BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Web Server

Volume 10(11)

PubstractHelper: A Web-based Text-Mining Tool for Marking Sentences in Abstracts from PubMed Using Multiple User-Defined Keywords

Chou-Cheng Chen[1] & Chung-Liang Ho[1,2,3]*

[1]Institute of Basic Medical Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [2]Department of Pathology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [3]Infectious Disease and Signaling Research Center, National Cheng Kung University, Tainan, Taiwan, Department of Medical Laboratory Science and Biotechnology, College of Medicine, National Cheng Kung University, Tainan, Taiwan; Chung-Liang Ho - Email: clh9@mail.ncku.edu.tw; *Corresponding author

Received October 29, 2014; Accepted November 14, 2014; Published November 27, 2014

Abstract:
While a huge amount of information about biological literature can be obtained by searching the PubMed database, reading through all the titles and abstracts resulting from such a search for useful information is inefficient. Text mining makes it possible to increase this efficiency. Some websites use text mining to gather information from the PubMed database; however, they are database-oriented, using pre-defined search keywords while lacking a query interface for user-defined search inputs. We present the PubMed Abstract Reading Helper (PubstractHelper) website which combines text mining and reading assistance for an efficient PubMed search. PubstractHelper can accept a maximum of ten groups of keywords, within each group containing up to ten keywords. The principle behind the text-mining function of PubstractHelper is that keywords contained in the same sentence are likely to be related. PubstractHelper highlights sentences with co-occurring keywords in different colors. The user can download the PMID and the abstracts with color markings to be reviewed later. The PubstractHelper website can help users to identify relevant publications based on the presence of related keywords, which should be a handy tool for their research.

Availability:
http://bio.yungyun.com.tw/ATM/PubstractHelper.aspx and http://holab.med.ncku.edu.tw/ATM/PubstractHelper.aspx.

Keywords: PubstractHelper, text-mining, reading helper

Background:
A large amount of information about biological functions and genetic research can be obtained by searching the PubMed database. While this database provides a list of boldface keywords to help users to search abstracts, it does not highlight the sentences which contain these keywords, and it does not provide any information about these keywords co-occurring in the same sentence. While many previous studies have developed websites and databases to help researchers in this context, these resources have a number of limitations, as outlined below.

EBIMed can be used to find genes and keywords related to proteins, Gene Ontology (GO) annotations, drugs and cancers that co-occur in the same sentences **[1].** The BSQA shows abstracts which contain keywords for genes and behaviors co-occurring in the same sentences **[2].** MeInfoText is a text-mining database that shows abstracts which contain co-occurring

keywords related to specific genes and methylation and/or cancers **[3].** Another tool for conducting database searches is PubTator which features a PubMed-like interface **[4].** This database enables entity-specific semantic searches and provides pre-annotations for computer-assisted biocuration, automatically applying text-mining tools to all articles with respect to genes, diseases, species, chemicals and mutations **[4].**

All of the text-mining systems mentioned above are aimed at achieving a comprehensive value-added database, with little attention paid to either the flexibility of the query interface or the ease of visualizing the search results. Since they are value-added databases which require periodical maintenance, they may not synchronize with the PubMed database. PubstractHelper retrieves abstracts directly from PubMed by means of its e-utility feature, ensuring perfect synchronization. By allowing searches of up to ten different groups of up to ten keywords each, all of which are user-defined, PubstractHelper also provides great flexibility. When at least one keyword from groups two to ten co-occurs in the same sentence with at least one keyword from group one, this sentence is highlighted in different colors for easier visualization.

