CBioC: BEYOND A PROTOTYPE FOR COLLABORATIVE ANNOTATION OF MOLECULAR INTERACTIONS FROM THE LITERATURE

C. Baral, G. Gonzalez, A. Gitter, C. Teegarden, and A. Zeigler

School of Computing and Informatics, Arizona State University
Tempe, AZ 85281, USA
Email: chitta@asu.edu, ggonzalez@asu.edu

G. Joshi-Topé
Northeast Biosciences, Inc
New York, NY, USA

In molecular biology research, looking for information on a particular entity such as a gene or a protein may lead to thousands of articles, making it impossible for a researcher to individually read these articles and even just their abstracts. Thus, there is a need to curate the literature to get various nuggets of knowledge, such as an interaction between two proteins, and store them in a database. However the body of existing biomedical articles is growing at a very fast rate, making it impossible to curate them manually. An alternative approach of using computers for automatic extraction has problems with accuracy. We propose to leverage the advantages of both techniques, extracting binary relationships between biological entities automatically from the biomedical literature and providing a platform that allows community collaboration in the annotation of the extracted relationships. Thus, the community of researchers that writes and reads the biomedical texts can use the server for searching our database of extracted facts, and as an easy-to-use web platform to annotate facts relevant to them. We presented a preliminary prototype as a proof of concept earlier[1]. This paper presents the working implementation available for download at http://www.cbioc.org as a browser-plug in for both Internet Explorer and FireFox. This current version has been available since June of 2006, and has over 160 registered users from around the world. Aside from its use as an annotation tool, data from CBioC has also been used in computational methods with encouraging results[2].

1. INTRODUCTION

There are about 15 million abstracts currently indexed in PubMed, with anywhere between 300,000 and 500,000[3] being added each year. To illustrate this problem, consider the following example. A search for the gene TNF alpha in PubMed yields 74430 articles (as of March of 2007) and 6193 review articles. Refining the search to TNF alpha and inflammation reduces this number to 15126 regular articles and 1757 review articles, still too many for a researcher to review. It would be significantly easier if he or she had access to a database that stores relevant nuggets of knowledge such as the relationship between genes and biological processes. The problem of constructing such a database has been recognized as one that needs to be solved to move forward into the great challenges of science for this century[4].

Currently, two approaches are used to extract such facts from biomedical publications: (i) human curation and (ii) development and use of automated information extraction systems. However, the constantly increasing number of articles and the complexity inherent to its annotation results in data sources that are continuously outdated. For example, GeneRIF (Gene Reference Into Function), was started in 2002, yet it covers only about 1.7% of the genes in Entrez[5] and 25% of human genes.

Automatic extraction and annotation seems a natural way to overcome the limitations of manual curation, and a lot of work has been done in this area, including the automatic extraction of genes and gene products[6], protein-protein interactions[7], relationships between genes and biological functions[8], and genes and diseases[9], among others. However, the reliability of the extracted information varies greatly and thus discourages the biologists from using it for their research.

CBioC represents a new approach to the problem through mass collaboration, where the community of researchers that writes and reads the biomedical texts will be able to contribute to the curation process, dictating the pace at which it is done. Automated text extraction is used

