3866–3868 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg538

Update on XplorMed: a web server for exploring scientific literature

Carolina Perez-Iratxeta[1,2,*], Antonio J. Pérez[3], Peer Bork[1,2] and Miguel A. Andrade[1,2]

[1]European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany, [2]Max Delbrück Center for Molecular Medicine, Robert Rösler-Str. 10, 13125 Berlin-Buch, Germany and [3]University of Malaga, Facultad de CC. Campus de Teatinos, 29071 Malaga, Spain

Received February 14, 2003; Revised and Accepted March 21, 2003

ABSTRACT

As scientific literature databases like MEDLINE increase in size, so does the time required to search them. Scientists must frequently inspect long lists of references manually, often just reading the titles. XplorMed is a web tool that aids MEDLINE searching by summarizing the subjects contained in the results, thus allowing users to focus on subjects of interest. Here we describe new features added to XplorMed during the last 2 years (http://www.bork.embl-heidelberg.de/xplormed/).

BACKGROUND AND GOALS

A scientist searching the scientific literature (for example, MEDLINE with Entrez at the NCBI's PubMed server, http://www.ncbi.nlm.nih.gov/entrez/) may initially retrieve an unmanageable number of references, typically hundreds. Even for very specific subjects, it is not always clear how to narrow the search to focus on the most relevant matches. For example, imagine you are a researcher interested in the possible role of the interaction between heparin and proteins in Alzheimer's disease, who queries the PubMed server with the terms 'Alzheimer and heparin'. This search returns presently >100 references to literature, of which only some mention proteins. Finding these currently requires manual examination of the abstracts which can be time-consuming. In such instances XplorMed can be useful (1).

XplorMed is a web tool that summarizes MEDLINE search results according to subjects and allows you to navigate through abstracts in an interactive fashion. Here we give details as to the use of XplorMed. A detailed tutorial is also available online (http://www.bork.embl-heidelberg.de/xplormed/example/).

INPUT TO XplorMed

There are two ways to provide input to XplorMed. You can type a PubMed query directly into our server or you can supply a file containing a set of abstracts. XplorMed can handle several abstract formats: MEDLINE (default), EndNote, XML and XplorMed (see page 1 of the tutorial for details).

A third way to query XplorMed is to start from literature linked to a particular entry from one of the MEDLINE, OMIM (2), SMART (3) or SWISS-PROT/SpTrEMBL (4) databases. Here you simply need to provide the identifier of the entry of your interest and the corresponding database name. The initial set of abstracts of each XplorMed session is kept in the server for a week, enabling you to recover your session. We recommend you start with sets of ~30 references, though the current maximum is 500 abstracts.

OVERVIEW OF AN ANALYSIS

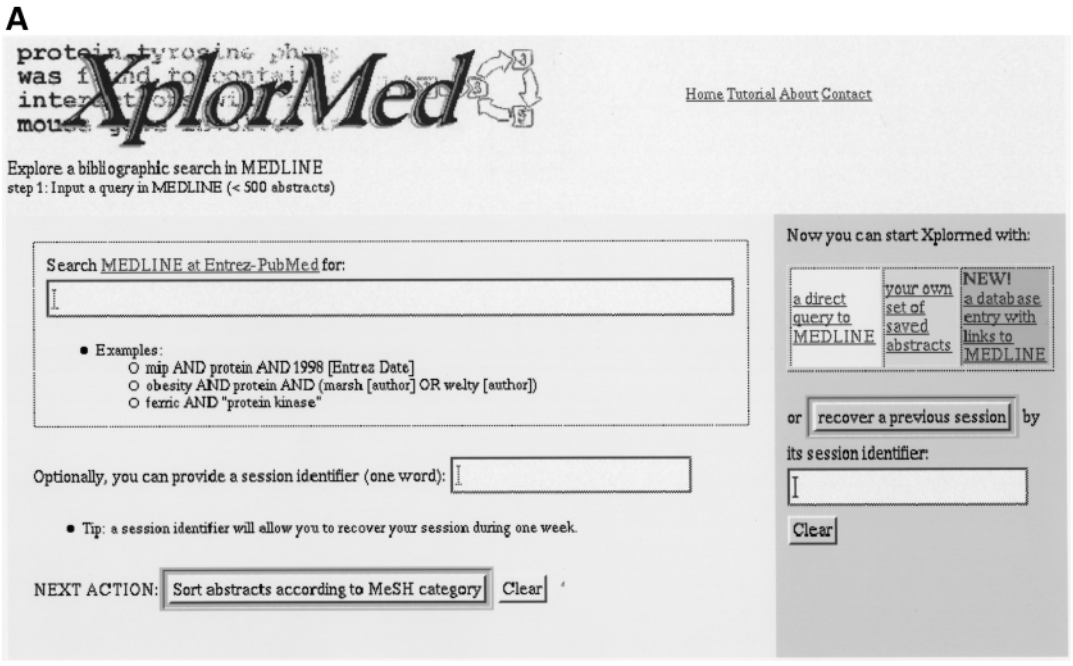
The first step involves a coarse overlapping clustering of the abstracts. References are classified into eight classes depending on their subject. Classes correspond to MeSH main categories, such as 'Anatomy', 'Organisms', 'Chemical and Drugs', 'Biological Sciences', etc. (see http://www.nlm.nih.gov/mesh/meshhome.html). You can impose an initial filtering to restrict the search to categories of interest and it is also possible to filter the search results by publication date (see page 2 of the tutorial).

