DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas043, doi:10.1093/database/bas043

Original article

Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II

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

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894 and [2]The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

***Corresponding author:** Tel: +1 781 271 7789; Fax: +1 781 271 2780; Email: lynette@mitre.org

Submitted 18 June 2012; Revised 24 August 2012; Accepted 2 October 2012

Manual curation of data from the biomedical literature is a rate-limiting factor for many expert curated databases. Despite the continuing advances in biomedical text mining and the pressing needs of biocurators for better tools, 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 for the BioCreative 2012 Workshop Track II. We received seven qualified contributions, primarily from model organism databases. Based on these descriptions, we 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. Compared to a survey done in 2009, our 2012 results show that many more databases are now using text mining in parts of their curation workflows. In addition, the workshop participants identified text-mining aids for finding gene names and symbols (gene indexing), prioritization of documents for curation (document triage) and ontology concept assignment as those most desired by the biocurators.

Database URL: http://www.biocreative.org/tasks/bc-workshop-2012/workflow/

Introduction

BioCreative (Critical Assessment of Information Extraction in Biology) is an international community-wide evaluation of information extraction applied to molecular biology (http://www.biocreative.org/). From its inception in 2004, BioCreative challenge evaluations have been developed in close association with the biocuration community to provide tools to assist in the curation of the biomedical literature (1–5). Challenge evaluation tasks over the years have included ranking of documents for curation based on presence of curatable information (‘document triage’), as well as extraction of genes and proteins from abstracts and articles (6,7) and their linkage to identifiers in standard biological resources (e.g. Entrez Gene, UniProt) (8–10). BioCreative has also addressed more complex tasks such as functional annotation for proteins in full-text articles using Gene Ontology (GO) terms, and extraction of protein–protein interactions (11–13).

A major goal of BioCreative has been to bring together the biocuration community and the text mining community to stimulate discussion between the curators—the end users of new information extraction and text-mining tools—and the developers of those tools, who need to become familiar with the needs and workflows of the biocurators.

To foster this communication, the BioCreative organizers held a workshop on ‘Text Mining for the Biocuration Workflow’ at the third International Biocuration Conference (Berlin, April 2009). In preparation for that workshop, the workshop organizers interviewed curators and elicited workflows for eight expert curated biological databases (14), 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 for both biocurators and text-mining developers. The workshop encouraged dialogue between these two communities who typically attend different meetings and do not have much opportunity to interact; it also enabled groups to identify potential partnerships. For the text-mining developers, the workshop provided an opportunity to hear curator priorities and to understand the overall workflow, including specific opportunities for the application of text mining. Curator priorities included support for document triage and the ability to curate from full text (and not just abstracts). For the curators, the workshop allowed them to communicate their workflow and to learn about the state of the art of text mining for biocuration. One of the interesting findings was that the detailed curation workflows elicited from the eight groups differed quite a bit—despite the fact that four were model organism databases (MODs). 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. Of the eight curation teams interviewed, most had experimented with text-mining tools, and several (mostly the older MODs) were using tools for search and browsing of the literature.

As a follow up to the workshop, the organizers did a survey of biocurators to determine how many groups were using text mining, and what curators' specific priorities were; these results are also reported in Ref. 14. At the time, almost 70% of curators surveyed reported experimenting with text mining, but less than half were using it. The 2009 workshop led to several additional publications on the integration of text-mining tools into the biocuration workflow. In Refs 15 and 16, 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 *et al*. (17) provided an overview of current text-mining methods for linking ontologies and protein–protein interactions to the biomedical literature, from their BioCreative experiences (3,12).

The BioCreative 2012 Workshop Track II on 'Curation Workflows' is a direct outgrowth of the 2009 workshop. The positive feedback from the biocurators led us to propose a track devoted explicitly to collecting workflows from multiple biological databases. This paper, together with the papers from the biocuration teams who participated in Track II, provides the next 'snapshot' of progress in providing text-mining tools to support biocuration. The workshop also provided an excellent opportunity for curators and text-mining developers to continue their interaction and mutual education.

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 interested in developing algorithms and tools to assist the curation process (shown in Table 1).