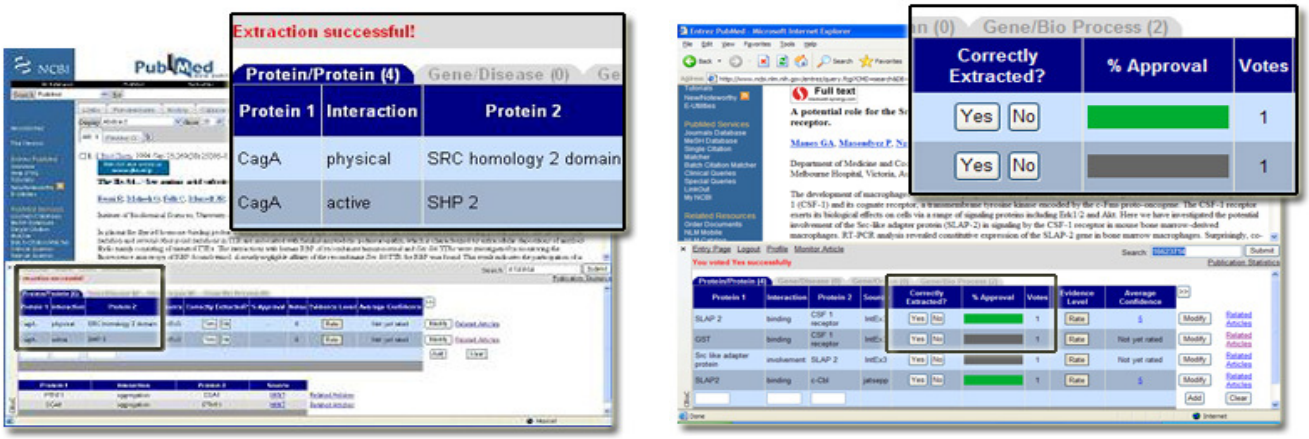

Figure 1. CBioC automatically launches the interaction web band at the bottom of the main window when a user visits PubMed, and displays the facts available for it. If the abstract has not been processed, extraction occurs "on the fly". The left image corresponds to the interactions display, allowing the user to tab through the different kinds of relationships (protein/protein, gene/disease, and gene/bioprocess). The right image shows the simplest annotation mechanism (a yes/no vote for "Correctly Extracted") and the agreement level (% Approval). Users may also modify and add interactions.

as a starting point to bootstrap the database but then it is up to researchers to improve upon the extracted data by modifications, additions of missed facts, and voting on the accuracy of extraction. It runs as a web browser extension and allows unobtrusive use of the system during the regular course of research, allowing the natural "checks and balances" of community consensus to take hold to resolve inconsistencies when possible, or to point out disagreements and controversial findings. Although most of the data in CBioC is currently from automatic extraction, users have contributed over 500 interactions which are currently being evaluated. This shows how with CBioC, small or large groups of researchers can easily annotate articles and find facts of interest to them.

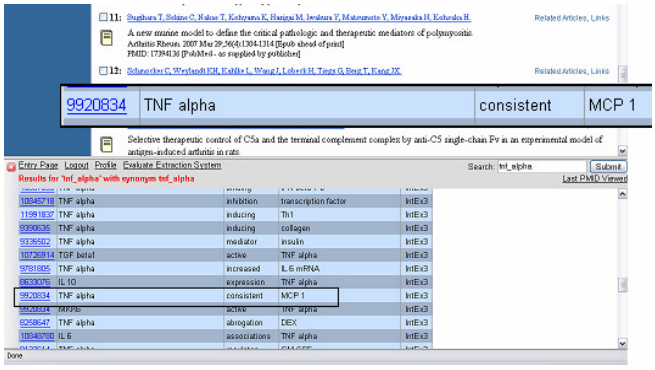

Figure 2. A CBioC search provides a simple way to browse through interactions involving a particular gene or the list of genes involved in a disease or biological process.

2. METHODS

CBioC is available for both Internet Explorer and Firefox, for PCs, Macs, and Linux machines. Once installed, CBioC runs unobtrusively, and when one visits the Entrez (PubMed) web site, CBioC automatically opens within a "web band" at the bottom of the main browser. Users that do not wish to install the plug in can get similar functionality by logging in from our home page.

CBioC uses a modified version of the extraction system IntEx[7] (dubbed IntEx3) that uses Natural Language Processing methods to extract protein-protein interactions, gene-disease relations, and gene-bioprocess relations.

