
Mining the botulinum research literature

Regina Z. Cer, Uma Mudunuri
and Robert M. Stephens

Advanced Biomedical Computing Center,
Information Systems Program,
SAIC-Frederick Inc., NCI-Frederick,
Frederick, MD 21702, USA
E-mail: cerr@mail.nih.gov
E-mail: mudunuriu@mail.nih.gov
E-mail: stephensr@mail.nih.gov

Frank J. Lebeda*

Research Area Directorate 2,
US Army Medical Research and Materiel Command,
504 Scott Street, Ft. Detrick, MD 20702-5012, USA
E-mail: Frank.Lebeda@us.army.mil

*Corresponding author

Abstract: The literature search section (botXminer) in the botulinum resource database (BotDB; <http://botdb.abcc.ncifcrf.gov>) contains a new query tool – ‘Batch Search’. In basic mode, a user can submit a list of words or phrases that are used to conduct a sequential search of the titles, abstracts and MEDLINE contents of clostridial neurotoxin citations. A table of results is returned that contains the numbers of citations found for each query term that are, in turn, linked to information from PubMed and MEDLINE for each citation. In matrix mode, a user can join two lists of user-defined terms in which all of the results from the combinatorial pairs are returned in a matrix. Each matrix element is linked to a citation list. In both modes, searches can be refined by selecting a range of publication years. Examples of ‘Batch Search’ matrices are statistically analysed for their quality by calculating their recall and precision values.

Keywords: batch searches; batch queries; information retrieval; recall; precision; F_1 -measure; biomedical literature; botulinum; combinatorial query; controlled vocabulary; database.

Reference to this paper should be made as follows: Cer, R.Z., Mudunuri, U., Stephens, R.M. and Lebeda, F.J. (2010) ‘Mining the botulinum research literature’, *The Botulinum J.*, Vol. 1, No. 3, pp.270–280.

Biographical notes: Regina Z. Cer is a Bioinformatics Analyst at the Advanced Biomedical Computing Center, SAIC-Frederick. She received her Master’s Degree in Computer Science with a bioinformatics concentration from the Johns Hopkins University in Baltimore, MD. Her current research interests include integrated biology, therapeutics, and graphics.

Uma Mudunuri is an IT Manager in the Bioinformatics Support Group at the Advanced Biomedical Computing Center, SAIC-Frederick. Her current research interests include literature and text mining, semantic web, data integration, visualising networks. She received her Master's Degree in Biotechnology from the University of Pune, India and a second Master's Degree in Bioinformatics from the University of Pennsylvania in Philadelphia, PA.

Robert M. Stephens received his PhD from the Department of Immunology and Infectious Diseases, Johns Hopkins University in Baltimore, MD. He is the Director of the Bioinformatics Support Group at the Advanced Biomedical Computing Center, SAIC-Frederick. He currently oversees several Bioinformatics projects covering NextGen sequencing, LIMS, microarray analysis, genomic annotations, data integration and has more than 100 publications in these areas.

Frank J. Lebeda is Deputy Director of the Combat Casualty Research Program and Lectures in Protein Bioinformatics at the Johns Hopkins University. He received his Doctorate in Pharmacology from the State University of New York at Buffalo. He has published in the areas of neuropharmacology and computational molecular biology, is a co-editor of *Biological Weapons Defense* (2005), has served on numerous government review panels, and is on the editorial board of *The Botulinum Journal*.

1 Introduction

The online botulinum resource, BotDB (Lebeda, 2004), is designed to provide specific structural and functional information on the clostridial neurotoxins using data gleaned from a variety of sources. Information within this resource is internally and externally linked to the biomedical literature and to relevant databases. Thus far, the BotDB provides structural information for 90 neurotoxins and other proteins from Protein Data Bank (PDB) files and associated websites that contain sequence and three-dimensional data. This resource offers pre-calculated analyses of secondary and tertiary structural data. The 'bontDBI' subsection contains citations and details about the inhibitory properties of low molecular weight compounds. Kinetic data from our modelling studies for the onset of neurotoxin effects (Lebeda et al., 2008) are also included. An IC_{50} -to- K_i converter is available (Cer et al., 2009) in addition to a variety of sequence analysis tools.