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

- AgBase (agricultural plants and animals),
- FlyBase (fruit fly),
- MaizeGDB (maize),
- MGI (mouse),
- TAIR (*Arabidopsis*),
- WormBase (*Caenorhabditis elegans*) and
- 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 (14).

Results

The workflows showed commonalities across the three stages identified in Table 2, as well as differences. Table 3 summarizes some of these comparisons. In the first stage, source collection, teams retrieve papers from PubMed and the main difference lies in the number of papers to be curated, which is heavily dependent on the curation resource of each individual group. Following that, the common practice is to identify relevant papers and assign curation priorities based on the content and gene/proteins mentioned in the paper abstract. Here, teams mostly differ in terms of paper selection criteria. Furthermore, in addition to identifying gene/proteins, some teams also search for biological entities such as cell types in this stage. The final step is full-paper curation. Despite commonalities such as the use of full text and controlled vocabularies, there exists a wide variety of differences between individual teams. For instance, because teams aim to capture different entities and relationships, different ontologies are used (details shown in Table 4).

All of the databases encoded a variety of biological entities using standard vocabularies and ontologies.

Table 1. Outline of issues for describing the curation workflow

Issue	Specific questions
Introduction	(i) Overall philosophy: what information is captured and from what sources? (ii) What use is being made of this information or is envisioned for this information? (iii) What is the current workflow of the operation, and where are automated methods used?
Encoding methods	(i) How is the information captured to make it machine readable? (ii) What entities are involved and how are they entered in the database? (iii) What relationships are involved and how are they symbolized? (iv) What standardized or controlled vocabularies are used? (v) Give examples of a variety of data elements and how they appear in the database
Information access	(i) When a curator runs into a problem or a difficult case, what kind of information is needed to solve it? (ii) What kind of internet searching is used most often in difficult cases? Dictionary? Wikipedia? Other database?
Use of text-mining tools	(i) What text-mining tools do you currently employ in your workflow and what problems do these algorithms solve for you? (ii) 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 could text mining overcome current bottlenecks in the existing pipeline)?

Table 2. Stages in the curation workflow

Curation stage	Sub-stage	Description
Sources	0	Collecting papers to be curated from multiple sources
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 (e.g. sentences, images)
	5	Check of record

Table 4 identifies (a subset of) common types of biological entities curated in the various databases. In particular, all of the databases used the GO (18) 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.

Generally speaking, MODs report that most papers contain all the information needed for making annotations. When there is a lack of sufficient information or a curator runs into a difficult case (e.g. an ambiguous gene name), the following steps are commonly used:

- Performing a BLAST search based on sequence information in the paper,
- Examining the supplementary files for additional details,
- Consulting relevant papers from the previously curated papers,
- Contacting the author for clarification and
- Searching information from other sources. Common ones include PubMed, Wikipedia, Textpresso (19), UniProt, etc.

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 (19) to search for specific classes of entities and/or to pre-assign certain classes of concepts (20). 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 extensions to handle gene function and biological process, both of which are 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).

Table 3. Commonalities and differences in the curation workflow stages

Curation stage	Commonalities	Differences
Source collection	• 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	• Database-specific selection criteria (e.g. species, gene/function, novelty) • Identification of additional bio-entities (e.g. anatomy, cell type)
Full curation	• Gene (function) centric • Use of full text • Use of controlled vocabularies and ontologies • Identification of experimental evidence • Contacting authors when needed	• Annotating database/species-specific entities and relationships • Annotating images (Xenbase)

Table 4. Common ontologies used across multiple curation databases (''X'' indicates ontology in use by the database in column header)

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

Discussion and Conclusions

One striking change from the 2009 results is that, as of 2012, the seven databases that participated in 2012 track are using text mining in at least some parts of their workflow. This contrasts with the 2009 survey, where less than half of the biocurators (46%) reported that they were currently using text mining. Although these two data points reflect reports from different (though partially overlapping) sets of curators, nonetheless it seems safe to conclude that there has been significant uptake of text-mining technologies incorporated into the biocuration workflow over the past few years.

