W748–W752 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl194

botXminer: mining biomedical literature with a new web-based application

Uma Mudunuri, Robert Stephens, David Bruining, David Liu and Frank J. Lebeda[1,*]

Science Applications International Corporation-Frederick, Advanced Biomedical Computing Center, SAIC-Frederick Inc., National Cancer Institute-Frederick, Frederick, MD 21702, USA and [1]US Army Medical Research Institute for Infectious Diseases, Integrated Toxicology Division, Fort Detrick, MD 21702-5011, USA

Received February 21, 2006; Revised and Accepted March 21, 2006

ABSTRACT

This paper outlines botXminer, a publicly available application to search XML-formatted MEDLINE® data in a complete, object-relational schema implemented in Oracle® XML DB. An advantage offered by botXminer is that it can generate quantitative results with certain queries that are not feasible through the Entrez-PubMed® interface. After retrieving citations associated with user-supplied search terms, MEDLINE fields (title, abstract, journal, MeSH® and chemical) and terms (MeSH qualifiers and descriptors, keywords, author, gene symbol and chemical), these citations are grouped and displayed as tabulated or graphic results. This work represents an extension of previous research for integrating these citations with relational systems. botXminer has a user-friendly, intuitive interface that can be freely accessed at http://botdb.abcc.ncifcrf.gov.

INTRODUCTION

The National Library of Medicine's (NLM) MEDLINE/PubMed biomedical literature database (http://www.nlm.nih.gov/bsd/licensee/2006_baseline_doc.html) is rapidly expanding and as of the 2006 edition, the baseline database contains more than 15 million citations. A number of papers have introduced tools to perform searches more advanced than those available with the current Entrez-PubMed user interface to mine the abstracts (1–4). To have more control in conducting searches, investigators have also started to reorganize the MEDLINE data into structured relational systems (5) or to exploit different data formats and create new databases (http://www.ebi.ac.uk/~lcwang/medline/index.htm).

These NLM citations are available in several formats (e.g. MEDLINE, ASN.1, and so on). Importing these citation data into a traditional, relational database limits the query capability to SQL (Structured Query Language) techniques. The eXtensible Markup Language (XML) format, however, is emerging as a *de facto* standard for the hierarchical organization of data in a highly granular structure, and provides a much richer query capability beyond what is available through traditional SQL. Oracle XML DB, available in Oracle 9i and 10g, recognizes this new format. It supports both the necessary massive storage requirements and the extensions to traditional SQL that enable a different kind of query which is hierarchical and highly granular. An early study using Oracle XML DB involved the evaluation of data-form standards and database technologies for medical informatics systems (6) (http://dcb.cit.nih.gov/publications/download/cims_performance.pdf). As described by Wang *et al*. (6), the set of utilities associated with XML DB permits data to be represented both as XML elements from XML documents and as cells within relational tables.

A recently developed website that includes a specialized, Oracle-based resource for the clostridial neurotoxins, BotDB (7) (http://botdb.abcc.ncifcrf.gov), has been extended to conduct literature searches of MEDLINE/PubMed using botXminer. The application of XML DB for botXminer was based, in large measure, on the documentation and thorough analysis of an original implementation using XML-formatted MEDLINE citations (http://www.ebi.ac.uk/~lcwang/medline/index.htm). An advantage offered by botXminer over PubMed is that it can make certain queries possible that cannot be performed using the Entrez-PubMed interface. The practical example herein illustrates how citations, containing user-supplied search terms, selected MEDLINE search fields and terms, are grouped and displayed as hyperlinked tabulated or graphic results. A more detailed technical description of the architecture associated with botXminer and other applications will appear elsewhere.

METHODS

Of the more than 15 million MEDLINE XML records created through the 2005 production year that were loaded into Oracle XML DB, a limited reference subset was generated

*To whom correspondence should be addressed: Tel: +1 301 619 4279; Fax: +1 301 619 2348; Email: frank.lebeda@amedd.army.mil

for botXminer using the search words 'botulinum' and 'tetanus'. This subset has 26 563 citations, a value that represents <1% of the total number that are available. The 'Literature' section of BotDB contains two botXminer-related options. The 'Search' option (data not shown) allows the user to query the standard fields of a MEDLINE output [author, terms in the title or abstract, PubMed identification number (PMID) and so on] by entering desired words or terms in the appropriate boxes and by selecting an optional range of publication years.

