

**Title:** blast sequences against a local version of BLAST  
Nadim Alkharouf, Ph.D.

## **Introduction**

BLAST (Basic Alignment Search Tool), allows rapid comparison of a sequence against a database of sequences. BLAST is fundamental to understanding how two sequences are related to one another. It is used to find homologs, discover coding regions (exons) in genomic DNA sequences, discover new genes or proteins, discover variants of genes or proteins, and also to explore protein structure.

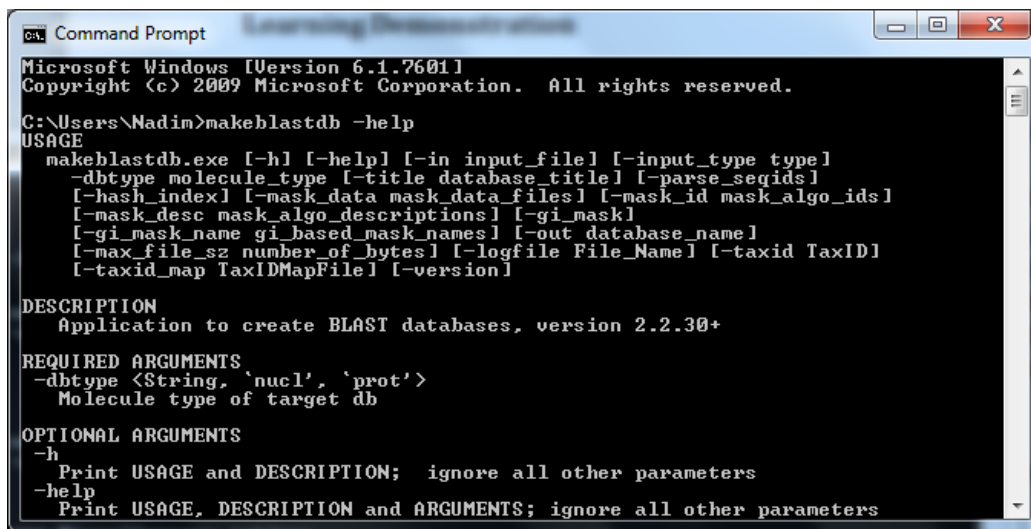
BLAST is one of the most widely used tools in molecular biology. You probably only used the web portal version however: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, which is very efficient and informative, but it does have limitations. For instance if one had 1000's of sequences to BLAST it would be almost impossible to use the web portal. That is why NCBI also provides a standalone version of BLAST, it includes all the executables for each type of BLAST.

In this learning demonstration we will: 1- Download and install the standalone version of BLAST and learn how to use it. 2- Learn how to format a sequence database to blast against. And 3- Write a Python script to call the BLAST executables and parse out the results.

## **Learning Demonstration**

- 1- Point your browser to:  
[http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=Download](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download) and download the latest standalone BLAST version.
- 2- Install BLAST by clicking on the downloaded file. Instructions to install BLAST can be found here for Windows PCs: <http://www.ncbi.nlm.nih.gov/books/NBK52637/>, or here for Mac/Linux: <http://www.ncbi.nlm.nih.gov/books/NBK52640/>
- 3- BLAST should install and create the necessary environment paths for you. In Windows BLAST is installed in C:\Program Files\NCBI\blast-#.#.##+\bin
- 4- Test the installation. Open up a command prompt window and type:

```
makeblastdb -help
```



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Nadin>makeblastdb -help
USAGE
    makeblastdb.exe [-h] [-help] [-in input_file] [-input_type type]
    -dbtype molecule_type [-title database_title] [-parse_segids]
    [-hash_index] [-mask_data mask_data_files] [-mask_id mask_algo_ids]
    [-mask_desc mask_algo_descriptions] [-gi_mask]
    [-gi_mask_name gi_based_mask_names] [-out database_name]
    [-max_file_sz number_of_bytes] [-logfile File_Name] [-taxid TaxID]
    [-taxid_map TaxIDMapFile] [-version]

DESCRIPTION
    Application to create BLAST databases, version 2.2.30+

REQUIRED ARGUMENTS
    -dbtype <String, 'nucl', 'prot'>
        Molecule type of target db

OPTIONAL ARGUMENTS
    -h
        Print USAGE and DESCRIPTION; ignore all other parameters
    -help
        Print USAGE, DESCRIPTION and ARGUMENTS; ignore all other parameters
```

If you see the options\argument list displayed, you know it's been installed and ready to be used. See snap shot above.

- 5- You can download a BLAST database to align your sequences against. They are available from this ftp site: <ftp://ftp.ncbi.nlm.nih.gov/blast/db/>. The databases include the nr database, swissprot, and sequence databases for human, mouse, rat...etc. These databases are already formatted\indexed for local BLAST.
- 6- If you have your own set of sequences that you want to align against, BLAST gives you the option of creating your very own sequence database.
- 7- Included with this learning demonstration are two files, one called strawberry.fa, which is a collection of nucleotide sequences from strawberry, and the other called testSequences.fasta. These are merely text files; you can open them in any text editor and view their content. We will create a sequence database from strawberry.fa and then use that to blast our test sequences against.
- 8- Create a folder on your computer and place the two download files above inside it, for instance you can create a folder called LD3 and place it on your C drive.
- 9- Using the command prompt window you opened earlier, change directory to that folder:

