doi:10.1093/nar/gkt441

PubTator: a web-based text mining tool for assisting biocuration

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

[1]National Center for Biotechnology Information, US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA and [2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, 701, Taiwan, R.O.C

Received March 1, 2013; Revised April 19, 2013; Accepted April 29, 2013

ABSTRACT

Manually curating knowledge from biomedical literature into structured databases is highly expensive and time-consuming, making it difficult to keep pace with the rapid growth of the literature. There is therefore a pressing need to assist biocuration with automated text mining tools. Here, we describe PubTator, a web-based system for assisting biocuration. PubTator is different from the few existing tools by featuring a PubMed-like interface, which many biocurators find familiar, and being equipped with multiple challenge-winning text mining algorithms to ensure the quality of its automatic results. Through a formal evaluation with two external user groups, PubTator was shown to be capable of improving both the efficiency and accuracy of manual curation. PubTator is publicly available at http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/.

INTRODUCTION

Current biomedical research has become heavily dependent on the online access to knowledge in expert-curated biological databases. Manual curation is often required to build these knowledge bases, which involves biocurators reading articles, extracting key findings and cross-referencing data. Biocuration has become an essential part of biological discovery and biomedical research (1–3). However, as the volume of biological literature grows rapidly, it becomes increasingly difficult for biocurators to keep pace with the literature because manual biocuration is a highly expensive and time-consuming endeavour. To help ease the burden of manual curation, there have been increasing efforts to use automatic text-mining techniques (4–12), including finding gene names and symbols, prioritizing documents for curation and assigning ontology concepts. In response to a call for participation in BioCreative 2012 Interactive Text Mining task (13), we developed PubTator, a web-based application that provides computer assistance to biocurators (14).

PubTator has several unique features that distinguish it from existing annotation and literature search tools (15–17), as it is designed specifically for the needs of biocurators who have limited text-mining experience. First, PubTator is a web-based system; thus, no installation is required and not restricted to any specific computer platforms. Second, PubTator is an all-in-one system that provides one-stop service for literature curation from searching and retrieving relevant articles to annotating selected articles. As such, user input can either be a search query or a list of PubMed articles. When manual curation is completed, users can readily download and export their annotations for database integration. Third, PubTator is designed in a PubMed-like interface, which many biocurators find it to be familiar and easy to use with minimal training required. Fourth, multiple competition-winning text-mining approaches have been integrated into PubTator for automatically identifying key biological entities (18,19). Hence, it provides state-of-the-art performance on generating automatic computer pre-annotations in computer-assisted biocuration. Finally, PubTator is adaptable to different annotation tasks and also allows its users to personalize their own annotation environment.

