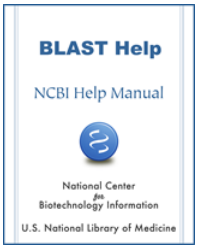
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Tao T, Madden T, Christiam C. BLAST FTP Site. 2011 May 29 [Updated 2020 Aug 30]. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



BLAST FTP Site

Tao Tao, PhD,[1] Tom Madden, PhD,[2] and Camacho Christiam[3]

Created: May 29, 2011; Updated: August 30, 2020.

The NCBI FTP server contains a BLAST-specific directory (https://ftp.ncbi.nlm.nih.gov/blast/). Through this directory, the standalone BLAST packages and a standard set of BLAST databases are available to the public for download through anonymous FTP. For faster download, the service is also available through the Aspera client for those users with the Aspera browser plug-in installed (https://www.ncbi.nlm.nih.gov/public/?blast/).

This document describes the subdirectories and file contents of the BLAST FTP directory. Technical details on how to use certain files, especially those under the db (database) subdirectory, are also provided.

Subdirectories under the BLAST FTP directory

There are several subdirectories under the BLAST FTP directory. Each stores a set of files with similar types of content. These subdirectories are summarized in Table 1.

Table 1. File content of subdirectories under the "/blast" FTP directory

Subdirectory	File content
db	Preformatted BLAST database files and FASTA sequence files (only for a few representative databases, kept under the /FASTA subdirectory)
demo	Various demonstration packages for software developers
documents	Preliminary documentation (mostly from software developers) and pointers to other documentation
executables	Different releases for standalone BLAST packages, including blast+
matrices	Different scoring matrices, only a selected subset are supported by blast
temp	Miscellaneous files
WGS_TOOLS	Perl scripts for generating WGS project-based database alias for TSA and WGS datasets, to be used with vdb blast
windowmasker_files	A collection of windowmasker files, organized into subdirectory named by the taxonomic ids

The "*/db*" subdirectory

This subdirectory contains a common set of preformatted BLAST database files in version 5 format. The FASTA sequences for a few widely used databases are stored under the "/FASTA" subdirectory. The contents of available preformatted databases are summarized separately according to their sequence nature and sources (for

Author Affiliations: 1 NCBI; Email: tao@ncbi.nlm.nih.gov. 2 NCBI; Email: madden@ncbi.nlm.nih.gov. 3 NCBI; Email: camacho@ncbi.nlm.nih.gov.

Corresponding author.

nucleotide databases). The databases provided for the cloud-based BLAST packages are under the "/cloud" subdirectory. The version 4 databases are kept in the "/v4" subdirectory, those entries will not be updated. The "/v5" subdirectory is a soft link back to the "/db" directory.

Table 2a. Preformatted protein database files

File name	Contents
landmark.tar.gz	The landmark database includes complete proteomes from a few selected representative genomes spanning a wide taxonomic range, the main database used by the SmartBLAST services.
cdd_delta.tar.gz	Condensed conserved domain database for use with deltablast protein searches.
nr.##.tar.gz	A collection of protein sequences with entries from GenPept, Swissprot, PDB, PRF, PIR and NCBI Reference Sequence (RefSeq) project.
pataa.tar.gz	Protein sequences from patents as supplied by USPTO. These entries are EXCLUDED from the nr database.
pdbaa.tar.gz	Protein sequences from PDB structure records' protein components.
refseq_protein.##.tar.gz	Protein sequences from NCBI RefSeq project.
swissprot.tar.gz	Protein sequences from the swiss-prot sequence database (last major update).
tsa_nr.##.tar.gz	Protein sequences from the Trascriptome Shotgun Assembly. Its entries are EXCLUDED from the nr database.
env_nr.##.tar.gz	Protein sequences from large environmental sequencing projects, e.g., Sargasso Sea, Acid Mine Drainage. Its entries are EXCLUDED from the nr database.

