BioQRator: a web-based interactive biomedical literature curating system

Dongseop Kwon[1], Sun Kim[2,*], Soo-Yong Shin[3], and W. 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

*Corresponding author: Tel: 301 496 2484, E-mail: sun.kim@nih.gov

Introduction

BioQRator (http://www.bioqrator.org) is a web-based annotation tool for biomedical literature. This tool was designed to support any task annotating entities and relationships. It is also one of the first web tools which support 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.

System Description

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 browsers such as Chrome, Safari and Firefox. Here is the scenario of how to use BioQRator.

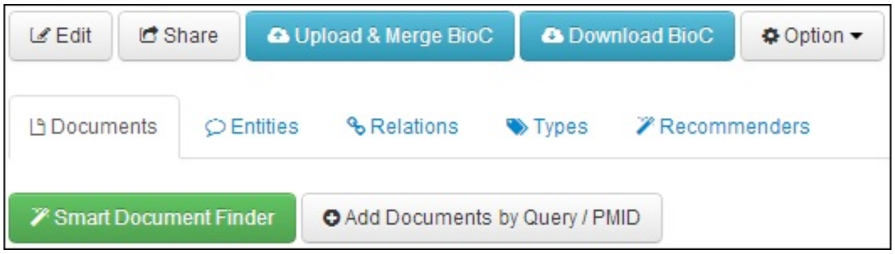
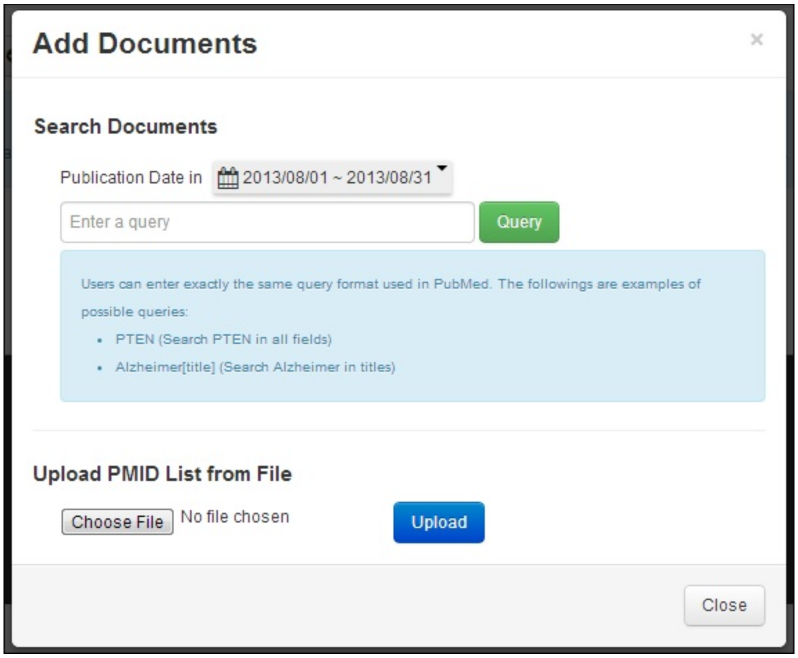


Figure 1. The main window of an empty collection.

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) Create entity and relation types: Given a collection, the next task is to associate with the collection those entity and relation types which will be used to annotate the collection. A user can create the entity and/or relation types by going to the "Types" tab (Figure 1) or during document annotation. The "Recommenders" tab (Figure 1) is used for setting up external resources to find relevant information for an entity name. However, BioQRator supports Entrez Gene and UniProt Recommenders in default.
4) Add documents
   A. Unless documents are loaded from a BioC file, a user should add documents into an empty collection. Currently, we provide two options: "Search documents with a PubMed query" and "Upload a PMID list from a file" (Figure 2). 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.
   B. Smart document finder (Figure 1): 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 select and add any documents of interest by clicking "Add to Collection" (Figure 3). The "Abstract" button below each document title can be used for a quick look at abstracts in the same window. The "PMID: 23775119" button is used for reading abstracts/full text through the PubMed service. "Mark as Irrelevant" is a special feature in the BioQRator search. If a document is marked as irrelevant, it will not appear in future retrieval results in the same collection.
   D. Adding documents is flexible. New documents or BioC files can be added to an existing collection any time.



