BioQRator: a web-based interactive biomedical literature curating system

Dongseop Kwon[1], Sun Kim[2], Soo-Yong Shin[3], John Wilbur[2]

[1]Department of Computer Engineering, Myoungji University, South Korea

[2]National Center for Biotechnology Information, National Institutes of Health, USA

[3]Department of Biomedical Informatics, Asan Medical Center, South Korea

Availability: http://www.bioqrator.org

Contact: sun.kim@nih.gov

1. Introduction

BioQRator is a web-based annotation tool for biomedical literature. This tool was designed to support any task annotating entities and relationships. It is also the first web tool which supports the BioC format [1] for annotation. For input, any documents in the BioC format and PubMed abstracts can be used. For output, annotated documents can be saved in a BioC format file as well. Our goal in the BioCreative IV IAT task focuses on the following two topics.

1) Develop a general-purpose annotation tool for entities and relationships. This tool is essentially a web interface which can be fully customized for a given task. To assist an annotation task, text mining resources can be utilized through the BioC format.
2) Apply and evaluate PIE *the search* [2] for a protein-protein interaction (PPI) annotation task. PIE *the search* is a web interface for searching PubMed literature for protein interaction information and the main method is based on a winning approach in BioCreative III [3]. In BioCreative IV IAT, the practical usability of PIE *the search* will be studied.

Here, we show basic functions of BioQRator and the performance of PIE *the search*. In addition, we propose a PPI annotation task for the BioCreative IV IAT task. The proposed task is based on the TRIP Database [4, 5], which contains manually curated protein-protein interactions for mammalian TRP (Transient Receptor Potential) channels. BioQRator and PIE *the search* will be used to assist TRIP DB annotations and it will be evaluated by annotation time and prediction accuracy. In addition, there is a possible collaboration with BioGrid [6], which will help improve the user interface of BioQRator.

2. BioQRator

BioQRator was designed as an easy-to-use tool to annotate any entities and relationships in text. In particular, most annotations can be done by a series of single mouse clicks (or drags) with simple typing. Since BioQRator was implemented using HTML5/CSS to support multiple browsers. It is compatible with the latest version of popular browsers such as Chrome, Safari and Firefox (Internet Explorer is only partially supported due to certain HTML5 compatibility issues). Here is the scenario of how to use BioQRator.

1) Sign in (or sign up if there is no account)
2) Create a collection: A user can create a collection by several different methods.
   A. From a web browser: A collection name is required. Source, date and key information can be optionally entered.
   B. From a BioC format file: All necessary information including pre-annotated documents is automatically loaded using the uploaded BioC file.
3) Add documents
   A. Unless documents are loaded from a BioC file, a user should add documents into an empty collection (Figure 1). Currently, we provide two options: “Search documents with a PubMed query” and “Upload a PMID list from a file”. Both options retrieve documents from PubMed, however retrieval results are sorted by PIE score in default. A higher PIE score means there is more possibility that the document may have PPI information.