Table 2b. Preformatted RefSeq nucleotide database files

File	Contents
16S_ribosomal_RNA.tar.gz	Microbial 16S RNA sequences from the RefSeq Targeted Loci project (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/).
refseq_rna.##.tar.gz	RNA sequences from NCBI RefSeq project, also included in the nt database.
refseq_euk_rep_genomes.##.tar.gz	Eukaryotic representative genomes from NCBI RefSeq project
refseq_prok_rep_genomes.##.tar.gz	Prokaryotic representative genomes from NCBI RefSeq project
refseq_viroids_rep_genomes.##.tar.gz	Viriods representative genomes from NCBI RefSeq project
refseq_viruses_rep_genomes.##.tar.gz	Viruses representative genomes from NCBI RefSeq project
human_genome.##.tar.gz	Current refseq human genome assembly (GRCh) with various database masking
mouse_genome.##.tar.gz	Current refseq mouse genome assembly (GRCm) with various database masking

Table 2c. Preformatted non-RefSeq nucleotide and target loci databases

File	Contents
nt.##.tar.gz	The nucleotide sequence database contains entries from traditional divisions of GenBank, EMBL and DDBJ. Sequences from bulk divisions, i.e., gss, sts, pat, est, htg, wgs, con, and environmental sequences are excluded. RefSeq genomic entries are also excluded.
patnt.##.tar.gz	Nucleotide sequences from patents as supplied by USPTO to GenBank, or from EU/Japan Patent Agencies through EMBL/DDBJ. Entries are **EXCLUDED** from the nt database.
pdbnt.##.tar.gz	Sequences for the nucleotide components of PDB structure records.
tsa_nt.##.tar.gz	A database with earlier non-project based Transcriptome Shotgun Assembly (TSA) entries. Project-based TSA entries are NOT included. Entries are **EXCLUDED** from the nt database.
ITS_*.tar.gz	Databases with collection fungal or eukaryotic Internal Transcribed Spacer sequences.
LSU_*_rRNA.tar.gz	Database with large submit rRNA sequences for prokaryotes and eukaryotes.

Table 2c. continued from previous page.

File	Contents
SSU_*_rRNA.tar.gz	A database with sequences small from fungi and eukaryotes
taxdb.tar.gz	A non-sequence database file containing taxonomic information for sequences in the preformatted databases providing common and scientific names for each entry.

Getting the preformatted database files

Preformatted BLAST database files offer several advantages over the FASTA files:

- The preformatted databases are broken into smaller volumes and therefore can be downloaded more readily with fewer errors
- A convenient Perl script (***update_blastdb.pl*** found in the bin directory of a locally installed blast+ package) is available to simplify the download of these preformatted databases
- Preformatted database files remove the makeblastdb formatting steps, and saves valuable processing time and diskspace
- Taxonomic information is encoded within the preformatted databases and can be used to limit the scope of a blast search, and sequence retrieval, and scientific name addition through the included taxdb files
- Sequences in FASTA format can be generated easily from the preformatted databases using the blastdbcmd utility when needed

Preformatted databases must be downloaded in binary mode, downloading through the *update_blastdb.pl* script is recommeded. An example command line for getting the preformatted refseq_rna nucleotide database and the session output are given below.

```
$ perl ../bin/blast+/update_blastdb.pl --passive --decompress refseq_rna
Connected to NCBI
Downloading refseq_rna (7 volumes) ...
Downloading refseq_rna.00.tar.gz... [OK]
Downloading refseq_rna.01.tar.gz... [OK]
Downloading refseq_rna.02.tar.gz... [OK]
Downloading refseq_rna.03.tar.gz... [OK]
Downloading refseq_rna.04.tar.gz... [OK]
Downloading refseq_rna.05.tar.gz... [OK]
Downloading refseq_rna.06.tar.gz... [OK]
Decompressing refseq_rna.00.tar.gz ... [OK]
Decompressing refseq_rna.01.tar.gz ... [OK]
Decompressing refseq_rna.02.tar.gz ... [OK]
Decompressing refseq_rna.03.tar.gz ... [OK]
Decompressing refseq_rna.04.tar.gz ... [OK]
Decompressing refseq_rna.05.tar.gz ... [OK]
Decompressing refseq_rna.06.tar.gz ... [OK]
```

