NAME

MuSeqBox − a program for **Mu**lti-query **Seq**uence **B**last **o**utput e**x**amination

SYNOPSIS

MuSeqBox [**−**] [**−i** *infname*] [**−g**] [**−n** *nhits*] [**−s** *nhsps*] [**−q**] [**−c** *crtfile*] [**−o** *outfname*] [**−t|h**] [**−l** *idsize*] [**−d** *dbsize*] [**−L** *ansize*] [**−k** *srsize*] [**−p** *pstyle*] [**−z**] [**−A** *pid mao scv outfan*] [**−F** *v5s v3s v5q v3q scv qcv outffl*] [**−I** *indel type outfas*] [**−M** *mov mex*] [**−R** *rps src outfrp*]

DESCRIPTION

MuSeqBox examines multi-query BLAST (1,2) output (supported programs: BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX). The current version of **MuSeqBox** has been tested to work with BLAST+ version 2.11.0+, available from http://blast.ncbi.nlm.nih.gov/. **MuSeqBox** will only process BLAST+ output in default format (BLAST+ output option -outfmt 0). Informative parameters of BLAST hits are saved in tabular form in either simple text or HTML format. The hit tables are optionally further analyzed with the program to produce subsets of BLAST hits according to user-specified criteria. For example, BLASTX output can be further analyzed to indicate queries that may be alternatively spliced transcripts (e.g., containing a large insertion or deletion), are likely to represent full-length coding sequences, appear to be chimeric transcript assemblies, or contain repeat structures. Users of the program should cite reference (3).

The **MuSeqBox** program was originally developed in the group of Prof. Volker Brendel at Iowa State University. Updates since 2011 have been entirely by VB at his current institution. Correspondence related to **MuSeqBox** should be addressed to

Volker Brendel
Department of Biology
Indiana University
212 South Hawthorne Drive
Bloomington IN 47405
U.S.A.
phone: (812) 855-7074
email: vbrendel@indiana.edu

References:

(1) 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.
Nucl. Acids Res. 25, 3389-3402.

(2) Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2008)
BLAST+: architecture and applications.
BMC Bioinformatics 10:421.

(3) Xing, L. and Brendel, V. (2001)
Multi-query sequence BLAST output examination with MuSeqBox.
Bioinformatics 17, 744-745.

OPTIONS

– Read input from standard input. This option is typically used to pipe BLAST output directly into MuSeqBox, eliminating the need to store potentially large BLAST output files.

–i *infname*

Read input from file *infname*. The input should either be NCBI BLAST output or simple text output from a previous application of **MuSeqBox** with the print option set to default mode (i.e., *pstyle*=4, see **–p** *pstyle* option below).

–g Retrieves the subject sequence species origin if provided in the BLAST database annotation in the form "[species name]" and adds it as a column in the **MuSeqBox** output.

–n *nhits*

For each query, select the *nhits* highest scoring BLAST hits (subject sequences) from the input for tabulation and processing. The default is to select all BLAST hits from the input (up to a maximum of 2000 [default value of MaxHITs in MuSeqBox.C; change and recompile if necessary]).

–s *nhsps*

For each hit (subject sequences), select the *nhsps* highest scoring HSPs from the input for tabulation and processing. The default is to select all HSPs from the input (up to a maximum of 10000 [default value of MaxHSPs in MuSeqBox.C; change and recompile if necessary]).

–q List in the **MuSeqBox** output also queries that did not match any database subjects in the BLAST search. By default, those queries are not tabulated in the **MuSeqBox** output.

–c *crtfile*

Read the user-set criteria for selection of queries with particular types of BLAST hits from the file *crtfile*. Queries are selected if any of its HSPs meets the numerical ranges specified for: query sequence length (*QLen*), HSP (highest-scoring segment pair) length (*HLen*), database sequence (subject) length (*SLen*), percent coverage of the subject sequence (*CovS*; *HLen* relative to *SLen*), percent coverage of the query sequence (*CovQ*; *HLen* relative to *QLen*), percent identity in the HSP (*Pid*), number of indels in the HSP (*Gaps*; for gapped alignment), alignment score (*Score*), and BLAST search expectation value (*Eval*). Criteria are specified with the logical operators <>, >, >=, <, and <=, and multiple specifications are combined with the logical AND.

The format of *crtfile* is as follows. The user may edit this file to affect changes in the Operator and Setting columns. The first three and the seventh settings take integer values, the others are floating point values. Selection of the operator <> de-selects any restrictions on this particular field.

No.	Description	Item	Operator	Setting
1.	Query_Length	QLen	>	300
2.	HSP_Length	HLen	<>	100
3.	Subject_Length	SLen	<>	300
4.	Subject_Coverage	CovS	>=	30.0
5.	Query_Coverage	CovQ	>=	20.0
6.	HSP_Percent_Identity	Pid	>=	80.0
7.	Number_of_Indels	NGap	<>	20
8.	Alignment_Score	Score	<>	100.0
9.	Expected_Value	Eval	<=	1e-10

−o *outfname*
Write output to file *outfname*. An extension name '.html' will be added if the **−h** option is selected. Output is written to standard output if the **−o** option is not selected.

