



## Standalone BLAST Setup for Unix

**Tao Tao, PhD**

NCBI

Tao@ncbi.nlm.nih.gov

### Introduction

NCBI provides command line based standalone BLAST programs based on the NCBI C++ toolkit as a single compressed package. The package, blast+, is available as ncbi-blast-initialed archives for a variety of computer platforms (hardware/operating system combinations) at:

`ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/`

The archives for UNIX, Linux and MacOSX are gzip-compressed tar files named using the following convention:

`ncbi-blast-#.##.##-CHIP-OS.tar.gz`

Here #.#.# represents the version number of the current release, while CHIP indicates the chipset and OS indicates the operating system. The archives and their target platforms are listed in the table below.

Table 1. Executable blast+ package available from NCBI			
Archive Name	Chipset	OS	File Type
ncbi-blast-2.2.23+-ia32-linux.tar.gz	Pentium chip	Linux, 32 bit	gzip'd tar archive
ncbi-blast-2.2.23+-sparc64-solaris.tar.gz	Sparc64	Solaris 10	gzip'd tar archive
ncbi-blast-2.2.23+-universal-macosx.tar.gz	ppc/intel	Mac OS X	gzip'd tar archive
ncbi-blast-2.2.23+-x64-linux.tar.gz	X64 chip	Linux, 64 bit	gzip'd tar archive
ncbi-blast-2.2.23+-x64-solaris.tar.gz	X64 chip	Solaris 10	gzip'd tar archive
ncbi-blast-2.2.23+.dmg	ppc/intel	Mac OS X	gzip'd disk image
ncbi-blast-2.2.23+-1.i686.rpm	Pentium chip	Linux, 32 bit	rpm
ncbi-blast-2.2.23+-1.x86_64.rpm	X64 chip	Linux, 64 bit	Rpm
Note: rpsblast databases are platform dependent.			

Installation process from the disk image for Mac OS X and the rpm package for Linux are different and will not be discussed here. Also, the BLAST packages based on NCBI C toolkit (legacy blast) are deprecated and its installation will be described briefly at the end of this tutorial.

### Downloading

The blast+ packages for various platforms should be downloaded through anonymous ftp using an ftp client, or other tools such as a web browser, wget, curl, etc. The example working session below demonstrates an ftp downloading process using the traditional ftp client under a Linux environment. Under Mac OSX, similar command line interface from the backend BSD Unix

is accessible by launching the Terminal. The Terminal program is generally under the Utilities folder.

## Steps

Steps to download the package are described below.

- Point a browser to this ftp directory:  
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/
- Right click on a desired archive and select "Save link as..." from the popup menu
- In the prompt, switch to a desired directory (folder) and click the "Save" button to save the archive to a desired location on the local disk

The above steps, as sample screenshots for the "ncbi-blast-2.2.23-win32.exe" archive, are given in Figure 2, where the first two steps are demonstrated by the top panel and the last step is demonstrated by the bottom panel.

```
$ ftp ftp.ncbi.nlm.nih.gov
Connected to ftp.wip.ncbi.nlm.nih.gov.
220-
Warning Notice!
This is a U.S. Government computer system, which may be accessed and used
[ ... extra warning message removed ... ]
There is no right of privacy in this system.
---
Welcome to the NCBI ftp server! The official anonymous access URL is ftp://
ftp.ncbi.nih.gov
Public data may be downloaded by logging in as "anonymous" using your E-mail
address as a password.
Please see ftp://ftp.ncbi.nih.gov/README.ftp for hints on large file
transfers
220 FTP Server ready.
Name (ftp.ncbi.nlm.nih.gov:tao): anonymous
331 Anonymous login ok, send your complete email address as your password.
Password: [note: enter your email address at this prompt]
230-Anonymous access granted, restrictions apply.
Please read the file README.ftp
230 it was last modified on Fri Mar 28 14:05:45 2008 - 716 days ago
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> cd blast/executables/blast+/LATEST/
250 CWD command successful
ftp> bin
200 Type set to I
ftp> get ncbi-blast-2.2.23+-ia32-linux.tar.gz
local: ncbi-blast-2.2.23+-ia32-linux.tar.gz remote: ncbi-blast-2.2.23+-ia32-
linux.tar.gz
229 Entering Extended Passive Mode (|||50335|)
150 Opening BINARY mode data connection for ncbi-blast-2.2.23+-ia32-
linux.tar.gz
(197354741 bytes)
100% |*****| 188 MB 35.71 MB/s
00:00 ETA
```

```
226 Transfer complete.
197354741 bytes received in 00:05 (35.70 MB/s)
ftp> bye
221 Goodbye.
$
```

To do BLAST searches on platforms lacking precompiled blast+ package will require compiling the BLAST source code. The source code file, "ncbi-blast-#.#.#-src" in either zip or gzipped tar format, is available from the same ftp directory as blast+ packages. Questions and feedbacks on source code compilation should be addressed to:

```
toolbox@ncbi.nlm.nih.gov
```