Add Documents

Search Documents

Publication Date in 2013/08/01 ~ 2013/08/31

Enter a query

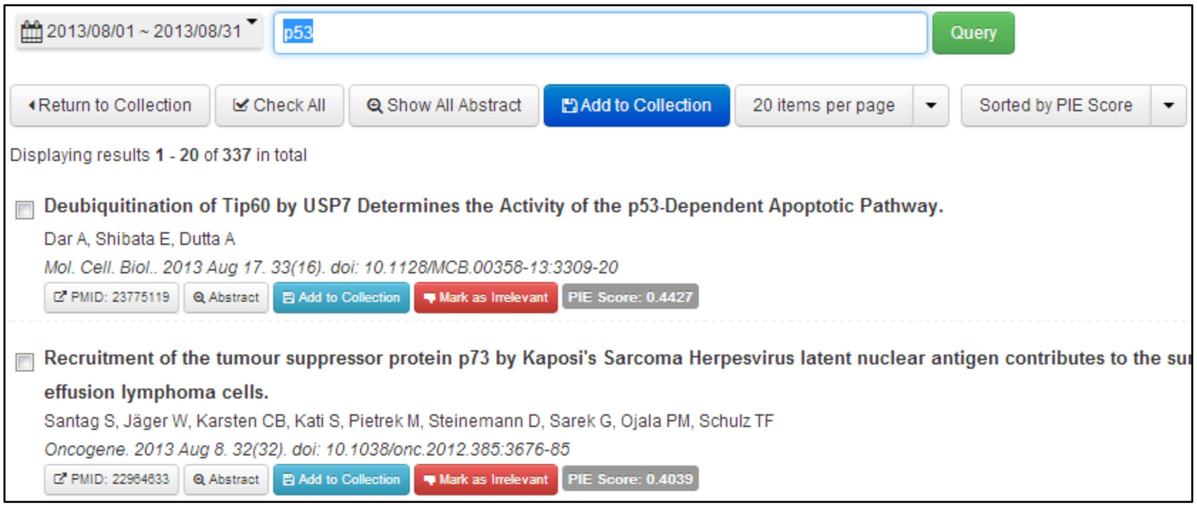
Users can enter exactly the same query format used in PubMed. The followings are examples of possible queries:

- PTEN (Search PTEN in all fields)
- Alzheimer[title] (Search Alzheimer in titles)

Upload PMID List from File

No file chosen

Figure 2. Adding documents to a collection.



2013/08/01 ~ 2013/08/31

p53

Displaying results **1** - **20** of **337** in total

□ **Deubiquitination of Tip60 by USP7 Determines the Activity of the p53-Dependent Apoptotic Pathway.**

Dar A, Shibata E, Dutta A

Mol. Cell. Biol.. 2013 Aug 17. 33(16). doi: 10.1128/MCB.00358-13:3309-20

PMID: 23775119 | Abstract | Add to Collection | Mark as Irrelevant | PIE Score: 0.4427

□ **Recruitment of the tumour suppressor protein p73 by Kaposi's Sarcoma Herpesvirus latent nuclear antigen contributes to the su effusion lymphoma cells.**

Santag S, Jäger W, Karsten CB, Kati S, Pietrek M, Steinemann D, Sarek G, Ojala PM, Schulz TF

Oncogene. 2013 Aug 8. 32(32). doi: 10.1038/onc.2012.385:3676-85

PMID: 22964633 | Abstract | Add to Collection | Mark as Irrelevant | PIE Score: 0.4039

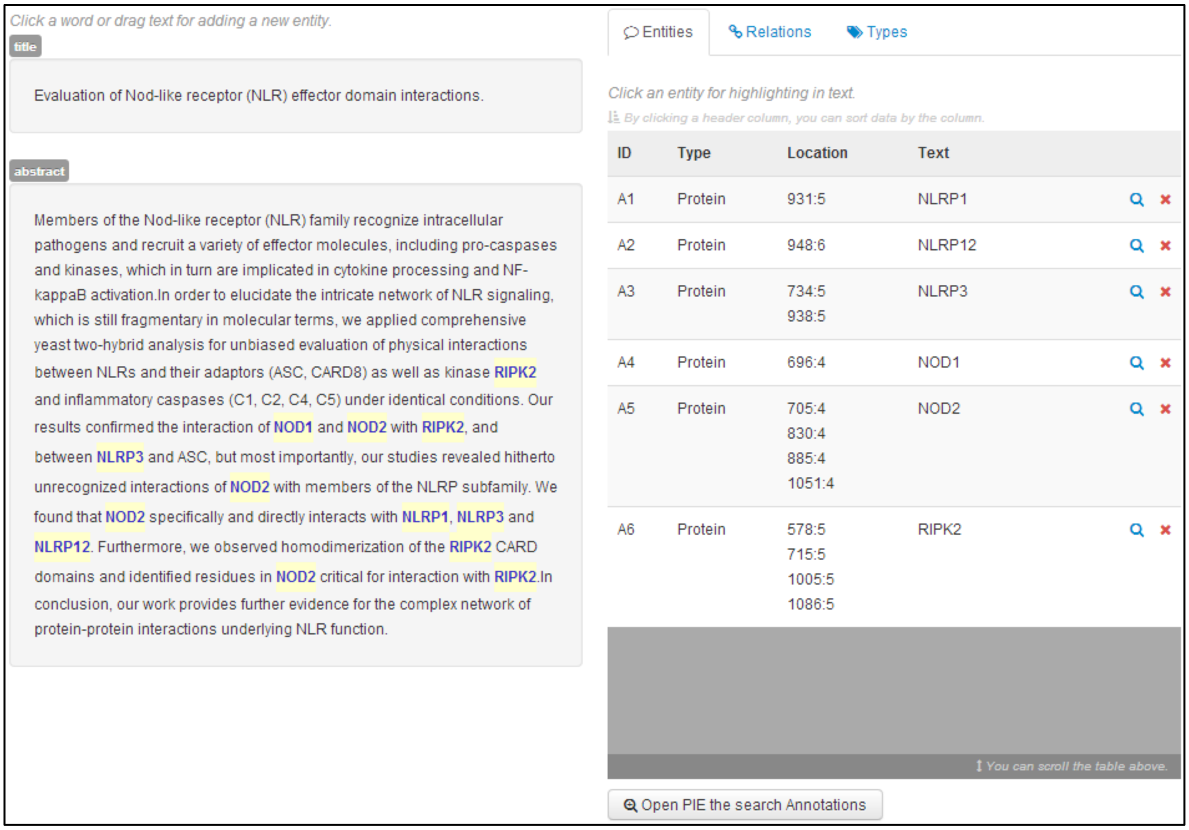
Figure 3. PubMed search results.