These BotDB resources and tools are linked to relevant literature citations. We have conducted advanced literature searches using botXminer (Mudunuri et al., 2006), a suite of tools linked to a specialised database of a MEDLINE/PubMed[®]-derived subset of clostridial neurotoxin citations. While collecting additional kinetic data for modelling the mechanism of action of botulinum neurotoxin from the clinical literature, it became apparent that more text mining capabilities were required than those offered by the Entrez/PubMed[®] and other web-based search tools, e.g., Chilibot (Chen and Sharp, 2004), GoPubMed (Doms and Schroeder, 2005), MedMiner (Tanabe et al., 1999), PubFinder (Goetz and von der Lieth, 2005), PubNet (Douglas et al., 2005), SciMiner (Hur et al., 2009), PubMatrix (Becker et al., 2003).

In this paper we introduce 'Batch Search', a new batch query tool that was developed and added to the 'Search' and 'Group Articles' options that were originally

created for botXminer. The basic mode of batch queries consists of executing a series of consecutive searches that are based on lists of user-supplied query terms contained in MEDLINE fields (title, abstract, author, journal, and chemical) and MeSH[®] terms (qualifiers and descriptors). We enhanced these features by adding a matrix mode for finding citations that contain co-occurrences of search terms within a single citation. This paper describes the two modes of operation (basic and matrix) of our 'Batch Search' tool that is located at <http://botdb.abcc.ncifcrf.gov/botXminer/batchSearch.jsp>. Examples are provided of its usage, and the results of statistical analyses that quantitatively assessed their relevancy.

2 Methods

2.1 Implementation

The 'Batch Search' tool is implemented within the BotDB (<http://botdb.abcc.ncifcrf.gov>) as a web application using an Oracle[®] database (Oracle9i Enterprise Edition Release 9.2.0.4.0), Java (JDK 1.5.0), Apache web server and Tomcat 4.1. Contained in botXminer are the 31,568 citations (as of September, 2009) in which the keywords 'botulinum' and 'tetanus' were used to identify and archive potentially relevant publications from the more than 19 million citations residing in MEDLINE/PubMed. Unlike some other PubMed search engines, there are no time restrictions for query submissions. The flow chart in Figure 1 shows the general search strategies of botXminer. 'Batch Search' was designed to extend the capabilities of the 'Search' and 'Group Articles' options in botXminer by conducting searches using lists of query words or phrases contained in titles, abstracts, and other features in MEDLINE and eventually, when accessible, the full text of the citation. This mining tool has an easy, intuitive web interface, does not require any registration and is open access. Links to Help sections include *Background*, *FAQs*, *Literature*, *Query Help*, *Links*, *Examples*, and *Sample Files* (Figure 2).

Figure 1 Schematic outline of botXminer. The three general types botXminer options ('Search', 'Batch Search' and 'Group Articles') are shown in relation to the Oracle XML database and the outputs containing PubMed and MEDLINE information (see online version for colours)

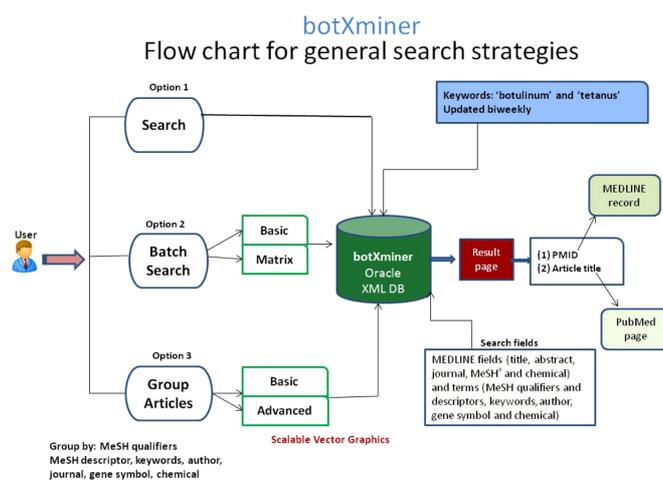


Figure 2 View of the ‘Batch Search’ basic mode page. The query box allows the user to type or paste in search terms. The user can narrow down searches further by either entering specific publication years or selecting a range of years. On the right-hand side is a list of links to the help sections (see online version for colours)

botXminer: [Batch Search \(basic mode\)](#)

- (1) Enter one or more search terms in the text box.
- (2) Optionally, narrow your search by either typing the years in the text box or by selecting the publication years from the provided range.
- (3) Click ‘Submit’.

Search term(s) (required)
Enter one search term per line.

food
iatrogenic
infant
wound

Publication Year(s) (optional)
Enter one year per line. [Download year list](#)

OR
Select from the provided range.