The complete options of this script (obtained using specific option "--help") are shown below.

```
$ perl update_blastdb.pl --help
NAME
    update_blastdb.pl - Download pre-formatted BLAST databases

SYNOPSIS
    update_blastdb.pl [options] blastdb ...

OPTIONS
    --decompress
      Downloads, decompresses the archives in the current working directory,
```

```
      and deletes the downloaded archive to save disk space, while
      preserving the archive checksum files (default: false).

    --showall
      Show all available pre-formatted BLAST databases (default: false). The
      output of this option lists the database names which should be used
      when requesting downloads or updates using this script.

      It accepts the optional arguments: 'tsv' and 'pretty' to produce
      tab-separated values and a human-readable format respectively. These
      parameters elicit the display of additional metadata if this is
      available to the program. This metadata is displayed in columnar
      format; the columns represent:

      name, description, size in gigabytes, date of last update (YYYY-MM-DD
      format).

    --blastdb_version
      Specify which BLAST database version to download (default: 4).
      Supported values: 4, 5

    --passive
      Use passive FTP, useful when behind a firewall or working in the cloud
      (default: true). To disable passive FTP, configure this option as
      follows: --passive no

    --timeout
      Timeout on connection to NCBI (default: 120 seconds).

    --force
      Force download even if there is a archive already on local directory
      (default: false).

    --verbose
      Increment verbosity level (default: 1). Repeat this option multiple
      times to increase the verbosity level (maximum 2).

    --quiet
      Produce no output (default: false). Overrides the --verbose option.

    --version
      Prints this script's version. Overrides all other options.

    --num_cores
      Sets the number of cores to utilize to perform downloads in parallel
      when data comes from GCS. Defaults to all cores (Linux and macos
      only).

DESCRIPTION
    This script will download the pre-formatted BLAST databases requested in
    the command line from the NCBI ftp site.

EXIT CODES
    This script returns 0 on successful operations that result in no
    downloads, 1 on successful operations that downloaded files, and 2 on
    errors.

BUGS
    Please report them to <blast-help@ncbi.nlm.nih.gov>
```

```
COPYRIGHT
    See PUBLIC DOMAIN NOTICE included at the top of this script.
```

Using the preformatted BLAST database files

The *--decompress* option of *updated_blastdb.pl* automatically decompresses and extract the archives of the requested database files. When manually downloading preformatted databases, those compressed archives must be downloaded in binary format using the passive mode, then inflated with ***gunzip*** or other decompress utilities. The working database files can then be extracted out of the resulting tar archive using *tar* program in Unix/Linux or *WinZip* and *StuffIt Expander* on Windows and Macintosh platforms, respectively.

Large databases are formatted in multiple gigbytes-sized volumes, which are named using the "name.##.tar.gz" convention. To reconstitute a given multi-volume database, all volumes with the same base database name are required. A database alias file, with ".pal" extension for protein or ".nal" extension for nucleotide, is provided to tie the volumes together. The database can be called using the base database name. For example, binary programs from the blast+ package can call the nt database using the command line option of "**-db nt**" option argument.

For proper setup of standalone blast+, it is recommended that database files be stored in a centralized directory, with the path to this directory be encoded by the BLASTDB variable. Details are available in the setup document for Windows and Mac/Linux/Unix.

Sequence files under the "/db/FASTA/" subdirectory

This subdirectory contains sequence files in the FASTA format. With preformatted databases readily available, only a few commonly used databases are available in this format.

Table 3. Protein database files under the /db/FASTA directory

File	Content
nr.gz	The FASTA equivalent of the nr.##.tar.gz database files
swissprot.gz	The FASTA equivalent of the swissprot.tar.gz database file.
nt.gz	The FASTA equivalent of the nt.##.tar.gz database files.

For local BLAST searches, the recommendation is to use the preformatted version given in the parent directory. For those without preformatted counterparts, the FASTA sequence file first need to be inflated using gunzip or other comparable utilities, the resulting file can then be formatted by *makeblastdb* from the blast+ package. Example command lines for formatting igSeqNt and igSeqProt are given below.

```
$ makeblastdb –in swissprot –dbtype prot –parse_seqids
$ makeblastdb –in nt –dbtype nucl –parse_seqids
```

For vector screening needs, get the FASTA sequences from this ftp directory and formatted them using makeblastdb:

```
https://ftp.ncbi.nlm.nih.gov/pub/UniVec/
```

Chromosome entries are available in the refseq_euk_rep_genomes and refseq_prok_rep_genomes databases. Organism-specific sequences are available in FASTA format under the genomes FTP directory:

```
https://ftp.ncbi.nlm.nih.gov/pub/genomes/all/ https://ftp.ncbi.nlm.nih.gov/pub/factsheets/
Factsheet_Assembly.pdf
```

The NCBI Datasets tool is another way to get the genomic data from NCBI. Please refer to this page for more information: https://www.ncbi.nlm.nih.gov/datasets

Database files under the "/db/v4/" subdirectory

This subdirectory contains the preformatted blast dtabase files in the older version 4 format. Those databases will not be updated. They are meant as a stop-gap measure to ease the transition to version 5 of the databases. Content description for this subdirectory will be skipped.

