Bioinformation Volume 5

open access

www.bioinformation.net Issue 5

Server

MULTBLAST: A web application for multiple BLAST searches

Taliah Mittler[1], Marcel Levy[2], Chad Feller[2], Karen Schlauch[1,*]

[1]Department of Biochemistry and Molecular Biology, University of Nevada, Reno; [2] Center for Bioinformatics, University of Nevada, Reno; Karen Schlauch - Email: schlauch@unr.edu; Phone: 775-784-6236; Fax: 775-784-1312; *Corresponding author:

Received September 19, 2010; Accepted September 28, 2010; Published November 1, 2010

Abstract:
Basic Local Alignment Search Tool, (BLAST) allows the comparison of a query sequence/s to a database of sequences and identifies those sequences that are similar to the query above a user-defined threshold. We have developed a user friendly web application, MULTBLAST, that runs a series of BLAST searches on a user-supplied list of proteins against one or more target protein or nucleotide databases. The application pre-processes the data, launches each individual BLAST search on the University of Nevada, Reno's-TimeLogic DeCypher® system (available from Active Motif, Inc.) and retrieves and combines all the results into a simple, easy to read output file. The output file presents the list of the query proteins, followed by the BLAST results for the matching sequences from each target database in consecutive columns. This format is especially useful for either comparing the results from the different target databases, or analyzing the results while keeping the identification of each target database separate.

Keywords: BLAST, Web application, Multiple BLAST searches.

Availability: The application is available at the URL: http://blastpipe.biochem.unr.edu/. A user name and password will be freely e-mailed upon request

Background:
Basic Local Alignment Search Tool, (BLAST) **[1, 2]** is the most frequently used algorithm for computing sequence similarity. It enables comparing a query sequence/s to a database/s of sequences and identifies sequences that are similar to the query above a defined threshold.

While there are a few BLAST applications that allow searching against several target databases, like TimeLogic's DeCypher® server **[3]** and ViroBLAST (which provides only a limited number of target databases) **[4],** they perform a single BLAST search against a combination of target databases. The present application launches a series of BLAST searches, searching with one query list of proteins against each of several requested target protein or nucleotide databases. The application merges the results into one output file, presenting the results from each target database in consecutive columns. This format is especially useful for either comparing the results from the different target databases, or analyzing the results while keeping the identification of each target database separate. MULTBLAST provides further advantages by pre–processing the query file and by allowing useful formatting options for the output file.

TimeLogic's DeCypher system (available from Active Motif, Inc.) offers a hardware-accelerated implementation of the BLAST algorithm (TeraBLAST™) **[3].** MULTBLAST utilizes these accelerated searches, thus allowing the completion of a large number of BLAST searches against large databases within a reasonable amount of time.

Methodology:
The application includes several Perl scripts (Perl v5.10.0) that perform the following tasks: upload and process the query file (including sorting, deleting duplicate identifiers, and when requested, retrieving sequences from the appropriate database), incorporate the user defined parameters into command files to be sent to the DeCypher server (ver. 8.5.0), launch a series of BLAST searches on the DeCypher server **[3],** utilizing DeCypher Client 8.5, and retrieve and combine the results into the requested output file. The searches are either protein to protein (Tera-BLASTP) or protein to DNA (Tera-TBLASTN; comparing protein sequences to nucleic sequences translated using all 6 reading frames).

Program input:
A web form prompts the user to enter job parameters and upload a query file used for all BLAST searches **(Figure 1).** The query file can be a FASTA formatted file with protein sequences from any specie/s or a personalized text file containing only a list of sequence identifier numbers, one Id on each line. If the user uploads a list of identifiers, the application will retrieve the corresponding protein sequences from the requested database. At this point the application can retrieve sequences from TAIR or, for *Caenorhabditis elegans*, from Ensembl (future plans include adding other databases). The user can select target databases against which the query will be compared from a list of 10 protein databases, including NCBI nr, Swiss-Prot, TrEMBL, UniRef100 and TAIR8, or 7 nucleotide database, including NCBI nt.

The user can define the following BLAST search parameters: maximum number of matches printed for each query id, significant threshold (only results at or better than this threshold, i.e., lower P value, will be reported) and gapped alignment processing method. Additionally, the following default settings are used: Weight matrix: BLOSUM62; Word size: 4; Query increment: 1; Extension threshold: 20; Open penalty: -11, Extend penalty: -1, Query filtered (masks repetitive elements in the data).

