# Collaborative Curation of Data from Bio-medical Texts and Abstracts and its integration

Chitta Baral, Hasan Davulcu, Mutsumi Nakamura, Prabhdeep Singh, Luis Tari, and Lian Yu

Department of Computer Science and Engineering,
Arizona State University,
PO Box 878809, Tempe, AZ 85287-8809, USA
{chitta,hdavulcu,mutsumi,prabhdeep,luis.tari,lianyu}@asu.edu

**Abstract.** We propose an inexpensive and scalable approach for curation that takes advantage of automatic information extraction methods as a starting point, and is based on the premise that if there are a lot of articles, then there must be a lot of readers and authors of these articles. Thus we provide a mechanism by which the readers of the articles can participate and collaborate in the curation of information.

## 1 Introduction

Besides the data that exists in various public and private databases, there is a much larger and ever increasing amount of information buried in existing biomedical articles. It is beyond human ability to read the various relevant articles and recall relevant findings of these articles for further research. Therefore, it becomes clear that the findings in these articles have to be culled and stored in a database such that the data can be integrated with other existing databases. The sheer volume of the articles and their constant growth makes it prohibitively expensive to employ *(and monetarily compensate)* human curators to read through the articles and cull the necessary knowledge/data buried in them. Nevertheless, such human curation (see for example [1,3-7,21]) has been tried for specific domains. Due to the issue of cost, many of the curated databases are proprietary with limited coverage.

In recent years an alternative approach of using automatic text extraction systems [2,8-20] has been proposed. Although good progress has been made in this area, the systems are not fool-proof. They at times infer incorrect information or miss out important information. Moreover, most existing systems focus on simpler data forms, such as identifying gene or protein names, simple interactions without context. Sometimes such simplicity may lead to inconsistency.

In this paper we propose a solution to the problem of curating information from the large and growing body of biomedical texts and abstracts. We propose a methodology where the community collaboratively contributes to the curation process. We use automatic information extraction methods as a starting point, and promote mass collaboration with the premise that if there are a lot of articles, then there must be a lot of readers and authors of these articles.

## 2 CBioC System Architecture

The two main components of our CBioC system are (i) the CBioC interface and (ii) the CBioC database. The user interacts with the CBioC system through the CBioC interface. When a user views a PubMed article, the CBioC interface is automatically invoked to display all the extracted interaction data relevant to the article. The user curates the extracted interaction data through voting. Depending on the access level, an user can also enter or modify data.

The CBioC interface has many subcomponents such as the automatic invocation component, the user and access management component, and the voting and other interactions component. Two auxiliary components of the system are (a) a suite of automated text extraction systems and (b) a data exchange system. The text extraction systems are used to automatically extract data from texts and abstracts and the data exchange system is used to download relevant data from existing databases (such as [7,9-13] ) and convert them to our format. This is illustrated in Figure 1 below.