The second option, 'Group Articles', is a unique query type. Initially, the user-provided search words or terms are searched for within each MEDLINE XML file. These results are then grouped by botXminer with terms contained in a select list of MEDLINE fields: the Medical Subject Headings® (MeSH®) qualifier or descriptor names, author, chemical, keyword and gene Symbol. An 'Advanced search' feature is also available that allows the user to search the common MEDLINE fields of journal title, MeSH and chemical in addition to the title and abstract. The option to specify a range of publication years is also applicable to the 'Group Articles' and the 'Advanced search'.

Queries may contain common logical operators for literature searches that are supported by Oracle Text: AND (&), NOT (~), OR (|) and ACCUM (,) (meaning either term is acceptable but both are preferred). Other operators use context grammar: NEAR (;) for proximity searches, and MINUS (−) for a lower preference. Wildcard symbols used include '%' for one or more wildcard characters, and '_' that allows a single wildcard character. A summary list with examples is provided on a help page.

Output graphics can presently be viewed in PNG, HTML or SVG formats and have so far been tested using Microsoft® Internet Explorer (IE) v.6, Netscape® v.8 and the Mozilla Firefox® v.1 browsers. The bitmapped PNG format only provides a static representation. The HTML format offers the advantage of being able to hyperlink the graphic points (rectangles and circles) to the citation summary information. Label information is presented as a pop-up label when moving the mouse over the rectangles and circles. For the SVG option, an interactive graphic view of the results is dynamically produced using the aiSee graphics package (http://www.aisee.com). The SVG format also has interactive hyperlinks while its associated plug-in (available for IE and Netscape®) provides the user with a drop-down feature menu to zoom in and zoom out, pan across the image, save the image, and other graph features. In the SVG format under IE, positioning the mouse over the rectangles and circles produces the label information within the 'window status bar'. At present, the Firefox browser does support the SVG format by using the aiSee software but does not support the drop-down feature menu. For these reasons, it is recommended that, at this time, the botXminer graphics be viewed in the SVG format using either the IE or the Netscape browsers.

USAGE

As a practical example, a search was conducted to find potential interactions between the SNARE proteins (three of which are substrates for the proteolytic botulinum neurotoxins) and other proteins. The substrates of interest [syntaxin-1A, synaptobrevin (VAMP) and SNAP-25] are intimately involved in the evoked release of neurotransmitter from synaptic vesicles. The retrieved citations for SNARE protein–protein interactions may, thus, help to predict what other cellular functions might be affected in addition to the toxin-induced blockade of neurotransmission.

A 'Group Articles' 'Advanced search' was used with the query: near ((snare, interact%), 5) (Figure 1, back panel). This search was designed to locate the search words 'snare' and 'interact%' (or interacts, interacting, etc.) that are separated by no more than five words. The 'abstract' field and the Group term 'chemical' were also used in this search.

This query resulted in a tabulated listing of 83 Group (chemical) terms that appeared in 18 MEDLINE XML files (Figure 1, middle panel). Clicking on the highlighted value of '18' produces a list of all the retrieved citations (Figure 1, front panel). Values for the 'Number of Articles' for a given term are hyperlinked to a second table that lists citations associated with the selected term (data not shown). The PMID numbers are hyperlinked to their corresponding MEDLINE pages, which are presented in a modified format, while article titles in this table are hyperlinked to their PubMed abstracts (data not shown).

Another way to visualize these data is in a network graph format (Figure 2, upper panel) with a method analogous to that used in ChiliBot (2) and PubNet (3). Before a graph is plotted, numerical characteristics of the graph are displayed. In this example, there are 83 Group terms (chemical names) and 859 connections (circles) between these terms. All pairs of chemical terms are connected by color-coded lines and circles that show the relative 'Number of Articles' for two co-occurring chemical terms. Each rectangle is hyperlinked to a list of citations containing the Group (chemical) term, whereas each circle is hyperlinked to a list of citations that contain the pair of terms associated with that connection (data not shown).

These graphical views also allow the user to quickly focus on and recognize patterns for the most frequent co-occurring terms. In this example, the cluster of terms near the top of the graph is examined more closely with the 'Zoom In' option (accessed through the drop-down menu with the right-hand mouse button) that is used with SVG (Figure 2, lower panel). In this panel, the pairs are only associated with a single paper (gray lines and circles). Nevertheless, from these citations and graphs it is evident that several protein–protein interactions between the SNAREs and other proteins including calmodulin, neurotransporters and other synaptic proteins were readily identified using this approach.

