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

Chih-Hsuan Wei1,2, Hung-Yu Kao2 and Zhiyong Lu1,*

1National Center for Biotechnology Information, US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA and 2Department 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 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

Published by Oxford University Press 2013. This work is written by US Government employees and is in the public domain in the US.
superlative performance in various text-mining competition events. Our entity recognition tools include GeneTUKit (19) for gene mention, GenNorm (18) for gene normalization, 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.

<table>
<thead>
<tr>
<th>Bio-entity</th>
<th>Text-mining tool</th>
<th>Nomenclature</th>
<th>F1 score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene (mention)</td>
<td>GeneTUKit</td>
<td>N/A</td>
<td>82.97</td>
</tr>
<tr>
<td>Gene (normalization)</td>
<td>GenNorm</td>
<td>NCBI Gene</td>
<td>92.89</td>
</tr>
<tr>
<td>Disease</td>
<td>DNorm</td>
<td>MEDIC</td>
<td>80.90</td>
</tr>
<tr>
<td>Species</td>
<td>SR4GN</td>
<td>NCBI Taxonomy</td>
<td>85.42</td>
</tr>
<tr>
<td>Chemical</td>
<td>A dictionary-based lookup approach</td>
<td>MeSH</td>
<td>53.82</td>
</tr>
<tr>
<td>Mutation</td>
<td>tmVar</td>
<td>NCBI dbSNP (rsii) or tmVar normalized forms</td>
<td>93.98</td>
</tr>
</tbody>
</table>

The reported F1 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).
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.
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.

**EVALUATION 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 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

Conflict of interest statement. None declared.

REFERENCES