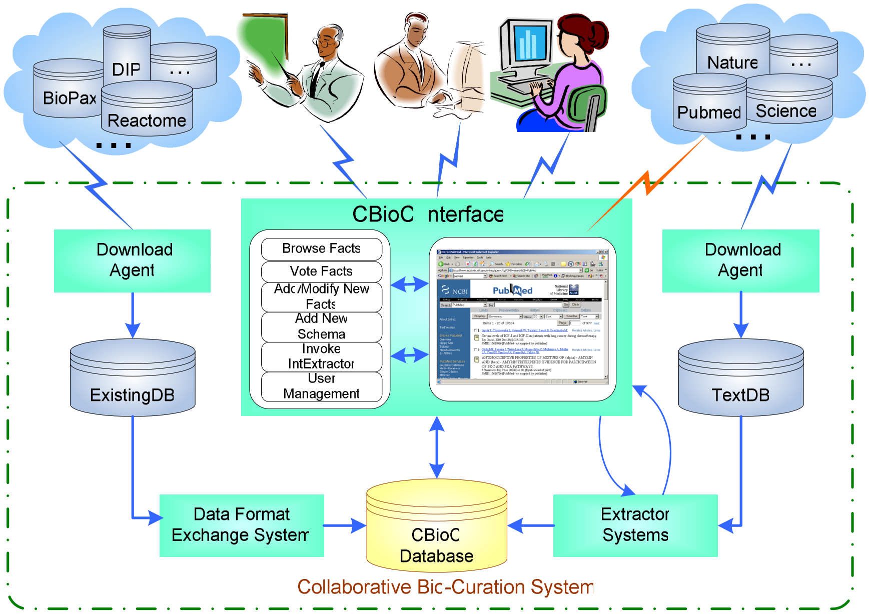
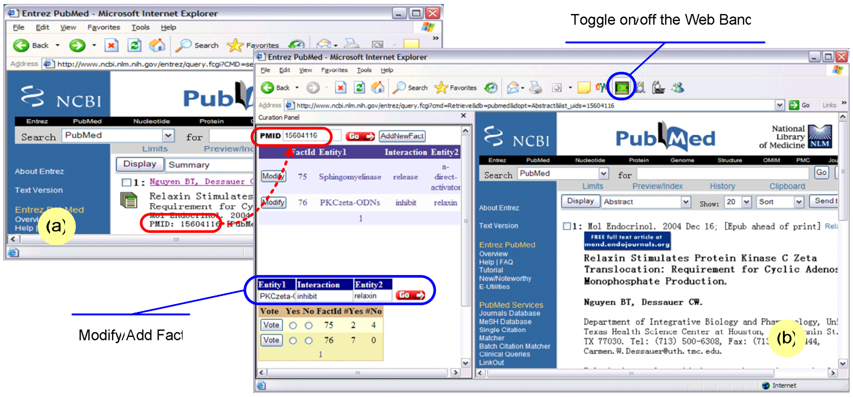


**Fig. 1.** Functional architecture of the CBioC System

We now illustrate the use of the CBioC system which also further illuminates on the architecture of the CBioC system.

**Installation and Invocation**: An important goal of ours is to make it easier for a researcher to participate in the collaborative curation. For that a researcher has to download our system and install it in her computer. Once the system is installed it watches the researcher's access of the web through Internet Explorer windows. Whenever the researcher accesses a web page from where she can access an article or an abstract, the CBioC system is invoked and an interaction frame is created, as shown below in Figure 2.



Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=se

□1: Nguyen BT, Dessauer C
Relaxin Stimulates
Requirement for Cy
Mol Endocrinol. 2004
PMID: 15604116 [PubMe

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15604116

PMID 15604116

| | FactId | Entity1 | Interaction | Entity2 |
|---|---|---|---|---|
| Modify | 75 | Sphingomyelinase | release | a-direct-activator |
| Modify | 76 | PKCzeta-ODNs | inhibit | relaxin |

1

| Entity1 | Interaction | Entity2 |
|---|---|---|
| PKCzeta-( | inhibit | relaxin |

| Vote | Yes | No | FactId | #Yes | #No |
|---|---|---|---|---|---|
| Vote | ○ | ○ | 75 | 2 | 4 |
| Vote | ○ | ○ | 76 | 7 | 0 |

1

□1: Mol Endocrinol. 2004 Dec 16; [Epub ahead of print] Rela

FREE full text article at mend.endojournals.org

Relaxin Stimulates Protein Kinase C Zeta Translocation: Requirement for Cyclic Adenos Monophosphate Production.

Nguyen BT, Dessauer CW.

Department of Integrative Biology and Pharm ology, Uni Texas Health Science Center at Houston, nin St. TX 77030. Tel: (713) 500-6308, Fax: (71 44, Carmen.W.Dessauer@uth.tmc.edu.

**Fig.2.** Automatic triggering of CBioC interaction frame

**System Implementation**: From the implementation angle, the CBioC system consists of three main parts: (i) Web forms and connection to database; (ii) WebBand and Browser Helper components, and (iii) Connector to Interaction Extractor, and is currently implemented for Internet Explorer in the client side and Linux-MySQL-Php on the server side. This is illustrated in Figure 3 below.