2007 2008

Submit Reset

[Batch Search \(matrix mode\)](#)

Background
FAQs
Literature
Query Help
Links
Examples
Sample Files

2.2 ‘Batch Search’ basic mode

The basic ‘Batch Search’ mode of botXminer allows the end-user to type or paste into the query box one or more terms to search the titles, abstract and MEDLINE contents of the subset of clostridial neurotoxin citations. Optionally, users may narrow their search further by either entering specific publication years (not shown) or selecting a range of publication years from the drop-down menus. Output lists contain the number of retrieved citations, authors of the publication and the journals in which the retrieved citations were published. Each citation has links to the corresponding MEDLINE and PubMed pages (Figure 3).

2.3 ‘Batch Search’ matrix mode

The matrix mode allows two sets of lists to be entered into two different query boxes (Figure 4). Using the ‘AND’ Boolean logical operator, concurrency will be sought for each pair of terms from both query lists within a single citation. Alternatively, the ‘OR’ Boolean logical operator can be used to find one or both terms in a single citation’s title, abstract or MEDLINE information. The query lists that are enumerated below were developed to demonstrate and evaluate the capacity of the ‘Batch Search’ matrix option.

Figure 3 Detailed outputs from 'Batch Search' basic mode. The summary table displays the categories for the number of citations (grey), authors (yellow) and journals (green) that were retrieved. Clicking on the query term 'food' with the number 101 from the matrix returned a list of citations. For each citation the paper title serves as a hyperlink to the PubMed page and the PubMed ID is a hyperlink to the MEDLINE page (see online version for colours)

BotDB [Home](#)

botXminer: Batch Search (basic mode) Results

Total no. of search terms submitted: 4
Publication date: 2007 - 2008

Search Term	Citations	Authors	Journals
food	101	375	69
iatrogenic	7	23	7
infant	203	941	124
wound	48	202	45
Total	359	1541	245

[Perform another Batch Search \(basic mode\)](#)

botXminer: Batch Search Results

Search terms - food
Publication Date: 2007 - 2008

1 to 101 of 101 articles

- 1 [19454118](#) Nelson R.
[Anal fissure \(chronic\)](#)
Clin Evid (Online). 2007 ;2007():
- 2 [19454107](#) Nicholas R, Chataway J.
[Multiple sclerosis](#)
Clin Evid (Online). 2007 ;2007():
- [19454072](#) Ferreira JJ, Sampaio C.
[Essential tremor](#)
Clin Evid (Online). 2007 ;2007():

BotDB [Home](#)

Article

997

PMID: 19454118
Owner: NLM
Status: In-Data-Review
DateCreated: 2009-05-20
ISSN: 1752-8328
Journal Volume: 2007
PubDate: 2007
Title: Anal fissure
ISO Abbreviation: Clin Evid (Online)
Article Title: Anal fissure (chronic)
Abstract: INTRODUCTION: Anal fissures are a common cause of anal pain during, and for 1-2 hours after, defecation. The cause is not fully understood, but low intake of dietary fibre may be a risk factor. METHODS AND OUTCOMES: We conducted a systematic review and aimed to answer the following clinical questions: What are the effects of non-surgical treatments for chronic anal fissure? What are the effects of surgical treatments for chronic anal fissure? We searched Medline, Embase, The Cochrane Library and other important databases up to January 2007. (Clinical Evidence reviews are updated periodically; please check our website for the most up-to-date version of this review). We included harms alerts from relevant organisations such as the US Food and Drug Administration (FDA) and the UK Medicines and Healthcare products Regulatory Agency (MHRA). RESULTS: We found 11 systematic reviews, RCTs, or observational studies that met our inclusion criteria. We performed a GRADE evaluation of the quality of evidence for interventions. CONCLUSIONS: In this systematic review we present information relating to the effectiveness and safety of the following interventions: anal advancement flap, and stretch dilation, botulinum A toxin-haemagglutinin complex alone or with nitrate, calcium channel blockers, internal anal sphincterotomy, and latex rubber bands.
Affiliation: Northern General Hospital, Sheffield, UK.
Author: Nelson Rick
Language: eng
Publication Type: Journal Article
ArticleID: 2007/08/15
Country: England

Figure 4 ‘Batch Search’ matrix mode webpage showing Query_1. See Methods for a description of Query_1 (see online version for colours)

botXminer: Batch Search (matrix mode)

- (1) Enter one or more search terms in the first text box.
- (2) Choose a ‘Connect term’: ‘AND’ or ‘OR’ from the drop down menu.
- (3) Enter one or more search terms in the second text box.
- (4) Optionally, narrow your search by selecting the publication years from the provided range.
- (5) Click ‘Submit’.

