Configuring BLAST

Tom Madden[1] and Christiam Camacho[⊠2]

Created: June 23, 2008; Updated: August 14, 2024.

The BLAST+ search applications can be configured by means of a configuration file or environment variables.

Configuring BLAST via configuration file

This can be accomplished with a configuration file named .ncbirc (on Unix-like platforms) or ncbi.ini (on Windows). This is a plain text file that contains sections and key-value pairs to specify configuration parameters. Lines starting with a semi-colon are considered comments. The application will search for the file in the following order and locations:

1) Current working directory (*)

2) User's HOME directory (*)

3) Directory specified by the NCBI environment variable

4) The standard system directory ("/etc" on Unix-like systems, and given by the environment variable SYSTEMROOT on Windows)

(*) Unless the NCBI_DONT_USE_LOCAL_CONFIG environment variable is defined.

The search for this file will stop at the first location where it is found and the configurations settings from that file will be applied. If the configuration file is not found or if the NCBI_DONT_USE_NCBIRC environment variable is defined, the default values will apply. The following are the possible configuration parameters that impact the BLAST+ applications:

Configuration Parameter	Specifies	Default value
BLASTDB	Path to BLAST databases.	Current working directory
DATA_LOADERS	Data loaders to use for automatic sequence identifier resolution. This is a comma separated list of the following keywords: blastdb, genbank, and none. The none keyword disables this feature and takes precedence over any other keywords specified.	blastdb,genbank

Author Affiliations: 1 Email: madden@ncbi.nlm.nih.gov 2 Email: camacho@ncbi.nlm.nih.gov

⊠ Corresponding author.

Table continued from previous page.

BLASTDB_PROT_DATA_LOADER	Locally available BLAST database name to search when resolving protein sequences using BLAST databases. Ignored if DATA_LOADERS does not include the blastdb keyword.	nr
BLASTDB_NUCL_DATA_LOADER	Locally available BLAST database name to search when resolving nucleotide sequences using BLAST databases. Ignored if DATA_LOADERS does not include the blastdb keyword.	nt
BLAST_USAGE_REPORT	Specifies whether or not usage information should be returned to the NCBI. Set this value to false to disable this feature.	true
WINDOW_MASKER_PATH	Path to windowmasker directory hierarchy.	Current working directory

The following is an example with comments describing the available parameters for configuration:

```
; Start the section for BLAST configuration
[BLAST]
; Specifies the path where BLAST databases are installed
BLASTDB=/home/guest/blast/db
; Specifies the data sources to use for automatic resolution
; for sequence identifiers
DATA_LOADERS=blastdb
; Specifies the BLAST database to use resolve protein sequences
BLASTDB_PROT_DATA_LOADER=custom_protein_database
; Specifies the BLAST database to use resolve protein sequences
BLASTDB_NUCL_DATA_LOADER=/home/some_user/my_nucleotide_db

; Windowmasker settings
[WINDOW_MASKER]
WINDOW_MASKER_PATH=/home/guest/blast/db/windowmasker
; end of file
```

Configuring BLAST via environment variables

Please note that the environment variables take precedence over any settings from the NCBI configuration file.

Environment Variable	Specifies
NCBI	Path to NCBI configuration file.
NCBI_DONT_USE_NCBIRC	If defined, no NCBI configuration file will be used.
NCBI_DONT_USE_LOCAL_CONFIG	If defined, no NCBI configuration file on the local directory or the user's HOME directory will be used
BLASTDB	Path to BLAST databases.

Table continued from previous page.

BLASTMAT	Path to scoring matrix files.
BATCH_SIZE	See "Controlling concatenation of queries" and "Memory usage" sections below.
NCBI_CONFIG__BLAST__X	Assuming X is any of the configuration parameters from the previous section, it serves the same purpose.
BLAST_USAGE_REPORT	Specifies whether or not usage information should be returned to the NCBI. Set this variable to false to disable this feature.

Controlling concatenation of queries

As described above, BLAST+ works more efficiently if it scans the database once for multiple queries. This feature is knows as concatenation. Unfortunately, for some searches the concatenation values are not optimal, too many queries are searched at once, and the process can consume too much memory. For applications besides BLASTN (which uses an adaptive approach), it is possible to control these values by setting the BATCH_SIZE environment variable. Setting the value too low will degrade performance dramatically, so this environment variable should be used with caution.

Memory usage

The BLAST search programs can exhaust all memory on a machine if the input is too large or if there are too many hits to the BLAST database. If this is the case, please see your operating system documentation to limit the memory used by a program (e.g.: ulimit on Unix-like platforms). Setting the BATCH_SIZE environment variable as described above may help.

Configure limit on the number of open files

On some systems (e.g., those running macOS), the default limit on the number of open files is set in a way that doesn't allow BLAST+ applications to operate. BLAST+ will display an error similar to the following:

```
Error memory mapping: /Users/johnny/nr.25.phr openedFilesCount=251 threadID=0
BLAST Database error: Cannot memory map /Users/johnny/nr.25.phr. Number of files opened: 251
```

To address this, try increasing the number of open files, e.g.:

```
ulimit -n unlimited
```

If the command above doesn't work, try specifying a positive integer as its argument, e.g.:

```
ulimit -n 65536
```