



MedBlast: searching articles related to a biological sequence

Qiang Tu¹, Haixu Tang² and Dafu Ding^{1,*}

¹Key Laboratory of Proteomics, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China and ²Department of Computer Science and Engineering, University of California, San Diego, CA 92093, USA

Received on January 6, 2003; revised on April 4, 2003; accepted on June 13, 2003

ABSTRACT

Summary: In the genomic era, researchers often want to know more information about a biological sequence by retrieving its related articles. However, there is no available tool yet to achieve conveniently this goal. Here we developed a new literature-mining tool MedBlast, which uses natural language processing techniques, to retrieve the related articles of a given sequence. An online server of this program is also provided.

Availability: Both online server and the program are available freely at <http://medblast.sibsnet.org>

Contact: dingdafu@server.shcnc.ac.cn

INTRODUCTION

The genome sequencing projects generate such a large amount of data every day that many molecular biologists often encounter some sequences that they know nothing about. Literature is usually the principal resource of such information. It is relatively easy to mine the articles cited by the sequence annotation; however, it is a difficult task to retrieve those relevant articles without direct citation relationship.

The related articles are those described in the given sequence (gene/protein), or its redundant sequences, or the close homologs in various species. They can be divided into two classes: 'direct references', which include those either cited by the sequence annotation or citing the sequence in its text; 'indirect references', those which contain gene symbols of the given sequence. A few additional issues make the task even more complicated: (1) symbols may have aliases; and (2) one sequence may have a couple of relatives that we want to take into account too, which include redundant (e.g. protein and gene sequences) and close homologs. Here we addressed these issues and developed the software MedBlast, which can retrieve the related articles of the given sequence automatically.

ALGORITHM

MedBlast uses BLAST (Altschul *et al.*, 1997) to extend homology relationships, precompiled species-specific thesauruses, a useful semantics technique in natural language processing (NLP) (Yandell and Majoros, 2002), to extend alias relationship, and EUtilities toolset [National Center for Biotechnology Information (NCBI); http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html] to search and retrieve corresponding articles of each sequence from PubMed. The algorithm is described as follows (Fig. 1):

- (1) *BLAST and Result Filter:* MedBlast takes a sequence in FASTA format as input. The program first uses BLAST to search the GenBank nucleic acid and protein non-redundant (nr) databases, to find those homologous and corresponding nucleic acid and protein sequences. Users can input the BLAST results directly, but it is recommended to input the results of both protein and nucleic acid nr databases. The hits with low *e*-values are chosen as the relatives because the low similarity hits often do not contain specific information. Very long sequences, e.g. 100 kb, which are usually genomic sequences, are discarded too, for they do not contain specific direct references. Users can adjust these parameters to meet their own needs. Redundant multiple protein sequences in one hit are all considered, because their annotation information are sometimes not redundant.
- (2) *Direct References Retrieve:* Afterwards, MedBlast uses the Elink tool to retrieve the articles cited by the sequences and the Esearch tool to search the articles citing the sequences.
- (3) *Indirect references Retrieve:*
 - (3.1) *Symbol Extraction:* MedBlast reads the sequence and obtains the gene symbols from the 'gene' tags of the 'CDS' features in the sequence file. The 'CDS' feature must overlap with the

*To whom correspondence should be addressed.