2.1. Usage

Consider a variation of the research scenario introduced before. A PubMed search for "TNF alpha atherosclerosis" returns over 900 abstracts. One of the abstracts (PMID 16814297), reports TNF-alpha modulates MCP-1, a common alias to CCL2. Expression of CCL2 has been found to be increased in cardiovascular diseases and is of high interest as a biomarker of atherosclerosis[10]. However, as of March 2007, none of the public curated databases had captured this important interaction, and any researcher that missed the article will probably not learn about it. CCL2 is involved in immunoregulatory and inflammatory processes[11]. Thus, that TNF-alpha modulates CCL2 as

reported in the article (supported by others, such as PMID 9920834) is significant, important to assess the relevance of TNF alpha with respect to atherosclerosis and for any systems biology simulations. Thus, relying solely on curated data could leave this piece of information out.

Consider the same scenario, but with CBioC installed. The user could start with a TNF alpha search in CBioC (see Figure 2). Quickly scrolling down through the listed interactions gives the researcher a general idea of the known relevant associated genes, even though some of them might not be accurate. If MCP 1 calls the researcher's attention, the rest of the interactions in that abstract can be quickly displayed by clicking on the PMID of the interaction of interest among the search results. A list of other articles that report the same interaction can be viewed by clicking on the "Related Articles" link.

2.2. Functionality

2.2.1. *Displaying data*

When one searches the PubMed database and displays a particular abstract, CBioC automatically displays the interactions found related to the abstract. If the abstract has not been processed by CBioC before, an extraction system runs "on the fly". CBioC also displays interactions found for the article in publicly accessible databases.

2.2.2. *Searching*

As a registered CBioC user, one can search the CBioC database for all facts related to a particular protein, gene, disease, or interaction word by simply typing the relevant term in the Search box within the CBioC web band. CBioC automatically expands a search term with known synonyms of the term. One can also display the facts available for a set of abstracts by typing a comma-separated list of their PMIDs in the search box. The search box also lets one see all the facts we have from a particular database by typing its name, such as "BIND" or "MINT".

2.2.3. *Modifying, and adding*

Registered CBioC users can vote on the accuracy of an extraction, modify the interactions, or add interactions that the extraction system missed. If the interaction seems correctly extracted, one can click the "Yes" button to approve. Otherwise, one can vote "No" or modify the data by clicking "Modify". If "Modify" is clicked, the data fields open up for editing. The user's screen id will be displayed in the "Source" column from then on, with the previous data stored and accessible via the "History" link. The modified information is then subject to community vote. Similarly, an interaction present in the abstract or in the full article, it can be entered in the last row.

3. RESULTS AND DISCUSSION

Although the CBioC system has moved well beyond its prototype stage, it is still considered a "beta" system and new features are being added. It is, however, functional. To date, over 4.5 million abstracts have been pre-processed, and CBioC does dynamic ("on the fly") extraction when a user views an abstract that has not been pre-processed. This is an important feature that gives users total control over which abstracts are to be processed. Additionally, we have incorporated interactions from BIND, GRID, MINT, DIP and IntAct.

A total of 261 distinct users have downloaded the CBioC plug-in, with 161 of them becoming registered users since June 2006, when CBioC was mentioned in Science Magazine's NetWatch [12]. Partial statistics for those that have chosen a personal title (such as "Doctor", "Professor", or "Researcher") during the registration process show our users include 53 doctors, 30 researchers, 17 professors, 8 post-docs and 40 students. Actions of registered users are tracked, and have so far yielded a total of over 500 curated interactions (either added, modified, or approved through a "yes" vote). Of course, this added to the more than 1.5 million relationships automatically extracted from text. As a point of comparison, at the time of its publication, IntAct[13] had 2200 interactions, most of them from high throughput experiments (not curated). Two years after its conception, MINT[14] had 2500 curated mammalian interactions, and was the largest publicly available dataset of curated entries at the time. It will be interesting to see how many curated interactions will CBioC have when it hits the 2 year mark in June of 2008. Table 1 shows statistics about content and user actions. About 55% of the votes confirm the automatic extraction

is correct (yes votes), an indicator of the extraction system precision. This use of community validation is another area to explore as value added by the CBioC platform.