SYSTEM DESCRIPTION

Pre-annotating PubMed articles using text-mining tools

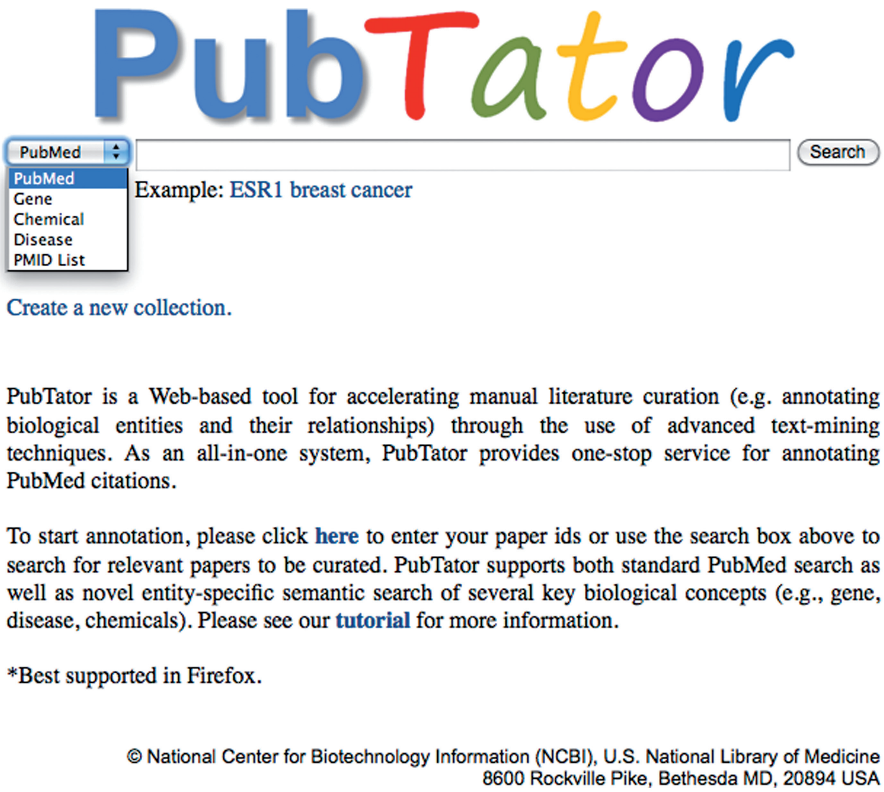
PubTator houses the entire content of PubMed and keeps current with nightly updates. To enable entity-specific semantic searches and provide pre-annotations for computer-assisted biocuration, automatic text-mining tools are applied to all articles with respect to genes, diseases, species, chemicals and mutations. More specifically, we not only find the occurrences of those entities in text but also map all entity mentions to standard database or controlled vocabulary identifiers as shown in Table 1. To ensure high quality of automatically processed results, we used tools that have been extensively evaluated for

*To whom correspondence should be addressed. Tel: +1 301 594 7089; Fax: +1 301 480 2288; Email: zhiyong.lu@nih.gov

Table 1. Text-mining tools used for pre-annotating bio-entities in PubMed articles

Bio-entity	Text-mining tool	Nomenclature	F_1 score (%)
Gene (mention)	GeneTUKit	N/A	82.97
Gene (normalization)	GenNorm	NCBI Gene	92.89
Disease	DNorm	MEDIC	80.90
Species	SR4GN	NCBI Taxonomy	85.42
Chemical	A dictionary-based lookup approach	MeSH	53.82
Mutation	tmVar	NCBI dbSNP (rs#) or tmVar normalized forms	93.98

The reported F_1 scores (http://en.wikipedia.org/wiki/F1_score) of different tools were either taken from their corresponding publications or assessed by us on public benchmarking datasets. MEDIC is a disease vocabulary created by Comparative Toxicogenomics Database. All other vocabularies are products of National Library of Medicine. Separate tools are used for identifying gene names in abstracts (mention) and assigning NCBI Gene identifiers to those mentions (normalization).



Example: ESR1 breast cancer

Create a new collection.

PubTator is a Web-based tool for accelerating manual literature curation (e.g. annotating biological entities and their relationships) through the use of advanced text-mining techniques. As an all-in-one system, PubTator provides one-stop service for annotating PubMed citations.

To start annotation, please click here to enter your paper ids or use the search box above to search for relevant papers to be curated. PubTator supports both standard PubMed search as well as novel entity-specific semantic search of several key biological concepts (e.g., gene, disease, chemicals). Please see our tutorial for more information.

*Best supported in Firefox.

© National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Figure 1. The PubTator homepage with five different search options.