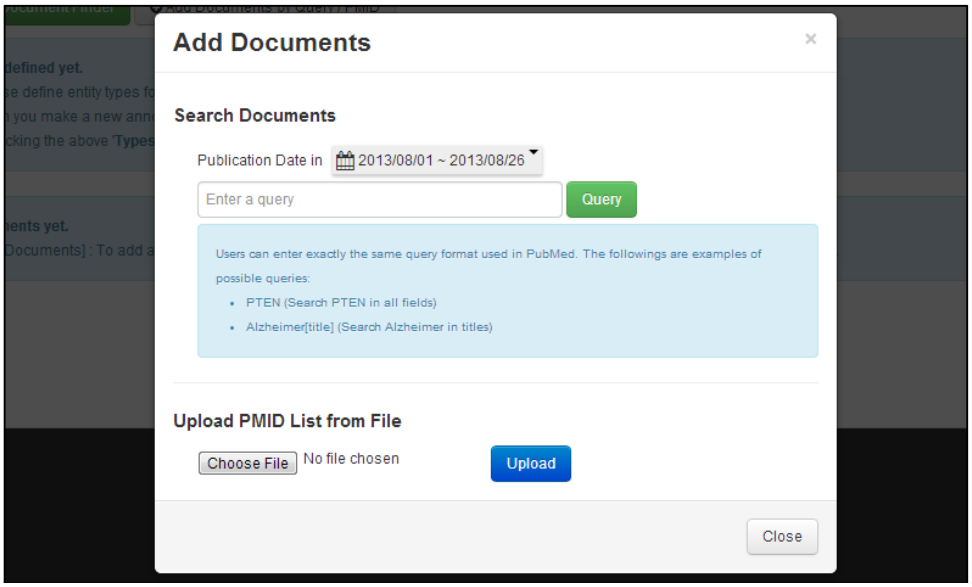

Figure 1. Adding documents to a collection.

For accessing PubMed IDs from a PubMed query, PubMed E-utils are used. Since this procedure increases the processing time significantly, we plan to change this mechanism in the near future.

B. Smart document finder: This is a convenient tool for periodically adding documents with a fixed query. A user will be able to set automatic document search weekly, monthly, quarterly, or even yearly.

C. After searching PubMed or PIE *the search*, a user can manually add any documents of interest by clicking "Add to Collection" (Figure 2).

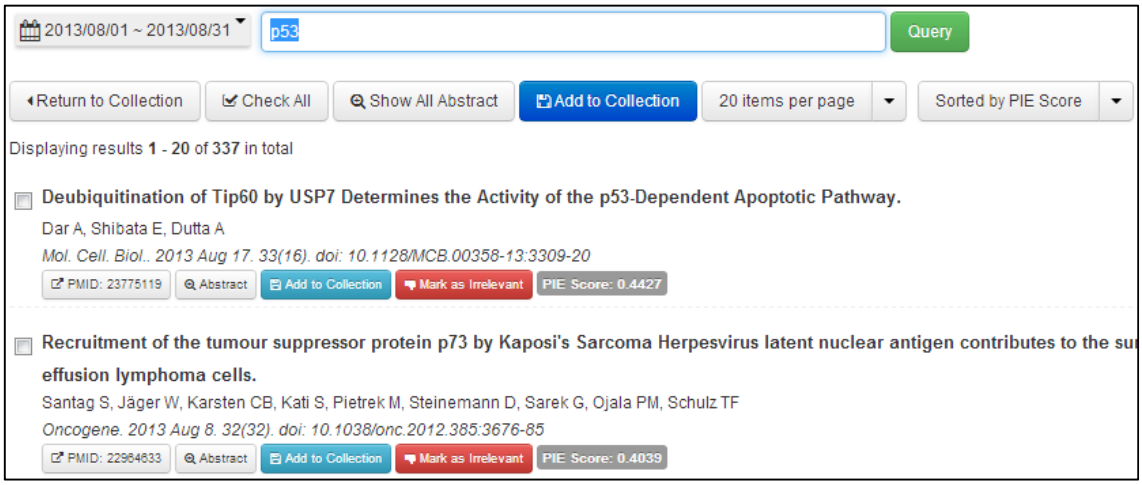

Figure 2. PubMed search results.

D. Adding documents is flexible. New documents or BioC files can be added to an existing collection any time.

4) Annotate documents

A. After adding a set of documents in a collection, a user can start annotating entities and relations. For uploaded BioC files, pre-annotated entities and relations will be automatically shown in the annotation window. However, for an entirely new collection, entity and relation types to be used should be defined first by using the "Types" tab.

B. For annotating an entity, a user can do a single click or drag the mouse to select the whole entity name. Once a mouse click or drag is done, a pop-up window will appear and a user can fill in necessary information. For normalizing gene/protein names, Entrez Gene and UniProt searches are provided in default. Entrez Gene or UniProt IDs can be easily assigned through this search process. Note that "Annotation ID" in this window is different from Entrez Gene or UniProt IDs. The annotation ID identifies the annotation uniquely and does not represent an ID in a database such as Entrez Gene or UniProt IDs.

C. For PubMed abstracts, pre-annotated PPI entities will be available. A user can use this information by clicking “Open PIE *the search* Annotations” (Figure 3).