Aside from the web interface, data from CBioC has also been used in computational methods with encouraging results[15]. We presented in a computational method to uncover possible gene-disease relationships that are not directly stated in an abstract or were missed by the initial mining of the literature. Ranked lists of genes obtained from the method reach precision of 98% for the top 50, and up to 92% for the top 200 genes, in contrast to about 70% accuracy by simple co-occurrence searches.

Table 1. CBioC statistics. The left table details the type of information stored in the CBioC database, accessible via term searches or by PMID. The right table details the number of actions by registered users (as of March 2007). Actions by non-registered users are not tracked. IntAct interactions are being updated, with over 130,000 becoming available soon. Users (excluding the development team) include 53 doctors, 30 researchers, 17 professors, 8 post-docs and 40 students.

CBioC Statistics			
Abstracts		**Integrated Data**	
Total Processed:	1,618,878	BIND Interactions:	114,685
With Interactions:	47%	GRID Interactions:	58,467
Interactions		MINT Interactions:	51,721
Total Protein/Protein:	972,769	DIP Interactions:	52,070
Total Gene/Disease:	301,547	IntAct Interactions:	6,734
Total Gene/Bio-Process:	251,233		

User Action	Total
Add interaction	163
Modify interaction	133
Rate interaction	71
Search	793
View (article)	3169
Vote (total)	370
Vote (yes)	207

References

1. Baral, C. et al. Collaborative Curation of Data from Bio-medical Texts and Abstracts and Its integration. in *Data Integration in the Life Sciences* 309-312 (Lecture Notes in Computer Science, San Diego, CA, 2005).
2. Gonzalez, G., Uribe, J.C., Tari, L., Brophy, C. & Baral, C. Mining Gene-Disease relationships from Biomedical Literature. in *Pacific Symposium in Biocomputing* (Maui, Hawaii, 2007).
3. Soteriades, E.S. & Falagas, M.E. Comparison of amount of biomedical research originating from the European Union and the United States. *BMJ: British Medical Journal.* **331** 192-194 (2005).
4. Emmott, S. Towards 2020 Science: a Report. in *Towards 2020 Science Workshop* (ed. Cambridge, M.R.) (2006).
5. Lu, Z., Cohen, K.B. & Hunter, L. Finding GeneRIFs via Gene Ontology Annotations. in *Pacific Symposium on Biocomputing* Vol. 11 52-63 (World Scientific Publishing Co. Pte. Ltd., Maui, Hawaii, USA, 2006).
6. Tanabe, L. & Wilbur, W.J. Tagging gene and protein names in biomedical text. *Bioinformatics* **18**, 1124-1132 (2002).
7. Ahmed, S.T., Chidambaram, D., Davulcu, H. & Baral, C. IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text. in *BioLINK: Linking Literature, Information and Knowledge for Biology* (Detroit, Michigan, 2005).
8. Koike, A., Niwa, Y. & Takagi, T. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics* **21**, 1227-1236 (2005).
9. Chun, H.-W. et al. Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning. in *Pacific Symposium on Biocomputing* Vol. 11 4-15 (2006).
10. Herder, C. et al. Chemokines and Incident Coronary Heart Disease. Results From the MONICA/KORA Augsburg Case-Cohort Study, 1984-2002. *Arterioscler Thromb Vasc Biol*, 01.ATV.0000235691.84430.86 (2006).
11. Entrez Gene entry for CCL2 (GeneID: 6347).
12. Leslie, M. NetWatch - Software: Annotate While You Read. in *Science Magazine* Vol. 312 1721 (2006).
13. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucl. Acids Res.* **32**, D452-455 (2004).
14. Arnaud Ceol et al. The (new) MINT Database. in *BITS 2004* (Padova, Italy, 2004).
15. Gonzalez, G., Uribe, J.C., Tari, L., Brophy, C. & Baral, C. Mining Gene-Disease relationships from Biomedical Literature: Incorporating Interactions, Connectivity, Confidence, and Context Measures. in *Pacific Symposium in Biocomputing* (Maui, Hawaii, 2007).