DISCUSSION

One goal in the development of botXminer was to mine the maximum amount of information from the MEDLINE XML files without doing any natural language processing. This development began with L. C. Wang's original documentation (http://www.ebi.ac.uk/~lcwang/medline/index.htm) that described how the whole collection of MEDLINE data can be stored inside Oracle XML DB in object-relational tables. Since

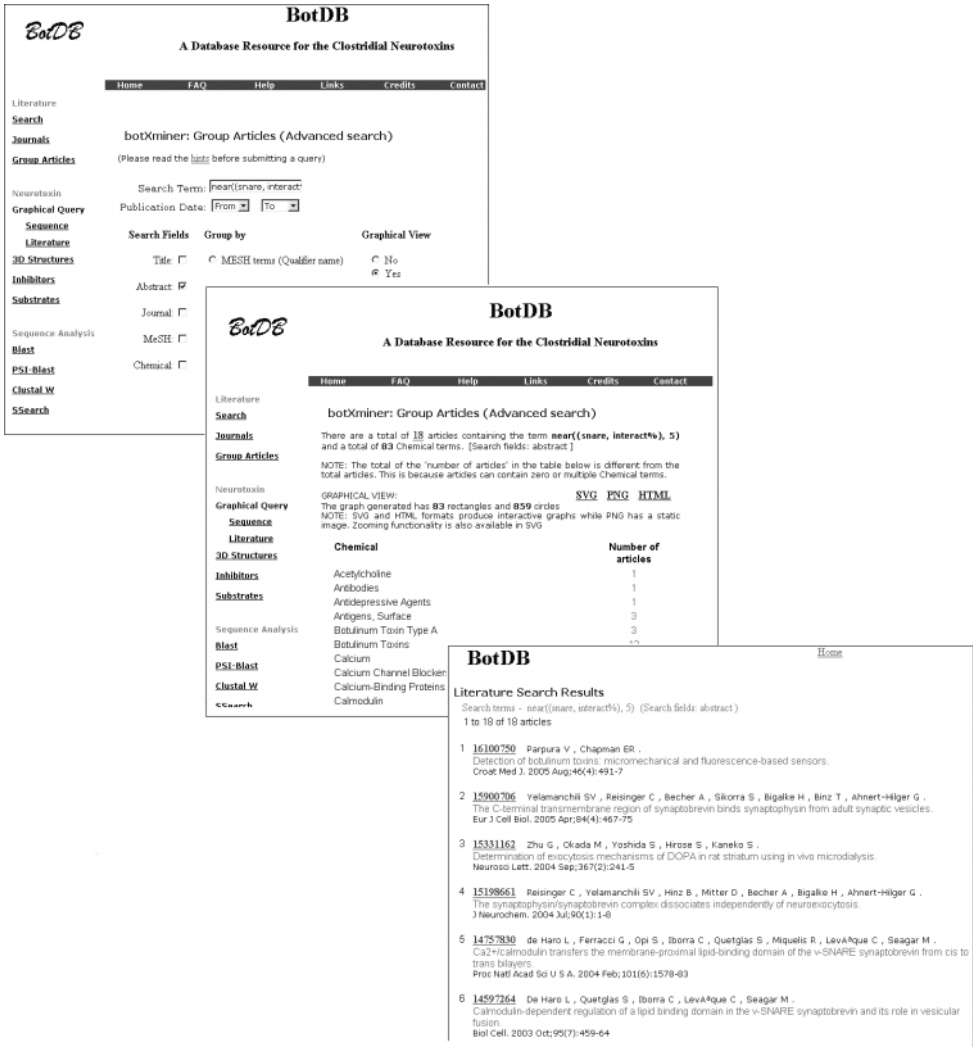

Figure 1. Literature search conducted in BotDB using botXminer and the Group Articles (Advanced search option). Back panel: screen shot of this BotDB search page. The end-user types in a desired word or term (words separated by logical or other operators) and selects one or more of the MEDLINE search fields and one of the six presently available terms. Middle panel: tabulated results are shown for the query 'near ((snare, interact%), 5)', the 'abstract' field and the 'chemical' term. The resulting list of chemical terms are tallied in the form of an alphabetically sorted table that can be expanded by clicking on a number. Front panel: the value of '18' associated with the total number of articles (middle panel) is hyperlinked to these citations.

entire MEDLINE XML files are stored in this database, the structure of these files is maintained, in contrast to writing new code to parse the data and to subsequently store them into relational tables. The schema-based, structured type data storage strategy is made more efficient in botXminer with Oracle's text indexing capability which makes proximity searches with operators such as 'NEAR', a context grammar operator, and wildcard searches feasible. The Group Articles feature of botXminer allows the user to look at a single, tabulated list of relevant grouped terms and to quickly retrieve citations of interest. Finally, botXminer has the first user interface for MEDLINE XML files that are stored in Oracle XML DB.