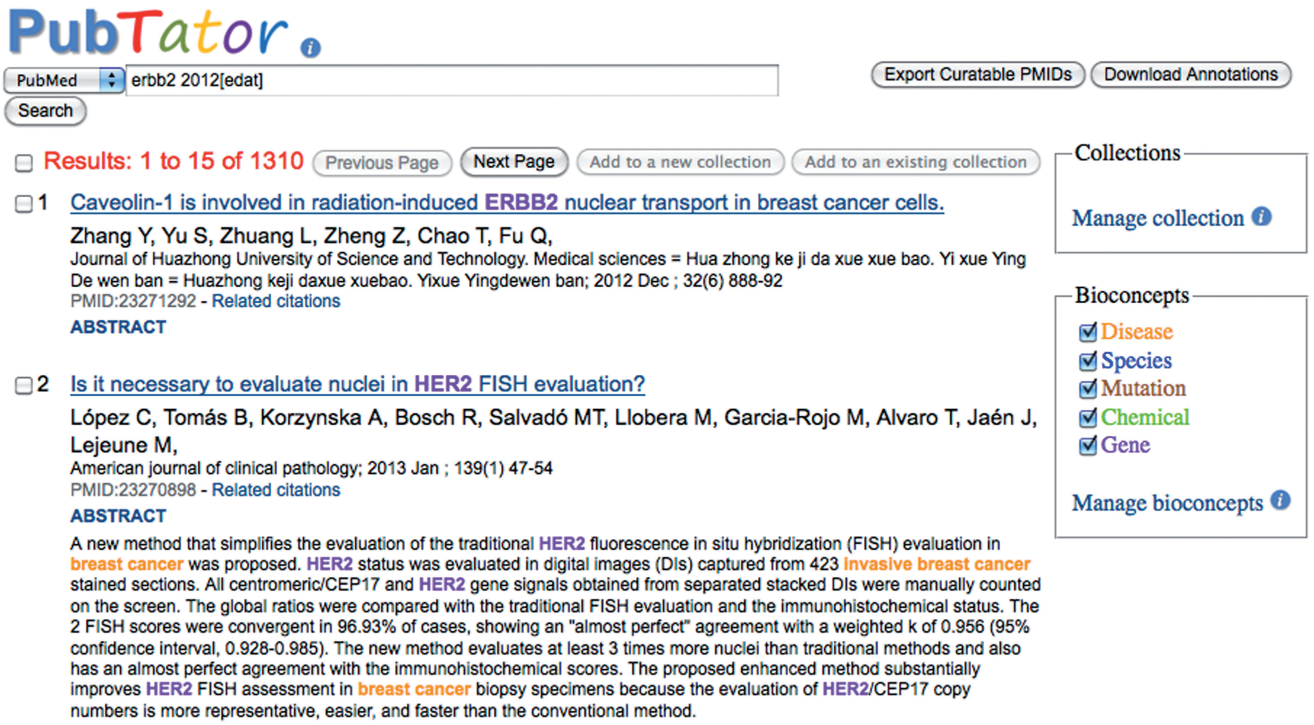
superlative performance in various text-mining competition events. Our entity recognition tools include GeneTUKit (19) for gene mention, GenNorm (18) for gene normalization, SR4GN (20) for species, DNorm (Leaman et al., 2013, under consideration; http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/DNorm/) for diseases, tmVar (21) for mutations and a dictionary-based lookup approach (8) for chemicals. SR4GN was also used for associating recognized species with their corresponding gene/protein mentions so that we were able to perform cross-species gene normalization in PubTator.

Search function in PubTator

PubTator supports both keyword searches and semantic searches with respect to specific bio-entities. As shown in Figure 1, five search options are currently available in PubTator:

- PubMed: return results identical to PubMed search results
- Gene: return all articles relevant to a specific gene or gene product
- Chemical: return all articles relevant to a specific chemical
- Disease: return all articles relevant to a specific disease or syndrome
- PMID List: return articles in the PubMed Identifier (PMID) upload order

The first search option (PubMed) is implemented using the NCBI's Entrez Programming Utilities Web service



□ Results: 1 to 15 of 1310

□ 1 Caveolin-1 is involved in radiation-induced ERBB2 nuclear transport in breast cancer cells.

Zhang Y, Yu S, Zhuang L, Zheng Z, Chao T, Fu Q,
Journal of Huazhong University of Science and Technology. Medical sciences = Hua zhong ke ji da xue xue bao. Yi xue Ying De wen ban = Huazhong keji daxue xuebao. Yixue Yingdewen ban; 2012 Dec ; 32(6) 888-92
PMID:23271292 - Related citations
ABSTRACT

□ 2 Is it necessary to evaluate nuclei in HER2 FISH evaluation?

