MBlastP

Version 1.4.2 04/15/2011 (Linux)

Usage

mblastp

-h - Prints this message.

-v - Prints MBlastP version.

mblastp -c <HashFileName> -d <DataDir> -e <outEValue[1.0E-05]> -f <OutFmt>

-m <MaxOutMSPs> -M <RAMMemory(in Gb)> -n <NumHashFiles>

-o <OutputFile[mblastx.out]> -q <QueryFile> -s <ScoringMatrix[BLOSUM62]>

-w <wordSize>

-c <HashFileName> - Use HashFileName for Hash Files. Default – 'ProtHash'

-d <DataDir> - Use DataDir for mblastp internal files.

Default - use $MBLASTX_DATADIR or current directory.This parameter overrides MBLASTX_DATADIR environment variable.

-e <outEValue> - Maximum EValue for MSPs. Default EValue is 1.0E-05.

-f <outFmt> - Output Format specifier.

T : Tab delimited (Default)

F : Free-form output

S : Tab Delimited with Query and Reference Strings suppressed.

-m <MaxOutMSPs> - Maximum number of MSPs to print. Default value is 10.

-M <RamMemory> - Amount of RAM Memory available in Gb. Default value is 48.

-n <NumHashFiles> - Number of Hash Files to process.

Default - get the number of hash files from Hash Header file.

-o <OutputFile> - Name of the output file. Default name is 'mblastx.out'.

File Name 'STDOUT' or '-' will force the output to standard output.

-q <QueryFile> - Name of the file containing queries.

-s <ScoringMatrix> - Name of the Scoring Matrix to use.

Default matrix is BLOSUM62.

-t <Threashold> - Minimum word score. Default word score is 26.

-T <NumThreads> - Number of threads to use. Default is 16

-w <wordSize> - Use wordSize for Hit Finding. Default is 15
Value should be 3, 6 - 18.

Sensitivity and Known Limitations

- In empirical testing, the default sensitivity of MBlastP running at word size 6 and word threshold 26 and single hit mode was found equivalent to NCBI BlastP running at word size 6 and word threshold 23 in the dual hit mode (default).
- Requires a minimum of 12Gb of RAM on the system.

Examples

1. Specifying the input query file
   mblastp *-q exampleQuery.fasta*

The -q parameter specifies the input query sequence data file, which is expected to be a valid FASTA file. This parameter specification is mandatory and there is no default value. In the above example, the file example Query.fasta is the name of the query file specified and is expected to be in the current directory.

The hash files for the Reference Sequence Database, prefixed '*ProtHash'* are expected to be in the current directory. In case the MBLASTX_DATADIR environment variable is set, the hash files for the Reference Sequence Database are expected to be in the directory pointed to by this environment variable.

2. Specifying the data directory for MblastP
   mblastp -q exampleQuery.fasta -d /home/mblastp/data

With the -d parameter, the hash files for the Reference Sequence database are expected to be in specified directory. This is irrespective of the setting for the MBLASTX_DATADIR environment variable. In the above example, the data directory used is */home/mblastp/data.*

3. Specifying the Hash File name
   mblastp -q exampleQuery.fasta -c nr_091003

With the -c parameters, the hash file prefix for the Sequence Reference File to be used by the MBlastP is specified. In the above example, the file prefix used is 'nr_091003', instead of the default 'ProtHash'. This parameter is useful is more than one set of hash files are created in the same directory.

4. Specifying the output format

By default, the output of MBlastP is in a tab-delimited format, closely resembling the first tab-delimited format produced by NCBI BlastP. This format can also be explicitly specified using *-f T*. The second possible format closely resemble the default free-form format of

NCBI BlastX. This format can be chosen by using *-f F* format specification, as in the following command line:

mblastp -q exampleQuery.fasta -f F

The third possible format is the default tab-delimited format with the Query and Reference portions of the HSP string suppressed, specified using *-f S*. This format is useful when run for a very large number of queries, reducing the size of the output file.

Undocumented Features for Sensitivity Considerations

(Confidential - For UMIGS only)

A few undocumented parameters are exposed as tunable filter options for the use of *UMIGS*. The following table lists the undocumented parameter combinations to be used in addition to other parameters for each sensitivity level and their corresponding expected X Factor:

Identity %age in HSP Expected X Factor Parameter Combination

50 525 -F I -t 26 -Z 5 -X 7 -Y 20

40 300 -F I -t 26 -Z 3 -X 5 -Y 12

30 150 -F S -t 22 -Z 2 -X 5 -Y 42