## Installation

To install, simply extract the downloaded package after placing it under a desired directory. This can be accomplished by a single tar command, or a combination of gunzip and tar commands.

```
tao@gizmo:/home/tao$ tar zxvpf ncbi-blast-2.2.23-x64-linux.tar.gz
```

or

```
tao@gizmo:/home/tao $ gunzip -dtar zxvpf ncbi-blast-2.2.23-x64-linux.tar.gz
tao@gizmo:/home/tao $ tar xvpf ncbi-blast-2.2.23-x64-linux.tar
```

Successful execution of the above commands installs the package and generates a new blast-2.2.23+ directory under "/home/tao" with bin and doc subdirectories, as well as a VERSION file. The bin subdirectory contains the programs listed below.

Table 3.1 Programs contained in blast+ package	
Program	Function
blastdbcheck	Checks database integrity
blastdbcmd	Retrieves sequences or other information from a BLAST database
blastdb_aliastool	Creates database alias
Blastn	Searches a nucleotide query against a nucleotide database
blastp	Searches a protein query against a protein database
blastx	Searches a nucleotide query, dynamically translated in all six frames, against a protein database
blast_formatter	Formats a web blast result using its assigned request ID (RID)
convert2blastmask	Converts lowercase masking into makeblastdb readable data
dustmasker	Masks the low complexity regions in the input nucleotide sequences
legacy_blast.pl	Converts a legacy blast search command line into blast+ counterpart and execute it
makeblastdb	Formats input FASTA file(s) into a BLAST database
makembindex	Indexes an existing nucleotide database for use with megablast

psiblast	Finds members of a protein family, identifies proteins distantly related to the query, or builds position specific scoring matrix for the query
rpsblast	Searches a protein against a conserved domain database (CDD) to identify functional domains present in the query
rpstblastn	Searches a nucleotide query, by dynamically translated it in all six-frames first, against a conserved domain database (CDD)
segmasker	Masks the low complexity regions in input protein sequences
tblastn	Searches a protein query against a nucleotide database dynamically translated in all six frames
tblastx	Searches a nucleotide query, dynamically translated in all six frames, against a nucleotide database similarly translated
update_blastdb.pl	Downloads preformatted blast databases from NCBI
windowmasker	Masks repeats found in input nucleotide sequences

## Configuration

Using the blast+ package installed above without configuration will be cumbersome - extraneous path prefixing to the program call and database specification will be needed since 1) the system does not know where to look for the installed BLAST programs, and 2) BLAST programs do not know which directory to search for the BLAST database. To smooth the execution of BLAST searches, two environment variables, PATH and BLASTDB, need to be modified and specified, respectively, to point to the corresponding directories.

Avoiding prefixing a path to the bin directory in every BLAST program calls requires modification of the existing \$PATH variable using the following command under bash, which appends the path to the new BLAST bin directory:

```
tao@gizmo:/home/tao$ PATH=/home/tao/blast-2.2.23+/bin
tao@gizmo:/home/tao$ export PATH
```

The equivalent command under csh is:

```
tao@gizmo:/home/tao$ setenv PATH ${PATH}:/home/tao/blast-2.2.23+/bin
```

The modified \$PATH can be examined using echo:

```
tao@gizmo:/home/tao$ echo $PATH
/usr/X11R6/bin:/usr/bin:/bin:/usr/local/bin:/opt/local/bin:/home/tao/blast-2.2.23+/bin
```

A better approach is to modify the login script to have the system do this automatically.

For the management of BLAST databases, a subdirectory named db should be created. For the above installation, the following command creates such directory under /blast-2.2.23+:

```
tao@gizmo:/home/tao$ mkdir ./blast-2.2.23+/db
```

A ".ncbirc" file under the home directory is needed to instruct BLAST programs to check this db directory to locate the specified target database. This file can be created using any text editor (jpic, nano or nedit) and it should have the following path specification.

```
[BLAST]
BLASTDB=/home/tao/blast-2.2.23+/db
```

Upon start, BLAST will read this file to get the path information it needs during BLAST searches. Without this file, BLAST will search the working directory, or whichever directory the command is issued from.