Database update

In general, BLAST databases are updated daily. There is no established incremental update scheme due to sequence removal and update. It is recommended that databases be downloaded at regular intervals to keep the content of local copy current. The ***update_blastdb.pl*** script can help streamline this download process. If the original database.##.tar.gz files are kept, this utility can automatically check the time stamps to determine if file refreshing is required or not.

Faster download through Aspera plug-in is also possible (downloadable from the Aspera Soft site under the download tab). A web interface for the NCBI FTP site is at: https://www.ncbi.nlm.nih.gov/public/. Aspera's commandline client ascp can be used to access Aspera indexed NCBI ftp site.

Contents of the "/blast/demo/" subdirectory

This directory contains technical presentations given by NCBI BLAST developers in scientific conferences and several tools and documents relevant to the BLAST service to deomonstrate how specific functions from NCBI's C-toolkit code can be used.

Table 4. Contents of the /blast/demo/ subdirectory

File/Dir Name	Content
QUICKBLASTP	Standalone kmer indexing tool and the kmer blastp tool. The algorithm is used by the web protein blast search when "Quick BLASTP (Accelerated protein-protein BLAST) " option is selected
README.quickblastp	Readme for quickblastp
quickblastp.tar.gz	Kmer-based quickblastp demonstration package
benchmark	Package with sample database and query for gauging the performance of BLAST releases on different platforms
bmc	Sequences used in the generation of BLAST search data for the blast+ paper published in BMC
igblast	Directory with igblast related set of scripts, see its README for more details
magicbkast_article	Supplemental materials for the magicblast paper and the binary used to generate the specs
blast_programming.ppt	PowerPoint presentation on BLAST programing
mt_tback.tgz	Multi-threaded traceback related test code
openmp_test.tar.gz	Openmp multi-thread related test code
parse_blast_xml.tar.gz	Demo package on parsing xml styled blast output
test_suite.tar.gz	An old set of test sequences with csh script
vecscreen	Binary program for vecscreen
*.ppt, *.pdf	Slides or posters presented by the BLAST group in various conferences
*.fsa	Miscellaneous sequences

Since NCBI is migrating to C++ code base provided by the C++ toolkit, some of the files from this directory could become obsolete or change without notice. Most of the functions demonstrated here are incorporated in the *blast_formatter* utility distributed in the blast+ package.

Contents of the "/blast/documents/" subdirectory

This directory contains mostly posters and other preliminary documentation from BLAST developers. For blast+ packages, a user manual along with the instruction for installation are available through NCBI bookshelf. Content description will be skipped. See the README for this directory for details: https://ftp.ncbi.nlm.nih.gov/blast/documents/README

Contents of the "/blast/executables/" subdirectory

This directory contains several subdirectories each for a set of BLAST distribution packages from a specific release. Binaries based on the new C++ toolkit are under the /blast+ subdirectory with the lastest release directly accessible through the /LATEST symbolic link. The only program file is **remote_fuser**, for use in database fetching in the cloud implementation.

Contents of the "/blast/executables/LATEST/" subdirectory

This directory is a symbolic link pointing to the LATEST release of BLAST+ programs built from the NCBI C++ toolkit. Packages for commom platforms available in different formats are summarized in Table 5 below.

Table 5. File content of the /blast/executables/LATEST/ subdirectory

File	Contents
ChangeLog	Changes introduced in this release
ncbi-blast-*src.*	blast+ source code in different formats
ncbi-blast-#.#.##+.x86_64.rpm	rpm installation package for PC running 64-bit Linux
ncbi-blast-#.#.##+-x64-linux.tar.gz	Tar archive for PC running 64-bit Linux
ncbi-blast-*.dmg	Disk image for Macintosh running 64-bit OSX
ncbi-blast-*-macosx.tar.gz	Tar archive for Macintosh running 64-bit OSX
ncbi-blast-*-win64.exe	Installer for PC running 64-bit Windows
ncbi-blast-#.#.##+-x64-win64.tar.gz	Tar archive for PC running 64-bit Windows

All non-source code archives or packages are equivalent. They contain a standard collection of standalone command line programs and accessory utilities for different platforms. Installation of the package enables local BLAST searches, custom database preparation from FASTA sequences, as well as sequence retrieval from existing databases formatted with the "-parse_seqids" argument.

Details on individual programs from the package as well as installation procedures are available for Windows and Mac/Linux/Unix.