−t|h Write results in either simple text format (**−t**; default) or HTML format (**−h**). If HTML format is selected, then the QueryID and SubjectID column entries are linked to NCBI Entrez queries (appropriate only if Query and Subject sequences are in standard GenBank format [exception: if the database is formatted with entries of type >gnl|UniProt|name, then the Subject entries are linked to the UniProt resource]). The Eval column entries are linked to NCBI BLAST to facilitate up to date database searches with the indicated queries.

−l *idsize*
Set column width (number of characters) for QueryID and SubjectID (default: 12; maximum: 50). Increase *idsize* if the entries in the formatted output do not align correctly.

−d *dbsize*
Set column width (number of characters) for Db column (default: 5; maximum: 16). Increase *dbsize* if the entries in the formatted output do not align correctly.

−L *ansize*
Set column width (number of characters) for Annotation (default: 36; maximum: 240).

−k *srsize*
Set column width (number of characters) for Source (default: 24; maximum: 50).

−p *pstyle*
1: print columns with text descriptions only; 2: print columns with numeric values only; 3: condensed print format, and 4: detailed print format (default).

−z Allow MuSeqBox to read BLAST output as input when the subject and/or query sequences are not in GenBank format. Note, however, that the links in the HTML output may not work. Additionally, table formatting will likely need to be corrected by specifying a suitable *idsize* parameter to the **-l** flag.

The following options are described for applications to post-processing of BLASTX output, but apply to other BLAST programs with obvious modifications:

−A *pid mao scv outfan*
Select queries that are globally highly similar to matching protein subjects. Required arguments: *pid*, minimal percent identity in each HSP; *mao*, maximal allowed overlap of selected HSPs; *scv*, cumulative percent coverage of the matched subject (sum of *CovS* for all selected non-overlapping and/or maximal allowed overlapped HSPs); *outfan*, file name to output the selected queries.

−F *v5s v3s v5q v3q scv qcv outffl*
Select queries that potentially encode full-length coding sequences (for BLASTX applications). Required arguments: *v5s*, maximal length of the variable 5'-terminal region (not contained in any HSP) of the subject sequence; v3s, maximal length of the variable 3'-terminal region of the subject sequence; v5q, v3q, similarly for the query sequence. Precisely, *-F v5s v3s v5q v3q* selects subject/query pairs for which the variable segment (not covered by HSPs) at the 5'-end is either at most *v5s* letters in the subject sequence OR at most *v5q* letters in the query sequence, AND the

variable segment at the 3'-end is either at most *v3s* letters in the subject sequence OR at most *v3q* letters in the query sequence. *scv*, cumulative percent coverage of the matched subject (sum of *covS* for all non-overlapping HSPs); *qcv*, cumulative percent coverage of the query (sum of *covQ* for all non-overlapping HSPs); *outffl*, file name to output the selected queries.

–I *indel type outfas*

Select queries that represent potential alternatively spliced transcripts. Required parameters: *indel*, the minimal size of the insertion or deletion (number of nucleotides for BLASTN, BLASTX, and TBLASTX; number of amino acids for BLASTP and TBLASTN); *type*, 1 for insertions in the query sequence (separated HSPs in the query, continuous HSPs in the subject - representing potential retained introns or additional exons), 2 for insertions in the subject sequence (continuous HSPs in the query, separated HSPs in the subject - representing potential skipped exons); *outfas*, file name to output the selected queries.

Note: Short indels may be represented by gaps within a single HSP when using gapped BLAST. Such cases may be detected by setting the *Gaps* criterion in the parameter file specified by the **–c** *crtfile* option.

–M *mov mex*

Indicate potential chimeric queries. Potential chimeric queries have two hits to different subject sequences, with the hits overlapping at most *mov* positions and the endpoint of the second hit extending at least *mex* positions beyond the endpoint of the first hit. Potential chimeric queries are identified in the output in lines starting with ' !Potential chimeraXY:', where 'XY' is '++', '+-', '-+', or '--' based on the orientation of the two hits. BLASTX hits of opposite sign are most likely chimeric queries resulting for example from fused transcripts during computational assembly.

–R *rps src outfrp*

Select queries that may contain repeats or align to database sequences with repeats. Required parameters: *rps*, minimal potential repeat size (number of nucleotides or amino acids, depending on the type of query and subject sequences); *src*, 1 to identify potential repeats in the query, 2 to identify potential repeats in the subject sequence; *outfrp*, file name to output the selected queries.

NOTES

A hardcopy of this manual page is obtained by 'man -t ./MuSeqBox.1 | lpr'.

You can change default settings of certain parameters in the MuSeqBox.C source code (#define's) and recompile.

VERSION

MuSeqBox-5.8 (13 January, 2024)

BUGS

- none known. Please report.

AUTHORS

Liqun Xing and Volker Brendel <vbrendel@indiana.edu>.