In comparison to other publicly available PubMed search applications, the differences exhibited by botXminer provide an opportunity for its descendants to enhance searches of the continually growing biomedical literature. In the tools described by Oliver *et al.* (5), MEDLINE XML files that have been downloaded are subsequently parsed and stored in schema-defined relational tables. Devising appropriate schema to optimally perform all queries is a challenging task. In contrast, PubFinder (4) stores abstracts in its database and creates a reference dictionary of commonly used words. The user needs to provide PMIDs of abstracts of interest which are processed for word frequencies and compared to the reference dictionary. This natural language application seems to be comparable at some level to PubMed's 'related articles' search feature. Applications, such as ChiliBot (1), PubGene (8) and MedMiner (9), require, as a minimal user-supplied input, a recognized gene name to conduct its PubMed search. GO-PubMed (2) uses the Gene Ontology vocabulary in its searches. Textpresso (10), as part of its knowledge retrieval strategy, uses a text-to-XML converter to systematically mark up sentences by its specialized ontology (that is presently based on *Caenorhabditis elegans*) from full-length, searchable articles.

The example query provided here illustrates the power of botXminer to help the user to efficiently and selectively search

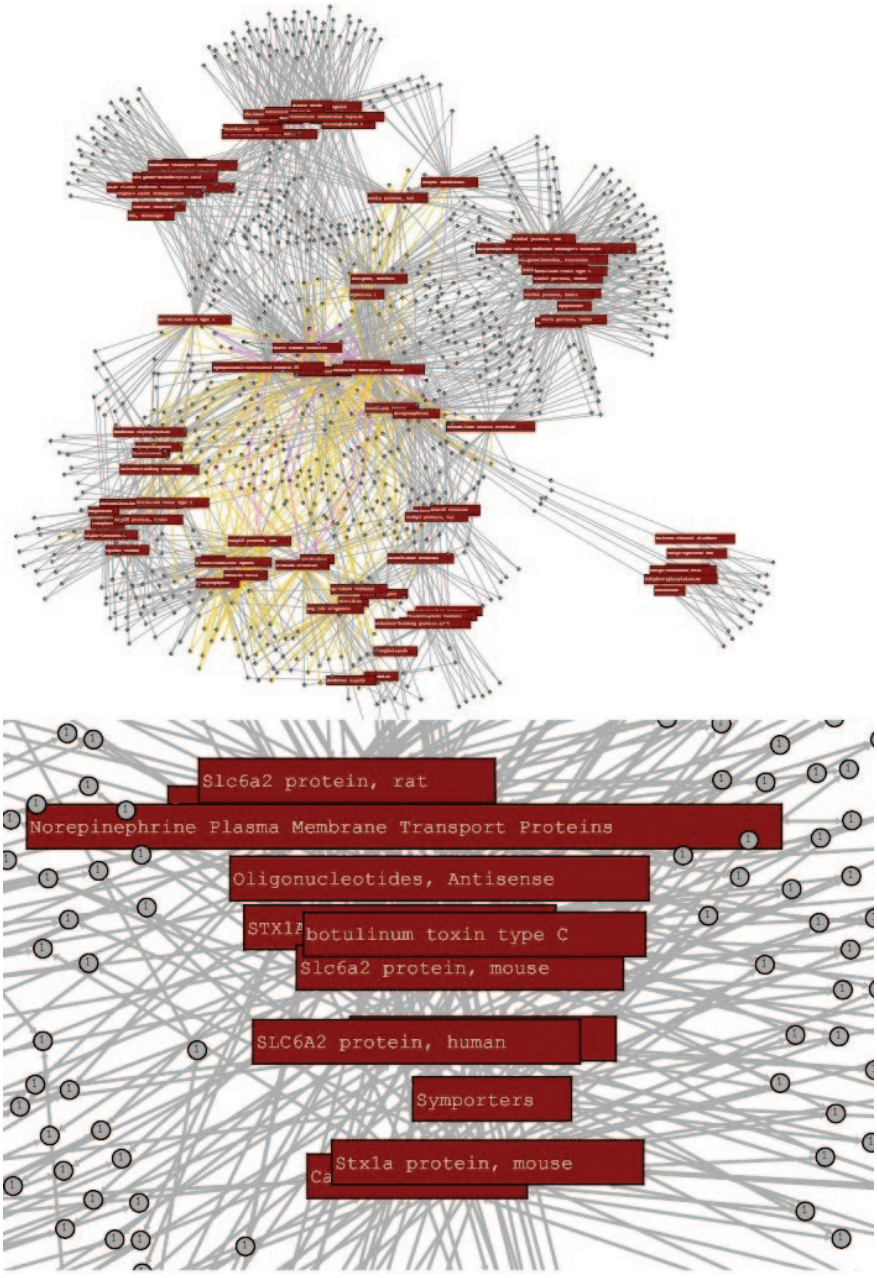

Figure 2. Network graphs show relationships between pairs of chemical terms. Clicking on the 'Graphical View' link (Figure 1, middle panel) produces a dynamic network graph (upper panel) that shows the interconnectedness of the chemical terms which, in turn, are linked to references (color-coded lines) with the number of co-occurrences labeled in the small circles. This graph displays all the terms from the table (Figure 1, middle panel) that are connected by the number of articles in which they co-occur. The lines and circles are color coded depending on the number of articles: gray, the terms co-occur in 1 article; yellow, co-occurrence in 2–5 articles; pink, co-occurrence in 6–10 articles; green, co-occurrence in >10 articles. Lower panel: magnified region of the cluster of terms that include norepinephrine transporters.