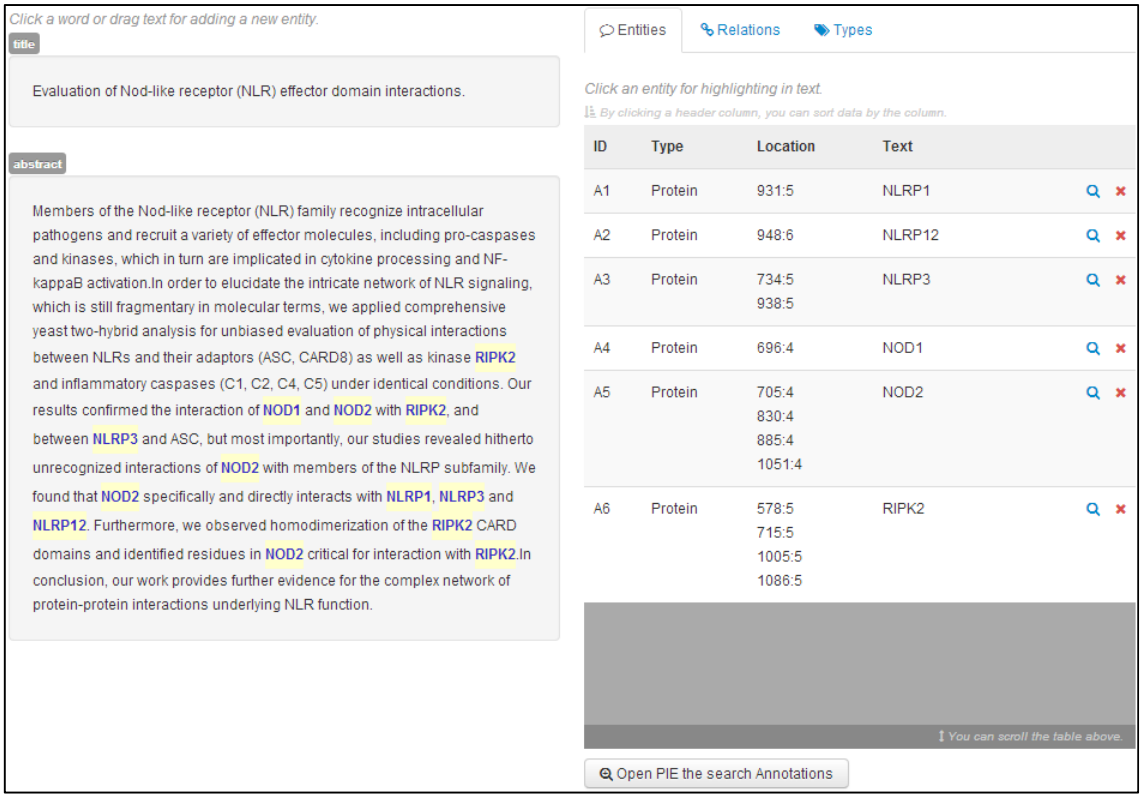

Figure 3. Annotating entities and relations.

5) Download a collection: Annotated documents in a collection can be saved as a BioC format file. BioC was developed to easily share text documents and annotations among different tools. Since BioCreative IV took the BioC initiative as one of its main tasks, we decided to fully support BioC as standard input and output file format.
6) Share a collection: A collection can be shared with other users. This function is enabled if other users are added through the “Share” button.

3. Performance of PIE *the search*

To support PPI annotations, article ranking and entity information from PIE *the search* was migrated to BioQRator. In previous work, we evaluated article ranking performance using the BioCreative III ACT (BC3) dataset [3]. For F1, MCC and AUC iP/R measures, PIE *the search* showed 0.6258, 0.5610 and 0.6834 respectively. However, the medians of BC3 participant results were 0.5353 F1, 0.4563 MCC and 0.5367 AUC iP/R. Table 1 shows the precisions of PIE *the search* at rank *N* for the BC3 test

set. Since PubMed abstracts can be sorted based on PPI scores in BioQRator, the performance at top-ranked documents is more important than overall classification performance in this regard. Hence, the table shows the usefulness of PIE *the search* as a PPI informative article search tool.