5) Annotate documents

A. After adding a set of documents to form 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.

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. Normally, a user does not need to assign annotation IDs because BioQRator automatically assign the IDs unless specified.

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 4).



Click a word or drag text for adding a new entity.

title

Evaluation of Nod-like receptor (NLR) effector domain interactions.

abstract

Members of the Nod-like receptor (NLR) family recognize intracellular pathogens and recruit a variety of effector molecules, including pro-caspases and kinases, which in turn are implicated in cytokine processing and NF-kappaB activation.In order to elucidate the intricate network of NLR signaling, which is still fragmentary in molecular terms, we applied comprehensive yeast two-hybrid analysis for unbiased evaluation of physical interactions between NLRs and their adaptors (ASC, CARD8) as well as kinase RIPK2 and inflammatory caspases (C1, C2, C4, C5) under identical conditions. Our results confirmed the interaction of NOD1 and NOD2 with RIPK2, and between NLRP3 and ASC, but most importantly, our studies revealed hitherto unrecognized interactions of NOD2 with members of the NLRP subfamily. We found that NOD2 specifically and directly interacts with NLRP1, NLRP3 and NLRP12. Furthermore, we observed homodimerization of the RIPK2 CARD domains and identified residues in NOD2 critical for interaction with RIPK2.In conclusion, our work provides further evidence for the complex network of protein-protein interactions underlying NLR function.

Click an entity for highlighting in text.

By clicking a header column, you can sort data by the column.

ID	Type	Location	Text
A1	Protein	931:5	NLRP1
A2	Protein	948:6	NLRP12
A3	Protein	734:5 938:5	NLRP3
A4	Protein	696:4	NOD1
A5	Protein	705:4 830:4 885:4 1051:4	NOD2
A6	Protein	578:5 715:5 1005:5 1086:5	RIPK2

You can scroll the table above.

Open PIE the search Annotations

Figure 4. Annotating entities and relations.

6) 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 the standard input and output file format.

7) Share a collection: A collection can be shared with other users. This function is enabled if other users are added through the "Share" button in a collection (Figure 1).

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 (2), we evaluated article ranking performance using the BioCreative III ACT (BC3) dataset (4). 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 (5) 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 (5), the Priority Model showed 0.9200, 0.9690 and 0.9440 for precision, recall and F1 scores respectively on the experiments using SemCat (6).

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, 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. Search results obtained from BioQRator will be compared with those from PubMed.
2) Annotate PPI-relevant interactions and normalize protein names: 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 and SYS were 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; 2012R1A1A2002804), respectively. SK and WJW were supported by Intramural Research Program of the NIH, National Library of Medicine.

References

1. Comeau,D.C., Dogan,R.I., Ciccarese,P. *et al*. (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, **2013**, bat064.
2. Kim,S., Kwon,D., Shin,S.-Y. *et al*. (2012) PIE *the search*: searching PubMed literature for protein interaction information. *Bioinformatics*, **28**(4), 597-598.
3. Kim,S. and Wilbur,W.J. (2011) Classifying protein-protein interaction articles using word and syntactic features. *BMC Bioinformatics*, **12**(Suppl 8), S9.
4. 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.
5. Tanabe,L. and Wilbur,W.J. (2006) A priority model for named entities. *Proceedings of the BioNLP Workshop on Linking Natural Language and Biology (LNLBioNLP '06)*, 33-40.
6. Tanabe,L., Thom,L.H., Matten,W., *et al*. (2006) SemCat: semantically categorized entities for genomics. *AMIA Annual Symposium Proceedings*, 754-758.