López C, Tomás B, Korzynska A, Bosch R, Salvadó MT, Llobera M, Garcia-Rojo M, Alvaro T, Jaén J, Lejeune M,
American journal of clinical pathology; 2013 Jan ; 139(1) 47-54
PMID:23270898 - Related citations
ABSTRACT

A new method that simplifies the evaluation of the traditional HER2 fluorescence in situ hybridization (FISH) evaluation in breast cancer was proposed. HER2 status was evaluated in digital images (DIs) captured from 423 invasive breast cancer stained sections. All centromeric/CEP17 and HER2 gene signals obtained from separated stacked DIs were manually counted on the screen. The global ratios were compared with the traditional FISH evaluation and the immunohistochemical status. The 2 FISH scores were convergent in 96.93% of cases, showing an "almost perfect" agreement with a weighted k of 0.956 (95% confidence interval, 0.928-0.985). The new method evaluates at least 3 times more nuclei than traditional methods and also has an almost perfect agreement with the immunohistochemical scores. The proposed enhanced method substantially improves HER2 FISH assessment in breast cancer biopsy specimens because the evaluation of HER2/CEP17 copy numbers is more representative, easier, and faster than the conventional method.

Figure 2. The PubTator search results page. Automatically computed entities are highlighted in colours. Unlike PubMed, article abstracts can be displayed here without going to a different page.

API (http://www.ncbi.nlm.nih.gov/books/NBK25500/). The next three semantic search options are based on pre-computed results of the different text-mining tools as shown in Table 1. As biological entities are often associated with multiple names, our semantic search feature allows users to retrieve all the articles relevant to an entity without having to enumerate the entire set of possible aliases (22). For instance, searching for the breast cancer gene ERBB2 will also retrieve articles containing only its alternative names such as HER2 (e.g. see Result 2 in Figure 2). The last search option (PMID List) is provided for users who already have a list of relevant articles for curation.

Same as PubMed, PubTator returns search results in the reverse chronological order for all search options except PMID List. However, only 15 results are returned per page in PubTator instead of 20 in PubMed, making it possible for users to glance at the abstract on the search results page as shown in Figure 2.

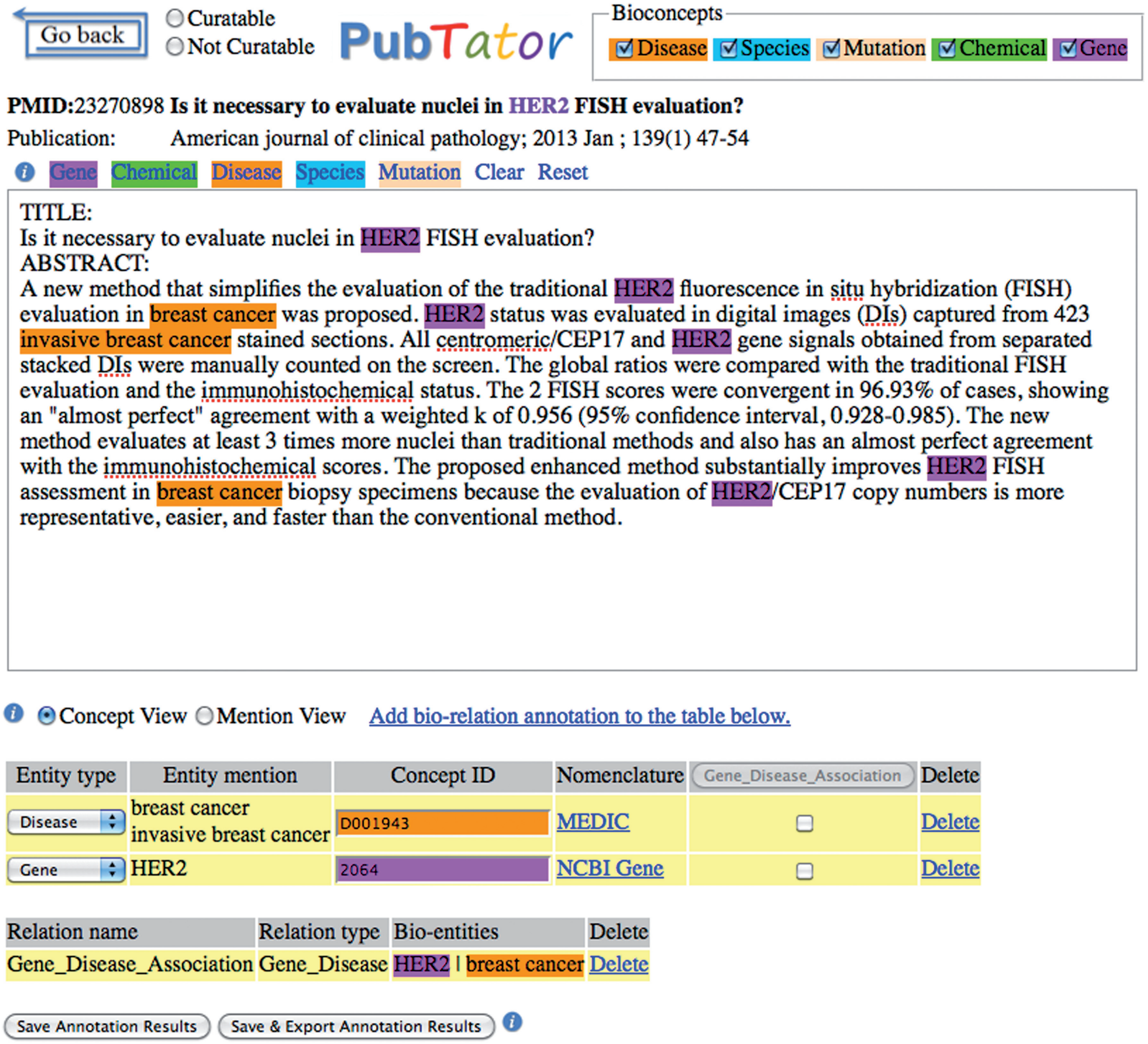
Different from PubMed, pre-computed biological entities are highlighted in each article when applicable: gene (purple), chemicals (green), diseases (orange), mutation (brown) and species (blue). A search filter (by taxonomy) is provided for those biocuration teams who work with a specific organism because by default we show results across all species.

