themselves very highly correlated ($\text{tau} > 0.9$, which is significant with $P < 0.01$). This was done for all filled surveys (38 total). It is clear from this table that task completion is very important, followed generally by the system's usability.

Finally, Table 8 gives an overall rating for each system for each group of questions (overall evaluation, task completion, system design, learnability, usability and recommendation) computed by taking the median value of all biocurator ratings for all questions in the group: overall evaluation (median of Questions 1–6), task completion (median of Questions 8–10), system design (median of Questions 11–14), learnability (median of Questions 15–19), usability (median of Questions 20–25) and recommendation (Question 7 alone). While systems generally score high for design, learnability and usability, it is clear from this table how important task completion is to the biocurators' overall experience.

All teams benefited to an extent by participating in Track III, especially by the feedback received from biocurators (pre- and at the workshop). As a result, some of the teams have improved or are improving their systems, and others have engaged new communities. For example, in PPInterFinder, the 'relation keywords' list was refined to decrease the false positives and new patterns were

added to PPI extraction methodology (22). PubTator (18) is planning to extend the bioconcepts being covered, as well as processing full-length articles, and T-HOD is extending the disease coverage (27). Finally, Textpresso engaged TAIR to similarly evaluate performance of Textpresso for cellular component curation of *Arabidopsis* gene products (14).

Discussion

The current IAT has been very challenging from multiple logistic aspects: recruitment and coordination of biocurators that can properly evaluate the systems; selection of datasets; issues of system readiness and data collection, formatting and processing. However, it is a great experience for both developers and users. Users are exposed to tools that may assist them in their curation; developers interact with potential users and learn about their real needs. We would like to point out that this activity covered different levels of annotation types that can be related to different biocuration strategies: document-centric and bio-entity centric biocuration as well as different level of granularity of the obtained results: from pure annotation relations between entities without textual evidence, and together with textual evidence at the level of phrases, sentences, passages and whole documents.

Below we describe some recommendations based on the lessons learned from this activity.

What is a biocuration task?

We found that it is important to align the views between biocurators and text mining groups on what constitutes a biocuration task, especially for those teams that do not work closely with biocuration groups. There are at least three aspects to consider which may influence the practical use of the text mining system: (i) in general, text mining systems should be more concerned with annotation guidelines as used in existing annotation workflows. We found that in some cases, even though teams provided guidelines, these did not follow necessarily the standards used by the representative databases (e.g. the definition of annotation types differs). On one hand, this may affect the system's performance as the biocurator has to be 'retrained' to the new guidelines, but tends to follow his own, but more importantly, the output of the system may be incompatible with the annotation standards, and therefore it may not be used; (ii) another critical aspect is the system capability to provide flexible options to improve the interpretation of the extracted data, as this is key in biocuration. For example, displaying an isolated sentence without pointers to explore additional context information makes it hard for human interpretation and validation of text mining results and (iii) the third one is the aspect of output validation. Many of the text mining systems required validation of the results at the sentence level (in many cases validating redundant information), whereas the biocurator decides at the abstract/document level. In this context, the implementation of sentence/text ranking methods to select the more informative and representative sentences becomes crucial. We plan to provide more guidance about these topics to the teams in the next BioCreative.

What to compare?

In this 'experiment', we compared manual versus system-assisted curation to have a common baseline, but we are aware that this may not represent how a biocurator does literature curation in their curation workflow, and therefore it may impact the efficiency of the manual task (time it takes to complete the task). We think that this approach is still very informative and along with the biocurators' feedback we should be able to better plan for BioCreative IV. It was very positive to find that many of the systems (4 out of 6) sped up curation in this comparison (Table 5). In addition, we had one case where a comparison between a curation tool and the same curation tool with text mining modules was made (Phenex with or without the CharaParser module). However, this option would be difficult to implement in a BioCreative setting if the intention is to engage a variety of users from different databases to try the systems. Alternatively, if systems could provide an interface where the user can activate or inactivate the use of a text mining tool while retaining other website functionalities, the result could be interesting to explore. Domeo (32) is an example of an annotation system in which the user can manually annotate a text or annotate automatically using a selected set of ontologies. In both cases, the annotations are saved in the same format (RDF) providing a way to easily compute metrics. This would also allow a better comparison of time on task. In this respect, ODIN (8) is a biocuration system that is able to record biocurators activities. We will explore some of these options in future discussion with the UAG.