```
cd C:\LD3
```

- 10- You can type the command *dir* in Windows or *ls* in Linux to view the list of files in that folder, you should see the file names mentioned above
- 11- To make a BLAST database out of strawberry.fa just type the command:

```
makeblastdb.exe -dbtype nucl -in strawberry.fa
```

The argument *-dbtype* specifies what type of sequences are going to be in the database, in this case it's a nucleotide (nucl) database. And the argument *-in* specifies the input file that contains the FASTA sequences, in this case its strawberry.fa.

If you look into your LD3 folder now you'll see additional files that were created, these are index database files that BLAST uses, all start with the name strawberry.fa

12- Once we create the database we are ready to blast\align sequences against it. BLAST comes in many different types, the main three are: *blastn* for nucleotide blast, *blastp* for protein blast, and *blastx* to search a nucleotide sequence against a protein database. A description of each type of blast and when to use each can be found here: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. All types of BLAST are included in the standalone version as .exe files.

13- Let's assume we want to find homologs for our sequences in *testSequences.fasta* and so will use *blastn*; because our query sequences are nucleotide and our database is also a nucleotide database.

14- *Blastn.exe* is the executable to run a *blastn* search in the standalone BLAST. To see the different options available for *blastn.exe* just type the command:

```
blastn.exe -help
```

15- You will see a LOT of options! These include options for e-value cut off, word size, matrix choice...etc. But the main options\arguments are: *-query* which specifies the sequence file you want to BLAST, *-db* which specifies the database you want to BLAST against, *-out* the name of the output file (you choose whatever name you like) and *-outfmt* for output format (choices include pairwise alignment formats to tabular formats).

16- Run this command in your command prompt window:

```
blastn.exe -query testSequences.fasta -db strawberry.fa -  
out blastOutput.txt -outfmt 0
```

17- The output file *blastOutput.txt* should have been created in your folder, open it in any text editor to view the results of the blast search.

18- Now that you've seen how easy it is to run your own BLAST searches, let's write a Python script to do the steps above and to also parse out the results for us.

19- Open up your favorite IDE or editor, and create a new Python source code file. Place the source code inside the folder with the BLAST database that you created above (or you will have to provide the full path to it in your code if it's somewhere else).

20- Use the *subprocess.call()* function, within the subprocess module – don't forget to *import subprocess*. We used this previously. We will use it here to create the database using Python (keep in mind that the newly created version will overwrite any older version).

- 21- Now use the `subprocess.call()` function to call `blastn.exe` and run BLAST like we did in the command prompt window above.
- 22- Save your code and run it. If it runs OK and no error messages were returned continue on to the next step.
- 23- The pairwise alignment format in the `blastOutput.txt` file is good, but is hard to read if you have 1000's of sequences in your report. Most biologists would prefer to look at a spreadsheet like file, that contains the query sequence\ID, it's best hit, the e-value, score and identities (for a description of what these mean see [blast documentation](#)). A tabular format like the one shown below is also easier to import and store in a relational database.
- 24- Use Python to parse out or convert the `blastOutput.txt` file so it's tabular in format, and looks like this (small sample of file for illustrative purposes only):

QueryID	SubjectID	Identstart	Identend	E-value	Score
s1	gene00001	159	159	3.00E-80	294
s2	gene00002	357	357	0	660
s3	gene00003	420	420	0	776
s4	gene00004	312	312	6.00E-165	577
s5	gene00007	126	126	5.00E-62	233
s6	gene00008	93	93	8.00E-44	172
s7	gene00009	192	192	2.00E-98	355
s8	gene00011	147	147	1.00E-73	272
s9	gene00012	213	213	4.00E-110	394
s10	gene00014	192	192	2.00E-98	355

- 25- You can either use the `-outfmt` options in BLAST to do this, or you can write your own regular expression. Try it both ways, using regex will give you lots of practice with how regex works.
- 26- We've already talked about the `re` module and the `search()` and `finditer()` functions. The trick here is to find the patterns and extract them. Play around with the regex till you get it right.
- 27- Let's modify the code to generalize it so that a user can specify what files he\she wants to use. Remember a good program is one that is flexible and anticipates user needs.
- 28- To that end, modify your source code now to ask the user for the sequence file that he\she would like to create a BLAST database from. Use statements to make sure the FASTA file exists.
- 29- Now ask the user for the FASTA file he\she wishes to BLAST against this database. Again test to see if the file name given to you exists in the source code's working directory (see notes on Files and Folder access under content for more info on this).

30- Save your code and test it on different files, and if possible on different systems as well.

31- Congratulations! You've just created an automated pipeline for a user to BLAST 1000's of his\her own sequences and retrieve the results in an easy to read report that he\she can open with Excel. That is what Bioinformaticians do day in and day out!

**Deliverables:**

Submit your complete source code file, with comments. Submit the Blast output file (pairwise alignment format), and the parsed output (tabular format) file.