Annotation function in PubTator

Currently, PubTator supports three annotation tasks: document triage, entity annotation and relationship annotation. In document triage, biocurators are engaged in selecting and prioritizing curatable articles based on the reading of the article. As a pre-step for full curation, users can readily identify curatable articles in two simple mechanisms using PubTator: First, a user can select articles from the search results by simply checking the box next to the articles (see Figure 2). Second, a user can indicate whether an article is curatable at the top of the annotation page (see Figure 3).

PubTator can be used for annotating bio-entities of any kind by following steps detailed in our online tutorial page (http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/tutorial/index.html#DefineBioconcepts). PubTator provides automated pre-annotations for five common types (shown in Table 1). As shown in Figure 3, pre-computed bio-entities are highlighted in colours in the text box and also displayed in the table below where both mentions and corresponding identifiers are stored. A user can modify and remove an existing annotation as well as insert a new one. To improve efficiency, once a new annotation is made to an entity, there is an option to propagate the annotation throughout the article for the same entity. Once completed, all annotations will be saved to our database for download.

Finally, PubTator can be used for annotating relationships between entities (http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/tutorial/index.html#DefineBiorelations). PubTator allows curators to specify the kind of relations they desire to capture from literature, which can be either between the same kind of entities, such as protein–protein interactions, or between different kinds, such as gene–disease relations. PubTator ensures that the entity types selected by the user are consistent with what is specified by the relationship definition.


Go back
○Curatable
○Not Curatable
PubTator
Bioconcepts
☑Disease ☑Species ☑Mutation ☑Chemical ☑Gene