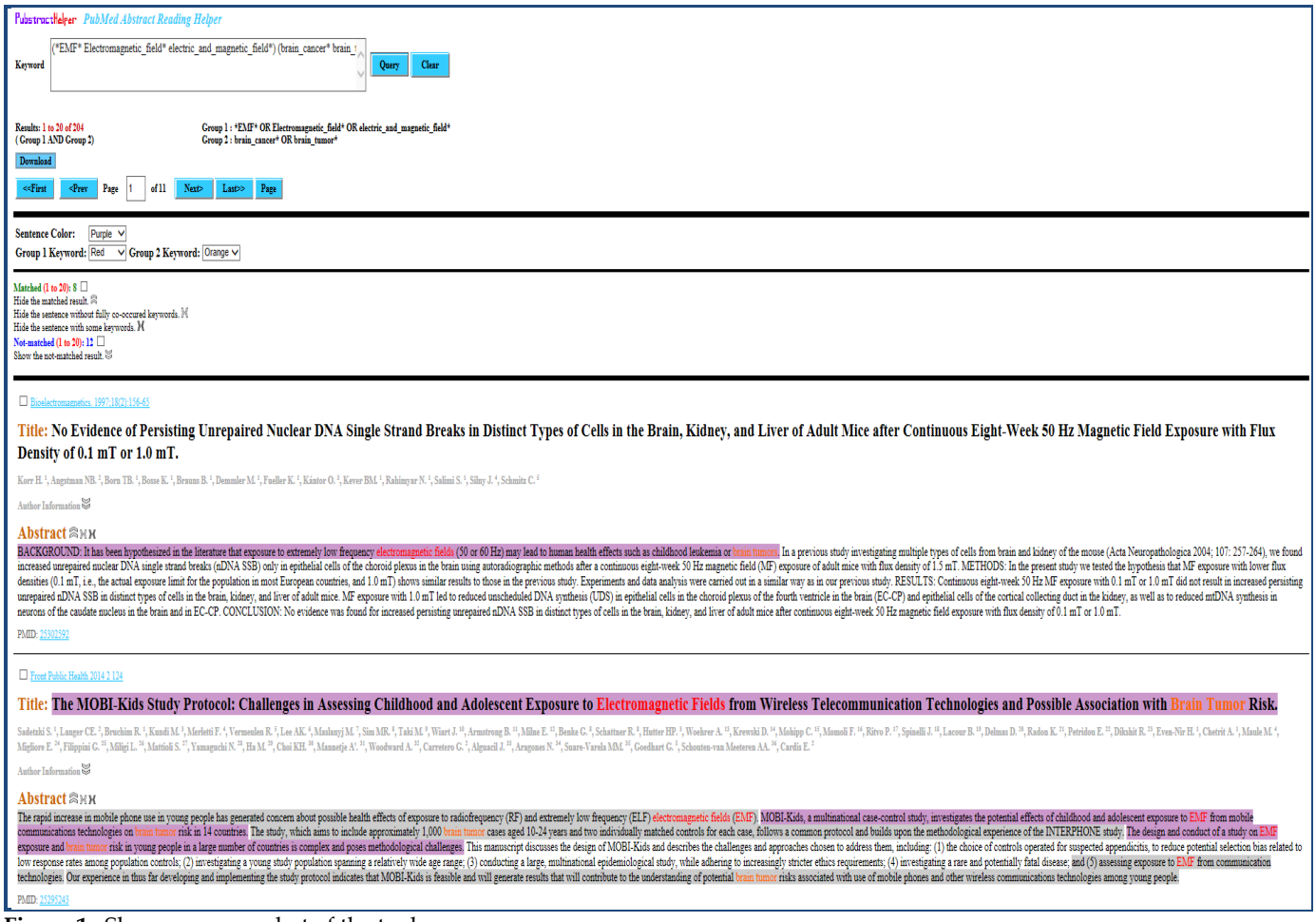

Figure 1: Shows a screen-shot of the tool.

Tool features:

Score calculation

PubstractHelper retrieves abstracts directly from PubMed using its e-utility feature. The abstracts are processed in 3 steps: the insertion of a symbol, the cutting of a sentence, and the scoring of a sentence **[5, 6].** The method for the PubstractHelper website recognizes as a distinct sentence any section of writing that is separated from any other section by ".", "?", "!", or ";" **[7].** The threshold score for each sentence is defined as:

$$Score[o] = P + \prod_i^n K_i$$

where $Score[o]$ is the score for sentence o. The i is an index of the keyword group. The n is the number of keyword groups, with the maximum value set at ten. The value of P is set at one, if at least one user-selected keyword (from any group) occurs in the sentence (default $P = 0$). The value of K_i is set at one, if at least one user-selected keyword in group i occurs in the sentence (default $K_i = 0$). If the value of $Score[o]$ is equal to or greater than 1, HTML tags are added to the beginning and end of the sentence. The tags are different for different scores (either 1 or 2), corresponding to different colors.

Tagged Abstracts

This section presents an example of using PubstractHelper to access abstracts in PubMed. Suppose that a user is interested in the relationship between "electromagnetic field" (EMF) and "Brain cancer". Then, a search is carried out in which group one contains "*EMF*", "Electromagnetic_field*" or "electric_and_magnetic_field*", and group two contains

"brain_cancer*" or "brain_tumor*". The results of this user query are shown in Figure. When this query was actually run with PubstractHelper, the number of abstracts in PubMed which contained a correlation between group one and group two was 204 on October 29th 2014. There are 1888 sentences in all abstracts, but only 82 sentences contain co-occurring keywords from groups one and two.

The user can enter various keywords with spaces between them into the textbox, and compound words can be entered using an underscore (e.g.: "electromagnetic_field"). The symbol "*" can be added in front of or behind a keyword, and the system then recognizes longer words which contain this keyword as a component or section. For example, if a user types "electromagnetic_field*" as the keyword, the system recognizes both "electromagnetic_field" and "electromagnetic_fields". The color of the keywords and the highlighted background of the sentences in which they co-occur can be changed. As shown in Figure, the user only needs to read the highlighted sentences which include the co-occurring keywords in the abstracts. In addition, the user can click to hide sentences in the abstracts which do not include the co-occurring keywords. Finally, the user can download the PMID or abstracts which he selects from the PubstractHelper website.

Conclusion:

Our website, PubstractHelper, is designed to enable researchers to quickly find key sentences in PubMed-listed abstracts by color-marking sentences with co-occurring keywords selected by users. It is a handy tool for biomedical research.

References:

[1] Rebholz-Schuhmann D *et al. Bioinformatics*. 2007 **26**: 237 [PMID: 17237098]

[2] He X *et. al. Nucleic Acids Res*. 2010 **38**: 175 [PMID: 20576702]

[3] Fang YC *et. al. BMC bioinformatics*. 2008 **9:** 22 [PMID: 18194557]

[4] Wei CH *et. al. Nucleic Acids Res*. 2013 **41:** 518 [PMID: 23703206]

[5] Albaraa Abuobieda M Ali *et al. JATIT & LLS*. 2011 **32:** 80

[6] Bird S & Liberman M, *Speech communication*. 2001 **33**: 23 doi:10.1016/S0167-6393(00)00068-6

[7] Matsuo Y *et. al. International Journal on Artificial Intelligence Tools (IJAIT)*. 2003 **13:** 157

Edited by P Kangueane

Citation: **Chen & Liang Ho,** Bioinformation 10(11): 708-710 (2014)