## Database Download

BLAST database is a key component of any BLAST search. To fully test the blast+ package thus installed, a functional database is needed. The following work session demonstrate the process of downloading and installation of the refseq\_rna.tar.gz, a pre-formatted BLAST database from NCBI.

```
tao@gizmo:/home/tao$ cd blast-2.2.23+/db
tao@gizmo:/home/tao$ ftp ftp.ncbi.nlm.nih.gov
Connected to ftp.wip.ncbi.nlm.nih.gov.
220-
Warning Notice!
[ ... Extra warning message removed for brevity ... ]
230-Anonymous access granted, restrictions apply.
Please read the file README.ftp
230 it was last modified on Fri Mar 28 14:05:45 2008 - 740 days ago
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> cd blast/db
250 CWD command successful
ftp> bin
200 Type set to I
ftp> get refseq_rna.tar.gz
local: refseq_rna.tar.gz remote: refseq_rna.tar.gz
229 Entering Extended Passive Mode (|||50279|)
150 Opening BINARY mode data connection for refseq_rna.tar.gz (857150245
bytes)
100% |*****| 817 MB 21.48 MB/s 00:00 ETA
226 Transfer complete.
857150245 bytes received in 00:38 (21.48 MB/s)
ftp> bye
221 Goodbye.
tao@gizmo:/home/tao/blast-2.2.23+/db$
```

Inflating the compressed archive and extracting the tar file will regenerate the files for this database. The example tar command and its output are given below. To save disk space, the refseq\_rna.tar.gz file can be removed after the installation.

```
tao@gizmo:/home/tao/blast-2.2.23+/db$ tar zxvpf refseq_rna.tar.gz
refseq_rna.nhr
refseq_rna.nin
refseq_rna.nnd
refseq_rna.nni
refseq_rna.nsd
refseq_rna.nsi
```

```

refseq_rna.nsq
tao@gizmo:/home/tao/blast-2.2.23+/db$ ls -l refseq_rna*
total 2101428
-rw-rw-r-- 1 tao sdesk 320523184 Apr 6 20:42 refseq_rna.nhr
-rw-rw-r-- 1 tao sdesk 25279444 Apr 6 20:42 refseq_rna.nin
-rw-rw-r-- 1 tao sdesk 16852896 Apr 6 20:42 refseq_rna.nnd
-rw-rw-r-- 1 tao sdesk 65876 Apr 6 20:42 refseq_rna.nni
-rw-rw-r-- 1 tao sdesk 87565216 Apr 6 20:42 refseq_rna.nsd
-rw-rw-r-- 1 tao sdesk 1829004 Apr 6 20:42 refseq_rna.nsi
-rw-rw-r-- 1 tao sdesk 840073418 Apr 6 20:42 refseq_rna.nsq
-rw-r--r-- 1 tao sdesk 857150245 Apr 6 21:13 refseq_rna.tar.gz
tao@gizmo:/home/tao/blast-2.2.23+/db$ rm refseq_rna.tar.gz
tao@gizmo:/home/tao/blast-2.2.23+/db$

```

The same procedure can be used to download other BLAST databases from NCBI. For regular batch database download/update, take advantage of the `update_blastdb.pl` script included in the package. This script can automatically download all the volumes of a large database.

## Execution and validation

With the above blast+ setup, BLAST programs installed under the "blast-2.2.23+/bin" directory can be invoked by name from any directory. In the example below, "blastn -help", invoked under the /home/tao directory, displays the program parameters of blastn to the console.

```

tao@gizmo:/home/tao> blastn -help
USAGE
blastn [-h] [-help] [-import_search_strategy filename]
[-export_search_strategy filename] [-task task_name] [-db database_name]
[-dbsize num_letters] [-gilist filename] [-negative_gilist filename]
[-entrez_query entrez_query] [-db_soft_mask filtering_algorithm]
[-subject subject_input_file] [-subject_loc range] [-query input_file]
[-out output_file] [-evaluate evaluate] [-word_size int_value]
[-gapopen open_penalty] [-gapextend extend_penalty]
[-perc_identity float_value] [-xdrop_ungap float_value]
[-xdrop_gap float_value] [-xdrop_gap_final float_value]
[-searchsp int_value] [-penalty penalty] [-reward reward] [-no_greedy]
[-min_raw_gapped_score int_value] [-template_type type]
[-template_length int_value] [-dust DUST_options]
[-filtering_db filtering_database]
[-window_masker_taxid window_masker_taxid]
[-window_masker_db window_masker_db] [-soft_masking soft_masking]
[-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
[-best_hit_score_edge float_value] [-window_size int_value]
[-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
[-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
[-outfmt format] [-show_gis] [-num_descriptions int_value]
[-num_alignments int_value] [-html] [-max_target_seqs num_sequences]
[-num_threads int_value] [-remote] [-version]

DESCRIPTION
Nucleotide-Nucleotide BLAST 2.2.23+

```

```

OPTIONAL ARGUMENTS
-h
Print USAGE and DESCRIPTION; ignore other arguments
-help
Print USAGE, DESCRIPTION and ARGUMENTS description; ignore other arguments
-version
Print version number; ignore other arguments

*** Input query options
-query <File_In>
Input file name
Default = '-'
-query_loc <String>
Location on the query sequence (Format: start-stop)
-strand <String, 'both', 'minus', 'plus'>
Query strand(s) to search against database/subject
Default = 'both'
...