PMID:23270898 **Is it necessary to evaluate nuclei in HER2 FISH evaluation?**

Publication: American journal of clinical pathology; 2013 Jan ; 139(1) 47-54

Gene Chemical Disease Species Mutation Clear Reset

TITLE:
Is it necessary to evaluate nuclei in HER2 FISH evaluation?
ABSTRACT:
A new method that simplifies the evaluation of the traditional HER2 fluorescence in situ hybridization (FISH) evaluation in breast cancer was proposed. HER2 status was evaluated in digital images (DIs) captured from 423 invasive breast cancer stained sections. All centromeric/CEP17 and HER2 gene signals obtained from separated stacked DIs were manually counted on the screen. The global ratios were compared with the traditional FISH evaluation and the immunohistochemical status. The 2 FISH scores were convergent in 96.93% of cases, showing an "almost perfect" agreement with a weighted k of 0.956 (95% confidence interval, 0.928-0.985). The new method evaluates at least 3 times more nuclei than traditional methods and also has an almost perfect agreement with the immunohistochemical scores. The proposed enhanced method substantially improves HER2 FISH assessment in breast cancer biopsy specimens because the evaluation of HER2/CEP17 copy numbers is more representative, easier, and faster than the conventional method.

● Concept View ○ Mention View Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	Gene_Disease_Association	Delete
Disease	breast cancer invasive breast cancer	D001943	MEDIC	□	Delete
Gene	HER2	2064	NCBI Gene	□	Delete

Relation name	Relation type	Bio-entities	Delete
Gene_Disease_Association	Gene_Disease	HER2 | breast cancer	Delete

Figure 3. The PubTator annotation page. The two radio buttons (Curatable/Not Curatable) at the top of the page is designed for document triage. The text box and the table below are used for entity annotation. The relationship table at the bottom of the page is for relationship annotation. In Mention View, each row corresponds to an entity mention. In Concept View (default), different mentions of the same concept (i.e. having the same identifier) are combined and displayed in the same row.

System adaptability

Instead of being a tool for a specific curation group, we aim to make PubTator adaptable to different curation needs. For instance, PubTator allows its users to define their own entity types and controlled vocabularies for annotating mentions and their corresponding concept identifiers, respectively. This is particularly useful for assigning gene and protein identifiers where curators from model organism groups may prefer using their own gene nomenclature (e.g. the Arabidopsis Genome Initiative locus identifiers) as opposed to the default NCBI Gene identifiers. However, for user-defined entity types or nomenclature, PubTator does not provide automatic pre-annotations.

In addition to PubMed articles, PubTator may also be used to process other types of biomedical text (e.g. annotating grants data). In such cases, the input text can be first uploaded to PubTator according to a specific format and then immediately processed by different text-mining tools on the fly.

EVALUTION RESULTS

PubTator has been formally evaluated through its participation in the interactive text-mining track of BioCreative 2012 workshop (13). PubTator improved both manual curation efficiency curation and accuracy in user studies for two curation tasks: document triage and gene indexing (23). After the task, the interactive text-mining track organizers conducted a survey to help identify strengths and weakness of different systems with regards to system design, usability and so forth. The survey results show that PubTator has top ratings in many aspects of biocuration from system design, to learnability, to usability. Overall, PubTator was the highest rated and most recommended among all participating systems (13).

CONCLUSIONS

There is an increasing need for automatic computer tools to assist many biocuration tasks, including prioritizing articles for full curation and annotating key biological concepts. PubTator was developed in response to these needs. In particular, PubTator provides users with many advanced text-mining tools through an easy-to-use graphical interface that is accessible through the web. Based on the previous user studies, we believe PubTator can provide practical benefits to biocurators in their routine curation work. Future work includes further improvement of existing text-mining algorithms and the integration additional text-mining tools for better support of ontology concept annotation, which was identified as a critical need in biocuration in recent studies (6,24). We also plan to investigate different search algorithms and full-text process in the future PubTator development.

ACKNOWLEDGEMENTS

The authors are grateful to Larry Smith, Don Comeau and Rezarta Dogan for their work on the PubTator prototype. They also thank John Wilbur, Sun Kim and Bethany Harris for helpful discussion, and Robert Leaman for proofreading the manuscript.