There may be several reasons for this, including the maturing of text-mining tools. There was also heavy representation of MOD curators participating in Track II of the 2012 workshop; some of these teams are making use of a sophisticated suite of open source software tools available through GMOD (http://gmod.org), including Textpresso. As noted above, Textpresso is being used in six of the seven databases, and its capabilities are being extended, in response to the needs of the MODs. Textpresso's success can be attributed to several factors: the developers came out of the model organism community (WormBase); it was developed as an open-source tool suite to support the MOD community; it has been built around the main ontologies in use in MOD curation; and the developers have supported a number of tool migrations to adapt Textpresso to new databases, resulting in a tool suite that is increasingly easy to tailor and insert into the workflow for additional databases.

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, we can now begin to quantify where curator time is spent. For example, Wiegers *et al.* (16) reported that in the CTD it was easy for biocurators to identify articles not appropriate for curation workflow; overall, CTD biocurators only spent 7% of their time on these (average of 2.5 min per rejected article versus 21 min on average for a curatable article), with 40% of articles designated as 'not appropriate'. Of course, the time savings is heavily dependent on the ratio of curatable to non-curatable documents presented: in situations where it is difficult and time-consuming to identify papers with

Table 5. Current uses of text mining and desired uses

Status	Specific use cases of text-mining tools
Current	• Finding gene names and symbols (gene indexing) • Querying full text with Textpresso • Assigning GO cellular component terms
Future/desired	• Improving gene indexing results • Performing document triage • Recognizing additional biological concepts (disease, anatomy) • Capturing terms from additional ontologies (e.g. GO, particularly molecular function and biological process) • Capturing complex relations such as gene regulation

curatable content, document ranking tools can be extremely valuable. Aerts *et al.* (21) reported that by using text-mining methods, they were able to prioritize some 30 000 papers containing unannotated *cis*-regulatory information within PubMed (out of millions of articles).

There has been some earlier work to quantify the impact and utility of text-mining tools for document ranking, indexing and curation (20–26). For example, the PreBIND system (22) was able to locate protein–protein interaction data in the literature; it was found to reduce task duration by 70%. Van Auken *et al.* (20) found that use of Textpresso for curating protein subcellular localization had the potential for significant speed up compared to manual curation (between 8- and 15-fold faster). Given the wider uptake of text-mining tools, it will be important to revisit this question and to build more sophisticated models of the costs and benefits of bringing tools into the workflow, including time spent on development/adaptation of tools to a specific database, as well as time spent training curators to use the tools.

To explore issues of how text-mining tools can assist curators, BioCreative created an interactive track starting with BioCreative III (27) and continued as Track III of the 2012 workshop (28). Findings from the earlier BioCreatives (2–4) suggested that text-mining tools could help with steps such as gene indexing or with mappings to specific ontologies (GO). In BioCreative II.5, authors had difficulties in linking genes and proteins to the correct specifies-specific Entrez Gene or UniProt identifiers, a task where an interactive tool could be very helpful. Providing such capabilities would make it possible to leverage additional resources, e.g. authors, for help with curation. The FlyBase curators have improved throughput in their system by asking authors to provide 'skim curation' of newly submitted articles—thus circumventing the need for triage and also speeding up the curation process (24). The success of Textpresso in curation of GO subcellular localization (20) is also a good example of helping the curator to find evidence and to create the correct mappings into a terminology or ontology.

As tools improve, we expect to see new insert points and new success stories. For example, Textpresso is working on capture of GO molecular function terms; such extensions may be facilitated by new tools on the ontology side, such as BioAnnotator (29). In addition, several of the systems, e.g. PubTator (30), in the Interactive Track (Track III) are working hand in hand with biological database curators to provide extraction of a wider range of biological entities (e.g. drugs, diseases), as well as extraction of relationships between these entities along with pointers to the underlying evidence.

We believe that BioCreative has been critical in bringing together the text mining and the biocurator communities; going forward, we expect to see increasing numbers of partnerships and increasing uptake of text-mining tools into curation workflows. This will require a balance between inserting tools tailored to the needs of a particular database and its workflow versus the need to develop generic text-mining tools that can be rapidly tailored to specific tasks. It has been a working hypothesis of BioCreative that by posing generic challenge tasks (bio-entity extraction and indexing, document ranking for triage, relation extraction); we can encourage the development of an inventory of capabilities that can then be rapidly adapted to the specific needs of biocurators. We plan to measure our success in BioCreative IV, in particular, by focusing on interactive systems, as well as improving interoperability of existing components.