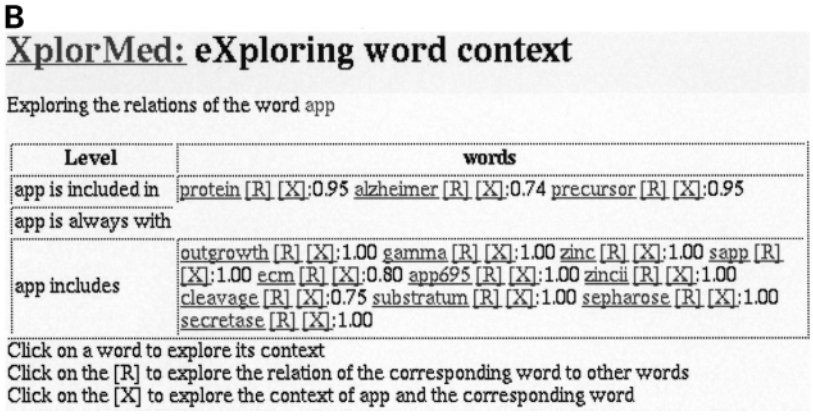
The next web page displays keywords in the selected abstracts. The method for computing keywords and relations between them can be found in literature (5). The list of extracted keywords provides a summary of the subjects within the query results and these are listed in order of relevance (more important concepts are listed first). Considering the above example of heparin and Alzheimer, XplorMed gives expected terms—'protein', 'heparin', 'alzheimer' and 'disease'—in addition to others that may be new to you, for example, 'tau' and 'app'.

At this stage, you can choose whether to go directly to the next step or to start a deeper analysis of the displayed subjects. The latter involves a context analysis of the subjects represented by the keywords and it is outlined briefly below (see Context Analysis of the Subjects). Alternatively, if you choose to go further, several groups or chains of closely related keywords are then presented to you.

You can modify the number of chains and their length by means of two parameters: *alpha* and *score* (see page 3 of

*To whom correspondence should be addressed. Tel: +49 6221 387 456; Fax: +49 6221 387 517; Email: cperez@embl-heidelberg.de





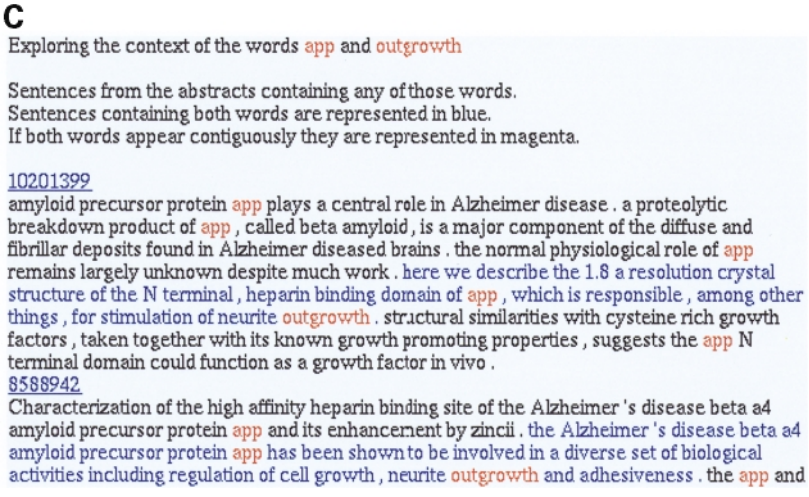


Figure 1. (**A**) XplorMed's home page. (**B**) Words related to 'app'. (**C**) Sentences containing the words 'app' and 'outgrowth'.

tutorial for details). Each chain is preceded by a number that indicates how many abstracts contain both words. By selecting one or more of these chains, you perform a sub-query of the original set. For example, suppose you are interested in protein domains that could bind heparin. Accordingly, you would inspect the pair {protein, domain}, which appears in 13 references. You can select an alternative or additional word chain if you do not find what you wanted among the proposals of the system.