Table 1. Ranking performance of PIE *the search*.

Top N	Precision
10	1.0000
50	0.9600
100	0.9400
200	0.9150
300	0.8467
400	0.8125
500	0.7680

For identifying gene/protein names, the Priority Model [7] is utilized in PIE *the search*. Since not all entities are important in PPI annotations, we only mark predicted gene/protein names which are used to identify PPI informative articles. In [7], the Priority Model showed 0.9200, 0.9690 and 0.9440 for precision, recall and F1 scores respectively on the experiments using SemCat [8].

4. Proposed Tasks for BioCreative IV Track 5 (IAT)

For BioCreative IV IAT, our focus is on two goals: the usability of BioQRator as a general-purpose annotation tool and the effectiveness of PIE *the search* as a supporting tool for PPI annotations. To achieve these goals, we propose a PPI annotation task for the TRIP DB (http://www.trpchannel.org).

The TRIP DB is a manually curated database of PPIs for mammalian TRP channel since 2010 [4, 5]. The TRIP DB contains 646 PPI pairs among 28 TRP channels and 394 cellular proteins by curating 369 articles as of June 2013. The contents of the TRIP DB are updated every two or three months depending on newly published TRP channel literature. The curators of the TRIP DB regularly check new publications using PubMed and annotate the relevant articles to extract PPI information based on the format in http://www.trpchannel.org/download. In addition, they collect diverse information using many external resources including HGNC, Entrez Gene, UniProt, DIP, IntAct, MINT, BioGrid, STRING, IUPHAR-DB, KEGG, OMIM and GO.

For using BioQRator as an annotation tool for the TRIP DB, the proposed tasks are as follows:

1) Search PubMed abstracts and sort the results based on relevance to PPI information: ranking performance can be used as an evaluation measure.

2) Provide an automatic annotation of genes/proteins and an editing capability: manual annotations are a time-consuming task. Reducing annotation time by using BioQRator is a main interest in the proposed task.
3) BioC compatibility: Supporting BioC as standard input and output format does not solve all the interoperability issues. Synchronizing locations of entities and character codes (e.g., UTF-8 and ASCII) among different tools is a crucial problem. We plan to address these issues by communicating with other BioC developers.

Acknowledgements

DK was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2012R1A1A2044389 and 2011-0022437). SSY was supported by a grant (2012-543) from the Asan Institute for Life Science, Seoul, Korea. SK and WJW were supported by Intramural Research Program of the NIH, National Library of Medicine.

References

1. BioC: a minimalist approach to interoperability for biomedical text processing. http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/
2. S. Kim, D. Kwon, S.-Y. Shin, and W. J. Wilbur (2012). PIE *the search*: searching PubMed literature for protein interaction information. Bioinformatics 28(4): 597-8.
3. S. Kim and W. J. Wilbur (2011). Classifying protein-protein interaction articles using word and syntactic features. BMC Bioinformatics 12(Suppl 8): S9.
4. Y.-C. Shin, S.-Y. Shin, I. So, D. Kwon, and J.-H. Jeon (2011). TRIP Database: a manually curated database of protein-protein interactions for mammalian TRP channels. Nucl. Acids Res. 39 (suppl 1): D356-D361.
5. Y.-C. Shin, S.-Y. Shin, J. N. Chun, H. S. Cho, J. M. Lim, H.-G. Kim, I. So, D. Kwon, and J.-H. Jeon (2012). TRIP Database 2.0: a manually curated information hub for accessing TRP channel interaction network. PLOS ONE 7(10): e47165.
6. BioGrid: a biological general repository for interaction datasets. http://thebiogrid.org/
7. L. Tanabe and W. J. Wilbur (2006). A priority model for named entities. Proceedings of the BioNLP Workshop on Linking Natural Language and Biology (LNLBioNLP ’06), pp.33-40.
8. L. Tanabe, L. H. Thom, W. Matten, D. C. Comeau, and W. J. Wilbur (2006). SemCat: semantically categorized entities for genomics. AMIA Annual Symposium Proceedings, pp. 754-758.