In conclusion, we have analyzed and reviewed curation workflow descriptions from seven independent curation groups. Based on this analysis, we have identified both common and database-specific aspects of literature curation between 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 mentioned above, working in close partnership between the biological database curators and the text-mining tool developers.

Acknowledgements

We would like to thank the other BioCreative 2012 organizers for helpful discussion and the Track II teams for their participation.

Funding

The US National Science Foundation Grant [DBI-0850319 (to L.H.)]; the NIH Intramural Research Program, National Library of Medicine, National Institutes of Health (to Z.L.). Funding for open access charge: The MITRE Corporation.

Conflict of interest. None declared.

References

1. Arighi,C.N., Lu,Z., Krallinger,M. *et al.* (2011) Overview of the BioCreative III Workshop. *BMC Bioinformatics*, **12**(Suppl. 8), S1.
2. 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.
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. 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.
5. Wu,C.H., Arighi,C., Cohen,K.B. *et al.* (2012) Editorial: BioCreative-2012 virtual issue. *Database*, doi: 10.1093/database/bas049.
6. Yeh,A., Morgan,A., Colosimo,M. *et al.* (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, **6**(Suppl. 1), S2.
7. Smith,L., Tanabe,L.K., Ando,R.J. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.*, **9**(Suppl. 2), S2.
8. Hirschman,L., Colosimo,M., Morgan,A. *et al.* (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6**(Suppl. 1), S11.
9. Morgan,A.A., Lu,Z., Wang,X. *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9**(Suppl. 2), S3.
10. Lu,Z., Kao,H.Y., Wei,C.H. *et al.* (2011) The gene normalization task in BioCreative III. *BMC Bioinformatics*, **12**(Suppl. 8), S2.
11. Blaschke,C., Leon,E.A., Krallinger,M. *et al.* (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, **6**(Suppl. 1), S16.
12. 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.
13. Krallinger,M., Leitner,F., Rodriguez-Penagos,C. *et al.* (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.*, **9**(Suppl. 2), S4.
14. Hirschman,L., Burns,G.A., Krallinger,M. *et al.* (2012) Text mining for the biocuration workflow. *Database*, doi: 10.1093/database/bas020.
15. Dowell,K.G., McAndrews-Hill,M.S., Hill,D.P. *et al.* (2009) Integrating text mining into the MGI biocuration workflow. *Database*, doi: 10.1093/database/bap019.
16. 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.
17. 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*, doi: 10.1093/database/bas017.
18. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
19. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
20. 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.
21. Aerts,S., Haeussler,M., van Vooren,S. *et al.* (2008) Text-mining assisted regulatory annotation. *Genome Biol.*, **9**, R31.
22. Donaldson,I., Martin,J., de Bruijn,B. *et al.* (2003) PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.
23. Alex,B., Grover,C., Haddow,B. *et al.* (2008) Assisted curation: does text mining really help? *Pac. Symp. Biocomput.*, **2008**, 556–567.
24. Karamanis,N., Lewin,I., Seal,R. *et al.* (2007) Integrating natural language processing with FlyBase curation. *Pac. Symp. Biocomput.*, **2007**, 245–256.
25. Wang,P., Morgan,A.A., Zhang,Q. *et al.* (2007) Automating document classification for the Immune Epitope Database. *BMC Bioinformatics*, **8**, 269.
26. 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.
27. Arighi,C.N., Roberts,P.M., Agarwal,S. *et al.* (2011) BioCreative III interactive task: an overview. *BMC Bioinformatics*, **12**(Suppl. 8), S4.
28. Arighi,C., Carterette,B., Cohen,K.B. *et al.* (2012) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database*, in press.
29. Jonquet,C., Shah,N.H. and Musen,M.A. (2009) The open biomedical annotator. *Summit on Translat. Bioinforma*, **2009**, 56–60.
30. 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*, doi: 10.1093/database/bas041.