```

Note: For installation without \$PATH modification, prefix the path to the program. For example, to execute the same command from /home/tao directory, use the following command instead, where the "." prefix denotes the current working directory:

```
tao@gizmo:/home/tao$ ./blast-2.2.23/bin/blastn -help
```

The following set of commands test the above installation by extracting a sequence from the installed refseq\_rna database and using this extracted sequence as a query in a blastn search. The search asks for two database hits returned in tabular format.

### Example Execution

In the command prompt, the working directory can be changed to "C:\blast-2.2.23+" by typing "cd\" followed by "cd blast-2.2.23+". If the first prompt is a drive other than "C:\", "C:" instead of "cd\" should be used first. Figure 5.2 contains example commands and their console output for a work session that tests a blast-2.3.23+ installation.

```

tao@gizmo:/home/tao$ blastdbcmd -db refseq_rna -entry nm_000249 -out
test_query.fa
tao@gizmo:/home/tao$ blastn -query test_query.fa -db refseq_rna -task blastn
-dust no -outfmt 7
-num_alignments 2 -num_descriptions 2
# BLASTN 2.2.23+
# Query: gi|263191547|ref|NM_000249.3| Homo sapiens mutL homolog 1, colon
cancer,
nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA
# Database: refseq_rna
# Fields: query id, subject id, % identity, alignment length,
mismatches, gap opens,

```

```

q. start, q. end, s. start, s. end, evaluate, bit score
# 2 hits found
gi|263191547|ref|NM_000249.3| gi|263191547|ref|NM_000249.3| 100.00
2662 0 0 1 2662 1 2662 0.0 4801
gi|263191547|ref|NM_000249.3| gi|114585959|ref|XM_001170433.1| 99.59
2666 7 1 1 2662 410 3075 0.0 4758
# BLAST processed 1 queries
tao@gizmo:/home/tao$

```

Note that the command lines and output wrap around.

## Setup Steps For Legacy blast

The original standalone BLAST package based on NCBI C-toolkit (legacy blast) is deprecated. The installation of legacy blast package for Windows differs from that for blast+ described above. The key differences are summarized below.

- a The legacy blast packages are located under a different ftp directory:

```
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/
```

- b The packages are named with this convention: blast-#.#.#-CHIP-OS.tar.gz, where #.#.# is the version, CHIP is the chipset, and OS is the operating system
- c The program names and functions are different (see Table 7 below for details)
- d Path to the extra data directory need to be specified in .ncbirc in this format:

Path to the extra data subdirectory needs to be specified in .ncbirc in this format

[NCBI]

```
DATA=/path/blast-#.#.#/data
```

Table 7. Programs contained in the legacy blast package	
Program	Function
bl2seq <sup>1</sup>	Directly comparing two FASTA sequences
blastall <sup>1</sup>	legacy blast containing the subfunction of blastn, blastp, blastx, tblastn, and tblastx
blastclust <sup>2</sup>	Clusters input FASTA sequences into related groups
blastpgp <sup>1</sup>	Standalone PSI-BLAST for search of distantly related protein sequences and generate position-specific matrices
copymat <sup>2</sup>	Copies blastpgp output for input to makemat
fastacmd <sup>1</sup>	Retrieves specific sequence or dumps the sequences from a formatted blast database
formatdb <sup>1</sup>	Convert FASTA formatted sequecne file into BLAST database
formatpsdb <sup>2</sup>	Format scoremat files into an RPSBLAST database
impala <sup>2</sup>	protein profile search program, mostly replaced by rpsblast
makemat <sup>2</sup>	Convert the copymat files into scoremat format, no longer needed by new blastpgp output
megablast <sup>1</sup>	Faster batch blastn program that uses greedy-algorithm. Works in contiguous or more sensitive discontinuous mode
rpsblast <sup>1</sup>	reverse PSI-BLAST program for searching against conserved domain database
seedtop <sup>2</sup>	Pattern search program

Note:  
<sup>1</sup> Those programs are re-organized into blastn, blastp, blastx, tblastn, tblastx, rpsblast, rpsblastx, psiblast, blastdbcmd and makeblastdb  
<sup>2</sup> Those programs have no blast+ counterpart at this time.

The commands for legacy blast, comparable to those given for blast+ in section 6, are:

```
blastall -
```

```
fastacmd -d refseq_rna -s nm_000249 -o test_query.fa
```

```
blastall -p blastn -i test_query.fa -d refseq_rna -F F -m 9 -b 2 -v 2
```

## Technical Assistance

Questions, feedbacks, and technical assistance requests should be sent to blast-help at:

```
blast-help@ncbi.nlm.nih.gov
```

Questions on other NCBI resources should be addressed to NCBI Service Desk at:

```
info@ncbi.nlm.nih.gov
```