The next web page provides an ordered list of abstracts; those likely to be most interesting according to your selection are highlighted on top (in our example, the papers dealing with the heparin binding domain). If you checked in the previous page the boxes for cross-linking to the corresponding databases, several hyperlinked symbols will label some abstracts (see Cross Linking to Molecular Biology Databases).

The filtered subset of papers can now be used as a new XplorMed starting point at the computation-of-keywords step (see above). Alternatively, you can expand this subset with new papers among their MEDLINE neighbors (see Expanding the Query through Related Bibliography). New keywords focusing more closely on your subject of interest will appear at this stage. The procedure can be performed repetitively and the recovery of the set of abstracts is possible at any stage.

CONTEXT ANALYSIS OF THE SUBJECTS

When the list of keywords is presented, you can explore both their meanings and relationships. By clicking on a word you can see all the sentences in the abstracts that contain that word and each sentence is linked to its MEDLINE abstract. In this way you can learn why a particular word is mentioned across the abstracts. Moreover, you can also discover interesting information by examining the words strongly related to a particular word (for example, 'app', see Fig. 1B). By clicking on the [R] next to each word, a window displaying closely related words (such as 'outgrowth' or 'zinc') will be shown. Clicking on the [X] near any related word (like 'outgrowth') shows the sentences containing either of the words ('app' or 'outgrowth') in abstracts containing both words [for example, 'The results indicate that the binding of APP to HSPG in the ECM may stimulate the effects of APP on neurite outgrowth.' (6)]. Words and sentences are highlighted in different colors for an easy identification (see page 3 of the tutorial for details). Clicking the button 'Explore the context of any word' allows you to do this kind of analysis in a more flexible way by typing other keywords of interest.

CROSS-LINKING TO MOLECULAR BIOLOGY DATABASES

As was mentioned above, the list of selected abstracts can be optionally hyperlinked to objects in several databases, currently MEDLINE, OMIM, SMART, SWISS-PROT and SpTrEMBL. The diverse symbols indicate the database and in the case of SWISS-PROT, the subject of the article, such as 'describes protein function', 'reports a 3D structure', etc. Note that the hyperlink to PubMed is always supplied, allowing you to check the content of the abstract. An additional symbol denotes review articles.

EXPANDING THE QUERY THROUGH RELATED BIBLIOGRAPHY

As mentioned above, once you have selected a subset of abstracts, it is possible to re-enter the analysis with the filtered set at the computation-of-keywords step. You can also expand this set of abstracts by retrieving neighbors from MEDLINE. Neighbors are those references that deal with the same (or similar) subject (7). To opt for this expansion you have to check the box at the bottom of the list of references. You can also change the number of neighbors included.

CONCLUSION

We have summarized how you can use the web tool XplorMed to deal more efficiently with MEDLINE literature. Because our server is being continually developed for the inclusion of new features, any suggestion from users is warmly welcomed and will be acknowledged.

ACKNOWLEDGEMENTS

We are grateful to the members of our group for their suggestions and to Robert B. Russell and to Seán I. O'Donoghue for comments to our manuscript. XplorMed uses in one step, *TreeTagger*, a part of speech tagger. We are grateful to Helmut Schmid (IMS, Stuttgart, Germany) for developing *TreeTagger* and making it publicly available.

REFERENCES

1. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2001). Xplormed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.*, **26**, 573–575.
2. *Online Mendelian Inheritance in Man, OMIM (TM)*. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000.
3. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
5. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002). Computing fuzzy associations for the analysis of biological literature. *Biotechniques*, **32**, 1380–1385.
6. Small,D.H., Nurcombe,V., Reed,G., Clarris,H., Moir,R., Beyreuther,K. and Masters,C.L. (1994). A heparin-binding domain in the amyloid protein precursor of Alzheimer's disease is involved in the regulation of neurite outgrowth. *J. Neurosci.*, **14**, 2117–2127.
7. Wilbur,W.J. and Yang,Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, **26**, 209–222.