FUNDING

Intramural Research Program of the NIH, National Library of Medicine. Funding for open access charge: U.S. National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Burge,S., Attwood,T.K., Bateman,A., Berardini,T.Z., Cherry,M., O'Donovan,C., Xenarios,L. and Gaudet,P. (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database (Oxford)*, **2012**, bar059.
2. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., Pierre,S.S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
3. Bourne,P.E. and McEntyre,J. (2006) Biocurators: contributors to the world of science. *PLoS Comput. Biol.*, **2**, e142.
4. Vishnyakova,D., Pasche,E. and Ruch,P. (2012) Using binary classification to prioritize and curate articles for the Comparative Toxicogenomics Database. *Database (Oxford)*, **2012**, bas050.
5. Névéol,A., Wilbur,W.J. and Lu,Z. (2012) Improving links between literature and biological data with text mining: a case study with GEO, PDB and MEDLINE. *Database (Oxford)*, **2012**, bas026.
6. Lu,Z. and Hirschman,L. (2012) Biocuration workflows and text mining: overview of the BioCreative Workshop Track II. *Database (Oxford)*, **2012**, bas043.
7. Rinaldi,F., Clematide,S., Garten,Y., Whirl-Carrillo,M., Gong,L., Hebert,J.M., Sangkuhl,K., Thorn,C.F., Klein,T.E. and Altman,R.B. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database (Oxford)*, **2012**, bas021.
8. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2012) Collaborative biocuration–text-mining development task for document prioritization for curation. *Database (Oxford)*, **2012**, bas037.
9. Auken,K.V., Jaffery,J., Chan,J., Müller,H.-M. and Sternberg,P.W. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (GO) cellular component curation. *BMC Bioinformatics*, **10**, 228.
10. Yu,W., Clyne,M., Dolan,S.M., Yesupriya,A., Wulf,A., Liu,T., Khoury,M.J. and Gwinn,M. (2008) GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, **9**, 205.
11. Dowell,K.G., McAndrews-Hill,M.S., Hill,D.P., Drabkin,H.J. and Blake,J.A. (2009) Integrating text mining into the MGI biocuration workflow. *Database*, **2009**, bap019.
12. Krallinger,M., Leitner,F., Vazquez,M., Salgado,D., Marcelle,C., Tyers,M., Valencia,A. and Chatr-aryamontri,A. (2012) How to link ontologies and protein-protein interactions to literature: text-mining approaches and the BioCreative experience. *Database (Oxford)*, **2012**, bas017.
13. Arighi,C.N., Roberts,P.M., Agarwal,S., Bhattacharya,S., Cesareni,G., Chatr-aryamontri,A., Clematide,S., Gaudet,P., Giglio,M.G., Harrow,I. *et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database*, **2013**, bas056.
14. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2012) PubTator: A PubMed-Like interactive curation system for document triage and literature curation. In: *Proceedings of the BioCreative 2012 Workshop*. Washington, DC, USA, pp. 145–150.
15. Neves,M. and Leser,U. (2012) A survey on annotation tools for the biomedical literature. *Brief. Bioinformatics*, December 18 (doi: 10.1093/bib/bbs084; epub ahead of print).
16. Müller,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.
17. Salgado,D., Krallinger,M., Depaule,M., Drula,E., Tendulkar,A.V., Leitner,F., Valencia,A. and Marcelle,C. (2012) MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, **28**, 2285–2287.
18. Wei,C.-H. and Kao,H.-Y. (2011) Cross-species gene normalization by species inference. *BMC Bioinformatics*, **12**, S6.
19. Huang,M., Liu,J. and Zhu,X. (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.
20. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PLoS One*, **7**, e38460.
21. Wei,C.-H., Harris,B.R., Kao,H.-Y. and Lu,Z. (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433–1439.
22. Divoli,A., Hearst,M.A. and Wooldridge,M.A. (2008) Evidence for showing gene/protein name suggestions in bioscience literature search interfaces. *Pac. Symp. Biocomput.*, 568–579.
23. Wei,C.-H., Harris,B.R., Li,D., Berardini,T.Z., Huala,E., Kao,H.-Y. and Lu,Z. (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database (Oxford)*, **2012**, bas041.
24. Laulederkind,S.J., Tutaj,M., Shimoyama,M., Hayman,G.T., Lowry,T.F., Nigam,R., Petri,V., Smith,J.R., Wang,S.-J., de Pons,J. *et al.* (2012) Ontology searching and browsing at the Rat Genome Database. *Database (Oxford)*, **2012**, bas016.