

BLAST+ now has a better BLAST database.



This document concerns some new BLAST features available in stand-alone BLAST. These changes do not affect the BLAST web interface.

Recent enhancements to BLAST+

We have made some recent enhancements to the BLAST+ applications that:

- 1.) Allow you to limit your search by taxonomy using information built into the BLAST databases.
- 2.) Have improved performance when limiting searches with an accession list.
- 3.) Can retrieve sequences by taxonomy from a BLAST database with `blastdbcmd`.

The new version of the BLAST databases (version 5) that supports the items listed above is now the default on the BLAST FTP site. We recommend use of BLAST+ 2.10.0 or later to take advantage of these new features.

We have also made some changes to the databases available on the FTP site. In addition to `nr`, `nt`, `swissprot`, etc, it now includes databases of high quality representative genomic sequences as well as databases based on [Targeted Loci](#) projects.

This document provides information about using these new databases. First, we provide information on how to access the latest BLAST executables and the version 5 databases from the BLAST FTP site. Second, we provide some examples that demonstrate new features associated with the version 5 databases. Finally, we demonstrate how quickly these searches run.

Executables and Databases

You can download the latest version of the BLAST+ executables at <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

You can find version 5 databases at <https://ftp.ncbi.nlm.nih.gov/blast/db> A description of the databases available on the FTP site can be found at <https://ftp.ncbi.nlm.nih.gov/blast/db/README>

Use `update_blastdb.pl` (included with the BLAST+ package) to download the version 5 databases, which are now the default:

```
>update_blastdb.pl --showall
Connected to NCBI
16S_ribosomal_RNA
18S_fungal_sequences
28S_fungal_sequences
Betacoronavirus
ITS_RefSeq_Fungi
ITS_eukaryote_sequences
LSU_eukaryote_rRNA
LSU_prokaryote_rRNA
SSU_eukaryote_rRNA
landmark
nr
nt
patnt
pdbaa
pdbnt
ref_euk_rep_genomes
ref_prok_rep_genomes
ref_viroids_rep_genomes
ref_viruses_rep_genomes
refseq_protein
refseq_rna
swissprot
taxdb

>update_blastdb.pl landmark --decompress
Connected to NCBI
Downloading landmark.tar.gz... [OK]
Decompressing landmark.tar.gz ... [OK]
```

The version 4 databases are no longer being updated, and we recommend moving to the new BLAST+ executables and version 5. A static copy of the last set of version 4 databases is at <https://ftp.ncbi.nlm.nih.gov/blast/db/v4/>

You will also need to install the EDirect command-line utility if you wish to look up TAXIDs via the command-line script, `get_species_taxids.sh` (see below). Instructions for installing EDirect are available at <https://www.ncbi.nlm.nih.gov/books/NBK179288/>. The shell script we provide will execute the actual EDirect commands. EDirect is only available for MacOSX and LINUX.

Examples

Limiting a search by taxonomy

In order to limit your BLAST+ search by taxonomy, you'll need to obtain the taxid(s) for your organism(s). A taxid is simply a number that specifies a node in the taxonomic tree. For example, 9606 is the taxid for human, 9989 is the taxid for rodentia, and 2 is the taxid for all bacteria. Taxids are preferable to organism names as the latter can be ambiguous. For example, bacteria is both a genus of insect as well as a superkingdom. BLAST will only accept taxids that are at or below the species level.

We provide a script to translate higher level taxids (e.g., Enterobacteriales) into a list of taxids that are at the appropriate level (details below). The same script can also be used to lookup (and disambiguate) taxids based upon a taxonomic name. The script is called `get_species_taxids.sh` and is part of the BLAST+ package. As noted above, you will also need to install the EDirect command-line utility if you wish to use `get_species_taxids.sh`. Instructions are at <https://www.ncbi.nlm.nih.gov/books/NBK179288/>.

You can limit your stand-alone BLAST+ search by specifying one or more (comma-delimited) taxids on the command-line, or you can specify a file containing multiple taxids. Some example command-lines:

Running a BLAST search with a higher level taxonomic node (Enterobacteriales):

```
get_species_taxids.sh -n Enterobacteriales

Taxid: 91347
rank: order
division: enterobacteria
scientific name: Enterobacteriales
common name:

1 matches found

get_species_taxids.sh -t 91347 > 91347.txids

blastn -db nt -query QUERY -taxidlist 91347.txids -outfmt 7 -out OUTPUT.tab
```



This example uses the `-taxidlist` parameter which takes a file as input.

Running a BLAST search with a species level taxid, human (taxid 9606):

```
blastn -db nt -query QUERY -taxids 9606 -outfmt 7 -out OUTPUT.tab
```



If you are not sure whether your taxid is at the species level or lower (or covers all such cases), it is safe to run `get_species_taxids.sh` and feed the output to BLAST.

Additionally, you may use the `-negative_taxids` and `-negative_taxidlist` options to exclude sequences by TAXID from your search.

Limiting a search by a list of accessions

You may use a list of accessions to limit a search with both the version 4 and version 5 databases. With the version 5 databases, we recommend preprocessing the the accession list. This process checks that the accessions appear to be real and produces a file optimized for use with BLAST. It is also possible to confirm that all the accessions are actually in your target database. The command lines below demonstrate the commands you will need.

```
blastdb_aliastool -seqid_file_in 9606.pacc          # 9606.pacc is a text file with protein accessions. This
command produces a file called 9606.pacc.bsl
blastp -db nr -query QUERY -outfmt "7 std staxid" -seqidlist 9606.pacc.bsl      # This command searches nr
limited to the accessions in the file 9606.pacc.bsl
```

Additionally, you may use the `-negative_seqidlist` option to exclude sequences by accession from your search.

Faster sequence lookups by accession

The version 5 databases use LMDB ([Lightning Memory-Mapped Database](#)) to quickly retrieve sequences by accession. You can still use `blastdbcmd` for the retrieval and the old parameters are still supported (examples below). There are two new parameters (`-taxids` and `-taxidlist`) to retrieve sequences by taxid, and the next example demonstrates the usage of the `-taxids` parameter.

```
blastdbcmd -db nt -entry u0001
blastdbcmd -db nt -entry_batch FILE_WITH_ACCESSIONS
blastdbcmd -db nr -taxids 9606 -outfmt "%a %T %S" -target_only          # retrieves all human entries
# %a prints the accession, %T prints the taxid, %S prints the scientific name
```



`-target_only` is used with the last `blastdbcmd` command to ensure that only accessions for human entries are presented. Otherwise, it will present all accessions on any sequence with at least one human accession. This is important since `nr` is a non-redundant database.

Speed of searches

We search AAC51230 (human MEN1 protein) against `nr` restricted to human proteins by various means and once without any restriction. All searches were run three times in a row and the lowest time was selected.

| Restrict nr by... | Runtime (seconds) |
|----------------------------|-------------------|
| TaxID | 11.7 |
| GI list | 13.2 |
| Accession list (version 5) | 12.1 |
| Accession list (version 4) | 18.6 |
| No restriction | 2,865.0 |

Problems/Feedback

Please send problems reports or feedback to blast-help@ncbi.nlm.nih.gov or [BLAST support](#).