for co-occurrences of grouped terms from MEDLINE XML files and to focus on the general problem of protein–protein interactions. Since we have created a database from the entire set of MEDLINE XML files, it is anticipated that a variety of other, specialized citation subsets (e.g. cytoskeletal proteins, signal transduction pathways and diseases such as cancer) will be developed that are similar in design to botXminer. It is further anticipated that botXminer will serve as a template for future applications that will mine the entire MEDLINE and other very large sources of biomedical information.

ACKNOWLEDGEMENTS

We thank Lichun C. Wang for kindly advising us during the initial stages of this effort and for her encouragement. Support for this project was provided to F.J.L. by the Defense Threat Reduction Agency (DTRA; D_X009_04_RD_B) and the Defense Advanced Research Project Agency (DARPA; 05-0-DA-008). This project has been also been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. Opinions, interpretations, conclusions, and recommendations

are those of the authors and are not necessarily endorsed by the U.S. Army or the Department of Health and Human Services. The mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government. Funding to pay the Open Access publication charges for this article was provided by DTRA.

Conflict of interest statement. None declared.

REFERENCES

1. Chen,H. and Sharp,B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, **5**, 147–159.
2. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*., **33**, W783–W786.
3. Douglas,S.M., Montelione,G.T. and Gerstein,M. (2005) PubNet: a flexible system for visualizing literature derived networks. *Genome Biol*., **6**, R80–R89.
4. Goetz,T. and von der Lieth,C.W. (2005) PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Res*., **33**, W774–W778.
5. Oliver,D.E., Bhalotia,G., Schwartz,A.S., Altman,R.B. and Hearst,M.A. (2004) Tools for loading MEDLINE into a local relational database. *BMC Bioinformatics*, **5**, 146–157.
6. Wang,S.A., Fann,Y., Cheung,H., Pecjak,F., Upender,B., Fazin,A., Lingam,R., Chintala,S., Wang,G., Kellogg,M., Martino,R.L. and Johnson,C.A. (2004) Performance of using Oracle XMLDB in the evaluation of CDISC ODM for a clinical study informatics system. In *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04)*, 24-25 June. Bethesda, MD, p. 594.
7. Lebeda,F.J. (2004) BotDB: a database resource for the clostridial neurotoxins. *Mov. Disord*., **19** ((Suppl. 8)), S35–S41.
8. Jenssen,T.-K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*., **28**, 21–28.
9. Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L. and Weinstein,J.N. (1999) MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, **27**, 1210–1217.
10. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*., **2**, e309.