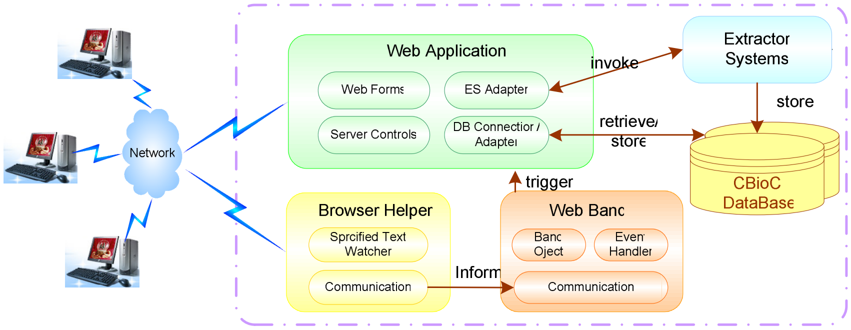


**Fig. 3** Implementation Architecture of CBioC System

**Acknowledgement**: Deepthi, Toufeeq, Haishan, Luping, Ning, Drs. Kambhampati, Das, Berens, Fukuda, Bittner, and Gaasterland and, NSF grant 0412000.

## References

[1] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., and Hogue, C.W. (2001) BIND-The biomolecular interaction Netwoek database. *Nucleic Ac. R.* **29**, 242-245.

[2] Rzhetsky, A. et al. (2004) Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics* **27**, 43-53.

[3] Stein, Lincoln (2002), Creating a bioinformatics nation, *Nature*, **417**, 119-120.

[4] Xenarios, I. and Eisenberg, D. (2001) Protein interacting databases. *Current Opinion in Biotechnology*. **12**, 334-339.

[5] KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

[6] BIND: Interaction Network Database, http://www.bind.ca

[7] HPRD: Human Protein Reference Database, http://www.hprd.org/

[8] Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998) Toward information extraction: identifying protein names from biological papers. *PSB 1998*, 707-718

[9] Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics.* 2002 Aug;18(8):1124-1132.

[10] Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proceedings of the International Conference on Intelligent System Molecular Biology*. 1999, 60-67.

[11] Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics. 2001* Feb;17(2):155-561.

[12] Novichkova, S., Egorov, S., and Daraselia, N. (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics. 2003* September. **19(13),** 1699-1706

[13] Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*. 2001, **17** Suppl 1:S74-82.

[14] Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P.A., Weng, W., Wilbur, W.J., Hatzivassiloglou, V., and Friedman, C. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform.* 2004 February, **37(1),** 43-53.

[15] Corney, D.P., Buxton, B.F., Langdon, W.B., and Jones, D.T. (2004) BioRAT: extracting biological information from full-length papers. *Bioinformatics*. 2004 November 22, **20(17)**, 3206-3213. Epub 2004 Nov 22.

[16] Temkin, J.M. and Gilder, M.R. (2003) Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinf.* 2003 Nov 1, **19(16),** 2046-2053.

[17] Chiang, J.H., Yu, H.C., and Hsu, H.J. (2004) GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics.* 2004 Jan 1, **20(1),** 120-121.

[18] Craven, M. and Kumlien, J. (1999) Constructing biological knowledge bases by extracting information from text sources. *Proceedings of International Conference on Intelligent System Molecular Biology.* 1999, 77-86.

[19] Bunescu, R., Ge, R., Kate, R.K., Marcotte, E.M., Mooney, R.J., Ramani, A.K., and Wong, Y.W. (2004) Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Journal Artificial Intelligence in Medicine* 2004.

[20] Ding, J., Berleant, D., Xu, J., and Fulmer, A. (2003) Extracting biochemical interactions from MEDLINE using a link grammar parser. *In Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03),* 467. IEEE Computer Society, 2003.

[21] www.biocurator.org