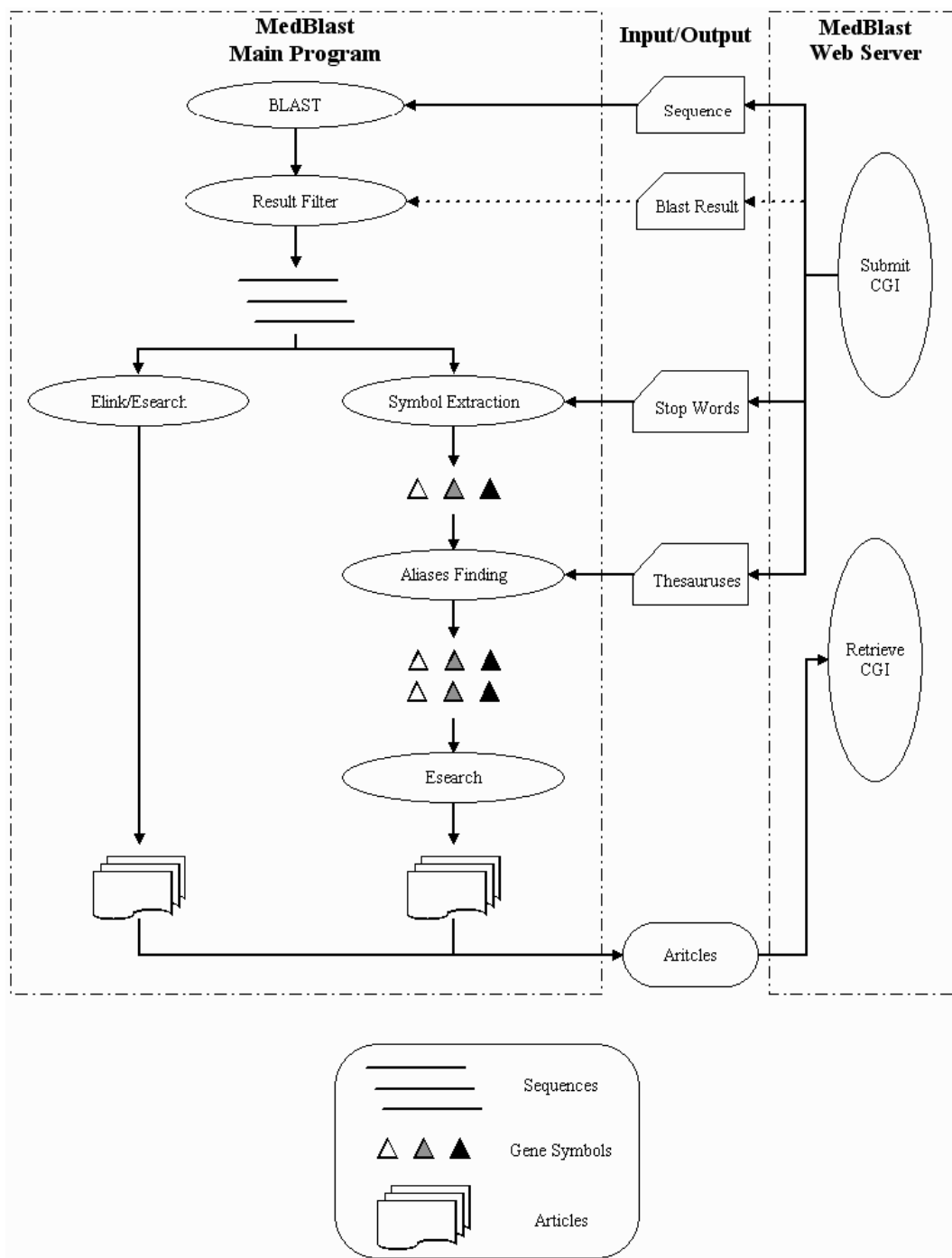


Fig. 1. The structure of MedBlast main program and web server. Users can input either the sequence or the BLAST results. The program has the default stop-word list and thesauruses, while users can also input their own stop-word list and thesauruses. The output is an HTML report of the articles.

HSP fragment of the hit, in order to neglect information from non-homologous fragments.

Sometimes, the value of 'gene' tag is not a meaningful symbol, for example, A, E + E', 14-3-3, ORF12, gene 27. We employ a set of

rules to filter these words, which include word length, non-word characters, ORF names, numbers and so on. Users may also define a stop-word list to discard particular words. Some tags contain a symbol followed by some other common words,

such as 'HSP60 gene'. In such cases the program extracts the symbol that is validated by the rules mentioned above, and discard those invalidated words.