System adaptability

This evaluation showed how the different systems could be adapted to assist the various database interests. 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 identify the appropriate PDF articles about *Dictyostelium* and to import the gene name and synonym vocabulary, among other things. Similarly, PubTator included TAIR's gene nomenclature to be useful for this database, and Charaparser was set up to work with the Phenex curation system. We should bear in mind that even when the systems are functional for a database, they may need some minor development to adapt it to a different user group. Therefore, we will be very mindful in allocating sufficient time for this in BioCreative IV.

Flexibility of Track III

To better learn the landscape of different tools, we opened this Track for any biocuration task. In fact, Table 1 shows how diverse were the tasks proposed by the participating systems. The activity was purposely designed to be flexible in many respects. First, regarding the metrics, we asked teams to suggest the appropriate metrics for their system. Most systems reported recall, precision and F-measure at different levels (sentence, document, etc). In addition, there were some particular measures for document ranking proposed as well as for inter-annotator agreement which might be interesting to explore as future standard metrics. Second, regarding system input, initially we proposed to have a PMID-centric approach for curation—given a set of documents, perform the task—because we wanted to expose the systems to a variety of examples. However, curation approaches vary so we also allowed other types of inputs. For instance, in PCS the input is a list of phenotypical characters in NeXML format (33), whereas other systems were gene-centric (e.g. PubTator, eFIP) or disease-centric (e.g. T-HOD). Finally, we did not request any specific format for the system output or the interface capabilities.

This flexibility, although it may increase the workload in terms of task planning, and data analysis, provides a great means to observe the approaches, standards and functionalities used by state-of-the-art systems. We believe results will aid in choosing appropriate metrics and standards for BioCreative IV challenges.

Engaging biocurators

Feedback from participating teams in the IAT of BioCreative 2012 workshop indicates that the participation of biocurators is one of the most valued aspects of this activity. In addition, recruitment of domain experts on the curation task is essential. In this regard, the coordinators from the IAT contacted many groups and consulted with teams to try to find the appropriate set of biocurators for each system, but the time frame, the lack of evident reward and other commitments were some of the barriers that prevented biocurators from participating in the pre-workshop evaluation. Based on this experience, we have now a better sense of the commitment needed by biocurators, and we expect that the UAG will serve not only by advising on BioCreative IV planning but also by providing insight on how to recruit biocurators or even serving as a source for biocurators. In this regard, the direct exposure of the UAG to the activity during the workshop has been an asset.

Finally, despite all the challenges, roadblocks and possible mistakes, both biocurators and text mining teams expressed interest in future participation in this activity. The IAT has served as a medium to experiment with different approaches to formally assess interactive systems. Similar to the BioCreative III IAT experience, we expect that the lessons learned will help to shape the future BioCreative IV task, not only to improve the IAT but also the challenge tracks that involve biocuration.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors would like to thank all the teams and the biocurators who participated in this activity.

Funding

National Science Foundation grant DBI-0850319; National Institutes of Health grant 5G08LM010720-02; The participation of Z.L. and W.J.W. was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The participation of M.K was supported by CONSOLIDER grant CSD2007-00050 and MICROME grant 222886-2.

Conflict of interest. None declared.

References

1. Burge,S., Attwood,T.K., Bateman,A. *et al*. (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database*, **2012**: article ID bar059; doi:10.1093/database/bar059.
2. Hirschman,L., Burns,G.A., Krallinger,M. *et al*. (2012) Text mining for the biocuration workflow. *Database*, **2012**: article ID bas020; doi:10.1093/database/bas020.
3. Hirschman,L., Yeh,A., Blaschke,C. and Valencia,A. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**, S1.
4. 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**, S1.
5. 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.
6. Arighi,C., Lu,Z., Krallinger,M. *et al*. (2011) Overview of the BioCreative III workshop. *BMC Bioinformatics*, **12**, S1.
7. Arighi,C., Roberts,P., Agarwal,S. *et al*. (2011) BioCreative III interactive task: an overview. *BMC Bioinformatics*, **12**, S4.
8. Rinaldi,F., Clematide,S., Garten,Y. *et al*. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database*, **2012**: article ID bas021; doi:10.1093/database/bas021.
9. Chin,J.P., Diehl,V.A. and Norman,K.L. (1988) Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of ACM CHI Conference on Human Factors in Computing Systems.* New York, pp. 213–8.
10. Likert,R. (1932) A technique for the measurement of attitudes. *Arch. Psychol.*, 1–55.