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5

Server

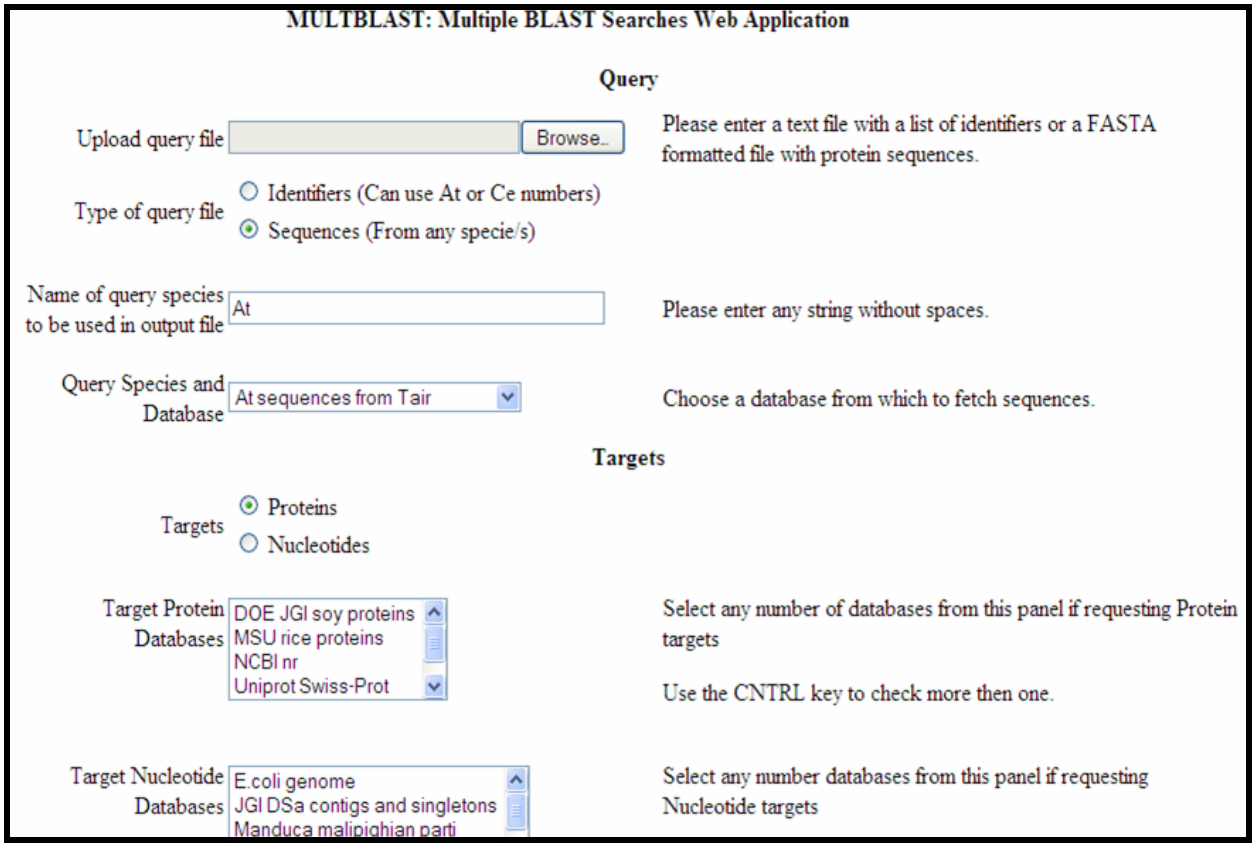

Figure 1: A screen shots of a portion of the web-application form.

Program output:

The user receives, based on his or her request, an email with either attachments of or links to the following two results files: 1) A final tab-delimited text file **(Table 1 see Supplemental material)** with a query identifier column (QUERYTEXT) followed by three columns for each target database: rank, significance level (E value) and the target sequence identifier (TARGETLOCUS). These three columns are the BLAST search results for the comparison between the respective query and target sequences. Results are based on matches between the query and each individual target database. No claims are made for matches between sequences of the different target databases. The file can be opened in Microsoft® Excel for easy viewing and further analysis. BLAST results can contain, for a given query sequence, more than one of the same target sequence. The user can request to remove or keep these duplicate target sequences. Additionally, the user can request to either print the query id in each row (useful for further analysis), or only print the first instance of each query id, leaving the query id column empty for all following rows of the same query id (allowing clearer viewing). 2) A log file for that run, containing the names of the input and output files, the target databases and the BLAST search parameters.

Future development:

Future developments will include adding other databases from which query sequences can be retrieved and additional target databases to search against, as well as incorporating an option to paste the query data into the application form in addition to the ability to upload a query file.

Acknowledgments:

We would like to thank Dr. Ron Mittler for initiating the idea behind this application as well as for his helpful comments and feedback and Dr. John Cushman and Richard Tillett for their helpful input and feedback.

References:

[1] SF Altschul *et al. J Mol Biol* **215**: 403 (1990) **[**PMID: 2231712]

[2] SF Altschul *et al. Nucleic Acids Res* **25**: 3389 (1997) [PMID:9254694]

[3] www.timelogic.com

[4] W Deng *et al. Bioinformatics* **23**: 2334 (2007) [PMID: 17586542]

Edited by Martin Gollery

Citation: Schlauch *et al.* Bioinformation 5(5): 224-226 (2010)

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5

Server

Supplementary material:

Table 1: Example of an output results file.

QUERYTEXT	RANK RICE	SIGNIFICANCE RICE	TARGETLOCUS RICE	RANK SOY	SIGNIFICANCE SOY	TARGETLOCUS SOY
AT1G01580.1	1	2.50E-125	LOC_Os04g48930.1	1	1.20E-228	Gm0149x00003
	2	2.50E-125	LOC_Os04g48930.2	2	1.00E-219	Gm0021x00288
	3	2.50E-125	LOC_Os04g48930.3	3	1.80E-176	Gm0052x00616
	4	1.30E-89	LOC_Os04g36720.1	4	1.30E-100	Gm0159x00046
				5	5.70E-93	Gm0025x00727
				6	2.60E-90	Gm0213x00025
AT1G20630.1	1	1.00E-257	LOC_Os03g03910.1	1	2.10E-266	Gm0237x00074.2
	2	4.50E-253	LOC_Os06g51150.1	2	2.70E-266	Gm0040x00008
	3	5.80E-245	LOC_Os06g51150.2	3	6.70E-265	Gm0153x00157
	4	5.90E-229	LOC_Os02g02400.1			
	5	5.30E-222	LOC_Os02g02400.2			

Table 1 is a portion of the results file for searching with *Arabidopsis* protein sequences against Rice (Rice Genome Annotation Project) and Soy (DOE JGI Soybean genome) protein databases, requesting to print only the first instance of each query id. Additional databases selected as targets would be displayed in following columns. The first Arabidopsis query (AT1G01581.1) yielded four rice sequence matches below the 0.001 significance threshold, and six soy sequence matches below the same threshold. No claims are made for matches between these rice and soy protein sequences.