Note The terms in text box one and those in text box two will be paired with your selected connect term. The resulting combined term will be used to search botXminer database.

2.4 Example queries

All botXminer queries using the ‘Batch Search’ matrix mode option were done for citations that were published in 2007 or 2008. Using logical operator notation for two lists (a_1, a_2, \dots, a_m) and (b_1, b_2, \dots, b_n) , the combinatorial query for a this option can be represented in a general form:

$$(a_1 \cup a_2 \cup \dots \cup a_m) \cap (b_1 \cup b_2 \cup \dots \cup b_n).$$

Each of the three test queries contained a short list of the types of botulism:

(food, iatrogenic, infant, wound)

and a short list of epidemiology-related terms:

(cases, incidence, frequency, occurrence)

These lists were designed for the example queries in this work and were not intended to be exhaustive.

Query_1 contained these two lists in which the equivalent, explicit query is

(food OR iatrogenic OR infant OR wound) AND (cases OR incidence OR frequency OR occurrence).

Query_2 contained these two lists, and only accepted those abstracts, titles and MEDLINE entries that did not contain the word 'tetanus', i.e., NOT tetanus or ~tetanus. The equivalent query is:

(food ~tetanus OR iatrogenic ~tetanus OR infant ~tetanus OR wound ~tetanus) AND (cases ~tetanus OR frequency ~tetanus OR incidence ~tetanus OR occurrence ~tetanus).

Query_3 contained both lists, NOT tetanus and also searched for the word 'botulism', i.e., AND botulism. The equivalent query is:

(food ~tetanus AND botulism OR iatrogenic ~tetanus AND botulism OR infant ~tetanus AND botulism OR wound ~tetanus AND botulism) AND (cases ~tetanus AND botulism OR frequency ~tetanus AND botulism OR incidence ~tetanus AND botulism OR occurrence ~tetanus AND botulism).

2.5 Statistical analyses

The numbers of citations that were returned by the query searches were categorised (Flynn, 1986) according to terminology used for information retrieval:

R_R Number of relevant, retrieved citations or true hits. The total number of relevant, retrieved citations is arbitrarily defined here by the results from Query_1.

R_N Number of relevant, not retrieved citations is from Query_2 or Query_3.

N_R Number of non-relevant, retrieved citations is from all three queries.

Commonly used statistics such as precision, recall and the F_1 -measure were calculated (Hur et al., 2009). Precision is the ratio of true hits (i.e., relevant, retrieved citations) to the number of citations retrieved:

$$p = \frac{R_R}{(R_R + N_R)}. \quad (1)$$

Recall is the ratio of true hits to the total number of relevant citations in the system:

$$r = \frac{R_R}{(R_R + R_N)}. \quad (2)$$

In this exercise, recall for Query_1 is equal to 1.00.

For a single measure of performance, an F -measure was calculated that equally weighed precision and recall

$$F_1 = \frac{2pr}{(p+r)}. \quad (3).$$

2.6 Potential sources of error

The citations within botXminer represent over 30,000 publications in more than 3800 of the ~5000 journals available in MEDLINE/PubMed. Because not all biomedical journals are indexed in that repository, it is likely that our approach will miss some relevant papers.

3 Results

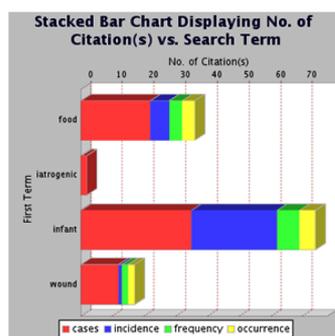
The output from Query_1 is displayed in a colour-coded matrix format and in a histogram representation (Figure 5). The numbers in the matrix table with a grey background represent the number of citations found for those two particular search term combinations, one from each query list. For easier visualisation, only the numbers with values greater than zero are hyperlinked to reduce the analysis time for a user. The numbers with a white background are the total raw number of citations per row (TRRC) or total raw number of citations per column (TCRC). The numbers with the green background are the total number of unique citations found for the vertically displayed query terms (TRUC). At the bottom of the matrix are numbers with a yellow background that represent the total number of unique citations for each term that is horizontally displayed at the top of the matrix (TCUC). Finally, the single number with a gold background contains the total number of unique citations for the query (TTUC). As in the basic mode, these values are linked to botXminer simple search results which contain links to the respective PubMed/MEDLINE pages.