11. Lewis,J.R. (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int. J. Human Comput. Interact.*, **7**, 57–78.
12. Davis,F.D. (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, **13**, 319–40.
13. Lund,A.M. (2001) Measuring usability with the USE questionnaire. *Society for Technical Communication Usability SIG Newsletter,* P. 2.
14. Van Auken,K., Fey,P., Berardini,T.Z. *et al*. (2012) Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase, and TAIR. *Database*, **2012**: article ID bas040; doi:10.1093/database/bas040.
15. 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 omponent curation. *BMC Bioinformatics*, **10**, 228.
16. Cui,C., Balhoff,J., Dahdul,W. *et al*. (2012) PCS for phylogenetic systematic literature curation. *Proceedings of the 2012 BioCreative Workshop. Washington DC, USA, 144–151.*
17. Balhoff,J.P., Dahdul,W.M., Kothari,C.R. *et al*. (2010) Phenex: ontological annotation of phenotypic diversity. *PLoS One*, **5**, e10500.
18. Wei,C.-H., Harris,B.R., Li,D. *et al*. (2012) Accelerating literature curation with text mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, **2012**: article ID bas041; doi:10.1093/database/bas041.
19. Huang,M., Liu,J. and Zhu,X. (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.
20. Wei,C., Huang,I., Hsu,Y. and Kao,H. (2009) Normalizing biomedical name entities by similarity-based inference network and de-ambiguity mining. *Ninth IEEE International Conference on Bioinformatics and Bioengineering Workshop: Semantic Biomedical Computing: 2009;* Taichung, Taiwan, pp. 461–466.
21. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PLoS One*, **7**, e38460.
22. Raja,K., Subramani,S. and Natarajan,J. (2012) PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database*, **2012**: article ID bas052; doi:10.1093/database/bas052.
23. Tudor,C.O., Arighi,C.N., Wang,Q. *et al*. (2012) The eFIP system for text mining of protein interaction networks of phosphorylated proteins. *Database*, **2012**: article ID bas044; doi: 10.1093/database/bas044.
24. Tudor,C., Schmidt,C. and Vijay-Shanker,K. (2010) eGIFT: mining gene information from the literature. *BMC Bioinformatics*, **11**, 418.
25. Hu,Z.Z., Narayanaswamy,M., Ravikumar,K.E. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.
26. Landis,R. and Koch,C. (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.
27. Dai,H.-J., Wu,J.C., Tsai,R.T. *et al*. (2012) T-HOD: Text-mined hypertension, obesity, diabetes candidate gene database. *Database*, **2012**: BioCreative Workshop Proceedings, Washington DC, pp. 121–131.
28. Dai,H.-J., Chang,Y.C., Tsai,R.T.-H. and Hsu,W.L. (2011) Integration of gene normalization stages and co-reference resolution using a Markov logic network. *Bioinformatics*, **27**, 2586–2594.
29. Tsai,R., Lai,P.-T., Dai,H.-J. *et al*. (2009) HypertenGene: extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *BMC Bioinformatics*, **10**, S9.
30. Vishnyakova,D., Pasche,E. and Ruch,P. (2012) Selection of relevant articles for curation for the comparative toxicogenomic database. *Proceedings of the 2012 BioCreative Workshop,* Washington DC, USA, pp. 31–38.
31. Kendall,M. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–89.
32. Ciccarese,P., Ocana,M. and Clark,T. (2012) Open semantic annotation of scientific publications using DOMEO. *J. Biomed. Semantics*, **3**, S1.
33. Vos,R.A., Balhoff,J.P., Caravas,J.A. *et al*. (2012) NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Sys. Biol.*, **61**, 675–689.