Note that blast+ does not provide a separate client-server tool. That function is built into individual blast programs, i.e. *blastn*, *blastp*, *blastx*, *tblastn* and *tblastx*, and can be invoked using the "-remote" option. The "-remote" option enables remote search against databases at NCBI using NCBI's computation resources. In addition, blast+ does not provide package equivalent to the decommissioned wwwblast package.

Contents of the "/blast/executables/blast+/" subdirectory

This subdirectory archives all releases of the blast+ package. Each release is under its own directory named using its version number. Available builds start at version 2.2.18, with verision 2.2.20 skipped. Optional version 5 database support began in release 2.8.0alpha.

Contents of the "/blast/executables/legacy.NOTSUPPORTED/" subdirectory

This subdirectory archives the releases of legacy blast package based on the NCBI C Toolkit. They are meant for historical references, NCBI no longer supports them. Each release is under its own directory with the version number as its name. Available builds start at version 2.0.7.

Packages with version number 2.2.10 or newer are packaged with a built-in directory structure to better organize the distributed contents.

Contents of the "/blast/executables/igblast/" subdirectory

This subdirectory archives the different releases of standalone igblat package (under the release subdirectory. The separate data files and database files required by the binary programs in this package are available in subdirectories separate from each releases.

Table 6. File content of the /blast/executables/igblast/release/ subdirectory

File	Contents
#.#.#	Different release directories
LATEST	Soft link pointing to the latest release subdirectory
database *	germline immunoglobulin gene sequences for mouse and rhesus monkey. Databases for other organisms are NOT provided due to license requirement.
Internal_data	Internal vdj gene information file for annotation need
optional_file	Additional gene annotation file
edit_imgt_file.pl *	A perl script for manipulating the deflines of IMGT immunoglobulin sequences so they can be used as input to make databases through makeblastdb

* Note: Internal_data, optional_file, and database directories are included in release 1.13.0 or later; use the files downloaded wit hthe release. Additionally, license requirement prevents NCBI from distributing the germline sequences for certain organisms as blast-ready databases. Instead, those sequences should be obtained from IMGT as FASTA. Convert the deflines of the IMGT sequences using the "edit_imgt_file.pl" script first before using makeblastdb to format the file into a igblast readable database.

Contents of the "/blast/executables/magicblast/" subdirectory

A subdirectory for the releases of a next generation sequence read mapper from NCBI. The LATEST directory maps to the current release's directory. Refer to the README file under the subdirectory for more details. Details technical description of the package is at: https://ncbi.github.io/magicblast/.

Contents of the "/blast/executables/rmblast/" subdirectory

This subdirectory archives the repeat masker BLAST. Only two releases, 2.2.27 and 2.2.28, are available. Refer to the readme for more details.

Contents of the "/blast/matrices/" subdirectory

This directory contains an extensive list of score matrices, most of which are experimental in nature and not supported by blast programs. The matrices can be grouped into PAM family of matrices, BLOSUM family of matrices, matrices for nucleotide and other miscellaneous matrices.

Table 7. Summary of matrices found in the /blast/matrices/ subdirectory

File	Contents
BLOSUM*	BLOSUM family of score matrices for protein alignment
PAM*	PAM family of score matrices for protein alignment
DAYHOFF*, GONNET*	Variants of PAM matrices for protein alignment
MATCH, IDENTITY	Simplified score matrices for protein alignment
NUC*	Nucleotide score matrices
PAM.tar.gz	C source code for generating PAM matrices

Contents of the "/blast/temp/" subdirectory

This is a directory used for testing purposes or other special needs that do not fall into the above categories.

Contents of the "/blast/WGS_TOOLS/" subdirectory

This directory provides two database alias generating tools for WGS and TSA datasets, respectively. These tools take an input taxonomic id and generate a database alias for project-based WGS or TSA datasets - those with four letter project prefix as listed at https://www.ncbi.nlm.nih.gov/Traces/wgs/. The alias file produced can be used with the vdb blast tools available from the sratoolkit, which is available for common platforms: https://www.ncbi.nlm.nih.gov/Traces/sra/?view=software. A handout describing the general usage of these vdb blast programs is available at:

ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_Local_SRA_BLAST.pdf

Getting Help

For details, please refer to documents under the Help tab of the BLAST homepage or the document directory under the BLAST FTP site.

Comments, questions and bug reports specifically relating to the BLAST programs and their usage should be sent to blast-help@ncbi.nlm.nih.gov.