Figure 5 Summary of the retrieved results from Query_1 from the 'Batch Search' matrix mode. The upper panel shows the numeric output of the matrix. The corresponding histograms in the lower panel are colour coded with respect to the search terms on the horizontal portion of the above matrix (see online version for colours)

botXminer: Batch Search (matrix mode) Results
 You submitted 4 first term(s) and 4 second term(s).
 They were joined with the selected connect term 'AND'.
 Publication date: 2007 - 2008

AND	cases	incidence	frequency	occurrence	TRRC ¹	TRUC ²
food	22	6	4	4	36	29
iatrogenic	2	0	0	0	2	2
infant	35	27	7	5	74	59
wound	12	1	2	2	17	15
TCRC ³	71	34	13	11		
TCUC ⁴	64	34	13	10	TTUC ⁵ = 98	

¹ TRRC = Total Row Raw Citations (may contain duplicates)
² TRUC = Total Row Unique Citations
³ TCRC = Total Column Raw Citations (may contain duplicates)
⁴ TCUC = Total Column Unique Citations
⁵ TTUC = Total True Unique Citations



It should be emphasised that while there is no limit to the number of search terms submitted, a user should not submit more than 200 × 10 search pairs. The dynamic histogram graph generated as in Figure 5 may be of poor resolution for such large matrices and may cause memory problems for the server.

3.1 Analysis

Retrieved citations were individually evaluated as being relevant or non-relevant according to the information content in the titles, abstracts or associated MEDLINE documentation (Table 1). Full electronic texts were obtained to evaluate those retrieved citations that contained ambiguous information.

Table 1 Query results and performance measures for three 'Batch Search' matrix mode searches for botXminer citations published in 2007 or 2008

Query no.	Query content	Retrieved citations	Relevant,	Relevant,	Non-relevant,	Recall (R_R)/ (R_R + R_N)	Precision (R_R)/ (R_R + N_R)	F_1
			retrieved citations (R_R)	non-retrieved citations (R_N)	retrieved citations (N_R)			
1	Botulism list AND epidemiology list	98	21	0	77	1.00	0.21	0.35
2	Botulism list AND epidemiology list ~tetanus	44	21	0	23	1.00	0.48	0.49
3	Botulism list AND epidemiology list ~tetanus AND botulism	19	18	3	1	0.86	0.95	0.93

The initial set of test terms of Query_1 yielded 98 hits that had low precision (0.21) and F_1 scores (0.35). By definition, the recall value was 1.00 for this first query. The submission of Query_1 resulted in 98 returned citations, 21 of which were relevant and 77 of which were non-relevant. The value of 21 was used as the total number of relevant citations for this exercise.

The results returned from Query_2 yielded 44 hits. The same number of relevant citations was retrieved as with Query_1 and therefore also had a recall value of 1.00. But the number of non-relevant citations was dramatically reduced from 77 to 23 citations which led to increased values of precision and F_1 scores, 0.48 and 0.49, respectively.

Further improvements of the precision (0.95) and the F_1 scores (0.93) were obtained with Query_3. This more focused search yielded 19 hits. On the other hand, the recall value was decreased to 0.86 because three of the relevant citations were missed.

4 Discussion

The 'Batch Search' tool described in this paper is a new extension to botXminer (Mudunuri et al., 2006), a PubMed mining application that is part of the BotDB clostridial neurotoxin resource. It was designed to be fast, efficient and reliable. The main advantage of 'Batch Search' mode over using the Entrez search engine provided by the National Library of Medicine is in the ease of operation because a user can submit multiple queries in one keystroke. Reducing the amount of typing is expected to lead to

fewer errors and to reduce the time and frustration involved in creating and submitting complex queries. As noted by others (Becker et al., 2003), submitting a series of uninterrupted queries diminishes the opportunity for the user to become distracted by the initial set of results. Another advantage for a botulinum or tetanus researcher to use 'Batch Search' is that the run time will be shorter than other PubMed applications because botXminer has pre-selected citations that contain the terms 'botulinum' and 'tetanus'. This filtering process avoids searching through more than 19 million biomedical citations in MEDLINE/PubMed. Instead, a user will be searching a much smaller subset. To filter the scope of the 'Batch Search' even further, specific dates or a range of publication dates can be used. A concerted effort was also made to provide a user friendly visualisation of the results. The intuitive matrix table and the colour-coded dynamic histogram graphical outputs were designed to help accelerate the search process.