- (3.2) *Aliases Finding*: Then, the program obtains the species name from the sequence file, looks up the symbols in the precompiled thesaurus of this species, to find other aliases.
- (3.3) *Esearch Query*: The corresponding genes/proteins in different species should be obtained by homology, not by alias, because highly homologous sequences are always the corresponding genes/proteins, while same symbols in different species may not be the same genes/proteins. Therefore, the program also tries to find the binomial species name of the sequence, and then gets the common name from NCBI taxonomy database.

Based on this information, MedBlast constructs the query word for the Esearch tool, in which the symbol and the binomial name can be in any field and common name should be in MeSH terms, for example, 'ATP5E AND ("Homo sapiens" OR human [mh])'.

- (4) *Output*: Finally, MedBlast makes a report in HTML format, including the tables summarizing all direct and indirect references. Redundant references are ignored.

IMPLEMENTATION AND RESULTS

The program is implemented in Perl, with bioperl (1.0.2, The Bioperl Project; <http://www.bioperl.org/>) and is tested under both Unix/Linux and Windows system. The program is freely available at <http://medblast.sibsnet.org>. We also maintain an online server at the same address so that users can try MedBlast without downloading. The interface of the web site is user-friendly and the documents are online.

In the current version, we have compiled the thesauruses of *Escherichia coli* [from EcoCyc (Karp *et al.*, 2002)], yeast [from Saccharomyces Genome Database (SGD) (Issel-Tarver *et al.*, 2002)], mouse [from Mouse Genome Database (MGD) (Blake *et al.*, 2002)] and human [from Human Nomenclature Database, Genew, (Wain *et al.*, 2002)]. More thesauruses will be available in the near future. Users can also define thesauruses for their own use.

A proper testing set for evaluation of the exact precision and recall of the program should contain all related articles about the given sequences, or their redundant sequences, or the close homologs in various species. Unfortunately, there is no such testing set now. We tested the program by the SGD data, which contains manually curated references of yeast genes. The testing set is randomly selected from those genes with more than 10 references, for sequences with fewer references may give unstable results. We tested 35 genes and

compared the MedBlast results with the references of each gene from SGD. The average recall is 75.1% and the precision is 47.0%. The testing set and the results are available at <http://medblast.sibsnet.org/evaluation.html>. Users should bear in mind that it is just a preliminary estimate, because SGD only collects those articles whose focuses are the single gene and particularly from yeast, while MedBlast will try to look for all related articles in all species.

In the current version, we employed only simple NLP techniques, such as ontology (thesaurus). MedBlast will be improved by using more comprehensive techniques to increase the accuracy, e.g. post-processing of the results to classify the references of large sequence projects, of gene function and so on. In addition, more thesauruses and ontology will increase the recall rate too. As a huge amount of literature and sequence data are generated in this genomic era, we believe MedBlast is a useful literature-mining tool for researchers who want to find literature references describing functions of their sequences.

ACKNOWLEDGEMENTS

We thank the staff at the Network Management Unit of Shanghai Institutes for Biological Sciences for their help in web site building. We also appreciate the staff at NCBI for their great work on EUtilities and kind help on our programming, the staffs of EcoCyc, SGD, MGD and HGNC for their great works and help on our thesauruses compiling. This work was supported by the National Natural Science Foundation of China (No. 39990600-03) and Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-2-07, KJCX1-08).

REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2002) The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. *Nucleic Acids Res.*, **30**, 113–115.
- Issel-Tarver,L., Christie,K.R., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S., Fisk,D.G. *et al.* (2002) Saccharomyces Genome Database. *Methods Enzymol.*, **350**, 329–346.
- Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
- Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
- Yandell,M.D. and Majoros,W.H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.*, **3**, 601–610.