The exercises performed here demonstrate that queries need to be balanced with respect to the search terms due to trade-offs between precision and recall scores. It is recommended that several search strategies are used to ensure that the maximum number of relevant citations is returned despite a potential increase in the number of returned non-relevant citations. The example queries in this paper retrieved no false positive citations. Therefore, only the statistical measures of precision and recall, were used. In binary classifications, quantitative measures such as selectivity, specificity, and areas under Receiver Operating Characteristic (ROC) curves can be used as evaluation metrics (Metz, 1978). It is expected that with more complex queries, binary classifications could be used to evaluate the returned citations.

The exercises also suggest a potential role for the 'Batch Search' feature in botXminer in developing controlled vocabularies. Indeed, we have created an initial list of terms that would be important to include in searches for epidemiologic data. These vocabularies along with the MeSH qualifier and descriptor names, can then be used together with user query terms to form customised database search strategies.

Acknowledgements

We thank Dr. Charles Millard for his intellectual contributions during the initial phases in the development of BotDB. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Army. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organisations imply endorsement by the US Government. None of the authors have any personal or financial conflicts of interest. Funding for this work was provided by the Defense Threat Reduction Agency, Joint Science and Technology Office-Chemical Biological Defense [Project 3.10043-07-RD-B] and by the National Cancer Institute, National Institutes of Health under contract HHSN261200800001E. Electronic versions for some of the citations used in the analysis for this paper were obtained at the National Library of Medicine, the Uniform Services University of the Health Sciences and the Sheridan Libraries at Johns Hopkins University.

References

- Becker, K.G., Hosack, D.A., Dennis Jr., G., Lempicki, R.A., Bright, T.J., Cheadle, C. and Engel, J. (2003) 'PubMatrix: a tool for multiplex literature mining', *BMC Bioinformatics*, Vol. 4, pp.61–66.
- Cer, R.Z., Mudunuri, U., Stephens, R. and Lebeda, F.J. (2009) 'IC₅₀-to-K_i: a web-based tool for converting IC₅₀ to K_i values for inhibitors of enzyme activity and ligand binding', *Nucleic Acids Res.*, Vol. 37, pp.W441–W445.
- Chen, H. and Sharp, B.M. (2004) 'Content-rich biological network constructed by mining PubMed abstracts', *BMC Bioinformatics*, Vol. 5, pp.147–159.
- Doms, A. and Schroeder, M. (2005) 'GoPubMed: exploring PubMed with the gene ontology', *Nucleic Acids Res.*, Vol. 33, pp.W783–W786.
- Douglas, S.M., Montelione, G.T. and Gerstein, M. (2005) 'PubNet: a flexible system for visualizing literature derived networks', *Genome Biol.*, Vol. 6, pp.R80.81–R80.10.
- Flynn, R. (1986) *An Introduction to Information Science*, Taylor & Francis Group/CRC Press, Boca Raton, FL.
- Goetz, T. and von der Lieth, C.W. (2005) 'PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts', *Nucleic Acids Res.*, Vol. 33, pp.W774–W778.
- Hur, J., Schuyler, A.D., States, D.J. and Feldman, E.L. (2009) 'SciMiner: Web-based literature mining tool for target identification and functional enrichment analysis', *Bioinformatics*, Vol. 25, pp.838–840.
- Lebeda, F.J. (2004) 'BotDB: a database resource for the clostridial neurotoxins', *Mov. Disord.*, Vol. 19, No. 8, Suppl, pp.S35–S41.
- Lebeda, F.J., Adler, M., Erickson, K. and Chushak, Y. (2008) 'Onset dynamics of type A botulinum neurotoxin-induced paralysis', *J. Pharmacokinet. Pharmacodyn.*, Vol. 35, pp.251–267.
- Metz, C.E. (1978) 'Basic principles of ROC analysis', *Semin. Nucl. Med.*, Vol. 8, pp.283–298.
- Mudunuri, U., Stephens, R., Bruining, D., Liu, D. and Lebeda, F.J. (2006) 'botXminer: Mining biomedical literature with a new web-based application', *Nucleic Acids Res.*, Vol. 34, pp.W748–W752.
- 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*, Vol. 27, pp.1210–1214, 1216, 1217.