

BLAST From the Command Line

Table of Contents

1.	<i>INTRODUCTION</i>	1
2.	<i>WHAT IS BLAST?</i>	2
3.	<i>WHY USE THE COMMAND LINE?</i>	2
4.	<i>INSTALLATION</i>	2
4.1	<i>MAC</i>	2
4.2	<i>UNIX</i>	4
5.	<i>THE SETS</i>	4
5.1	<i>WHERE TO FIND A SUBJECT SET</i>	4
5.2	<i>WHERE TO FIND A QUERY SET</i>	5
6.	<i>SIMPLE BLAST OF DOWNLOADED EXAMPLES</i>	5
6.1	<i>SIMPLE NUCLEOTIDE BLAST RESULTS</i>	6
7.	<i>FURTHER READING</i>	7
8.	<i>CITATIONS AND REFERENCES</i>	7
9.	<i>MSX1 CONTENTS</i>	8

1. INTRODUCTION

The intention of this manual is to introduce the reader to the concept of BLAST, specifically its use locally through a command line interface. Technical details of what BLAST is will be explored including some of its uses in life sciences. Several methods of BLAST will be explored before details of how to install it on different operating systems will be explained in detail.

The remainder of the manual will explain how to find sequences, how to use BLAST through the command line, and the different variables that can be used. The subject base used will be that of the stickleback. The query will be a gene the author is familiar with. The purpose of using this combination is to determine if the gene sequence is detected in the stickleback genome. The goal of running the BLAST on these data is to see if these gene sequences are detected in the stickleback genome and to calculate the similarity of the ortholog gene.

This manual could not possibly be considered an exhaustive lesson on all of the functionality of a tool such as BLAST. This manual will wrap up with an explanation of some of the most used optimizations as well as selected further reading and finally the references used in this manual.

2. WHAT IS BLAST?

BLAST stands for Basic Local Alignment Search Tool¹ and is used to find similar sequences in two or more sequences. The sequences can be that of nucleotide or primary protein sequence. The sequences themselves can be anything from a small primer to an entire genome. The smaller the query however the more places it will most likely align to on the subject. BLAST is often a first tool used to analyze an unknown sequence.

3. WHY USE THE COMMAND LINE?

This manual will instruct how to install and use BLAST on a local machine. BLAST is available on NCBI online as well as many other websites. The question may arise about why a scientist would want to use the terminal when so many online resources are available.

Running BLAST locally is a valuable resource for reproducible science. A BLAST script can be shared or even provided in a publication that can be followed and analyzed by other users. A library can be created by the user that may not be published in the NCBI database. Private libraries of unpublished genomes or orthologous genes can be used to align locally. Using BLAST on a local machine also allows the user to process larger jobs much quicker than online BLAST could facilitate.

4. INSTALLATION

Installing BLAST for command line is simple but slightly different depending on which operating system is used. Mac and UNIX are similar. Windows does not ship with a command line interface so one will need to be installed.

4.1 MAC

Open a terminal by using spotlight search (⌘+Space) and typing terminal before hitting enter. Confirm the terminal is in a directory that BLAST will be installed in. For the purpose of this manual BLAST will be installed at ~/BLAST.

```
cd ~
mkdir BLAST
cd ~/BLAST
```

Use curl to download BLAST to the installation directory

```
curl ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.9.0/ncbi-blast-2.9.0+-x64-macosx.tar.gz --output ncbi-blast.tar.gz
```

A terminal window with a dark background and light green text. The prompt is a green arrow pointing to a tilde (~). The user enters 'mkdir BLAST', then 'cd ~/BLAST', and finally 'curl ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.9.0/ncbi-blast-2.9.0+-x64-macosx.tar.gz --output ncbi-blast.tar.gz'. The output shows a progress bar for the download, with a green arrow pointing to the word 'BLAST' at the bottom.

The command is broken down into four parts.

- `curl` – curl is software that refers to Client URL and can be used to download anything that has a URL attached to it. (Jones, 2009)

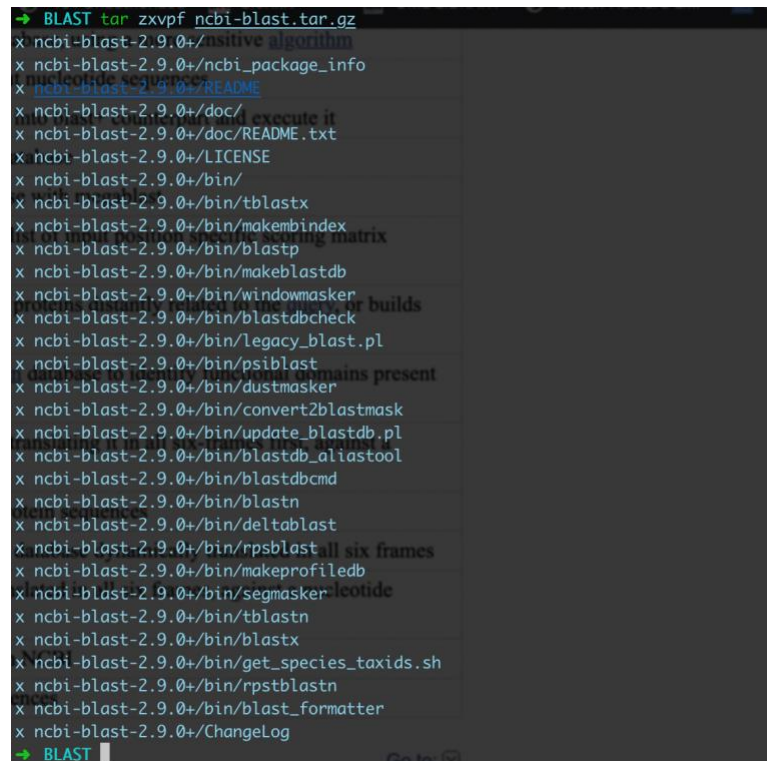
¹https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

- `ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.9.0/ncbi-blast-2.9.0+-x64-macosx.tar.gz` – This is the location on the internet that the file is located.
- `--output` – This is called a shell redirect. Its purpose is to point to where the file should be saved.
- `ncbi-blast.tar.gz` – This is the name and location the file will be saved as. In this example no directory is given so the file is saved at the current working directory which was set to `~/BLAST` in the first step.

Once downloaded the software needs to be unpacked using the tar command.

```
tar zxvpf ncbi-blast.tar.gz
```

- `tar` – Tar is a software package for Tape Archive although tape isn't used anymore to archive files the software is still very valuable for handling archive data and for transferring file systems. This file is archived with tar as well as gunzipped
- `zxvpf` – These are called options and there are more than just these five. These five however stand for the following options.
 - `z` – sends the tar file through gzip first
 - `x` – extracts the folder
 - `v` – shows a progress bar if the function is going to take a long time
 - `p` – keep permissions of archive
 - `f` – filename of the archive
 There are many other options that can be used²
- `ncbi-blast.tar.gz` – This is the name and location of the tar file.



```

→ BLAST tar zxvpf ncbi-blast.tar.gz
x ncbi-blast-2.9.0+/nsitive algorithm
x ncbi-blast-2.9.0+/ncbi_package_info
x ncbi-blast-2.9.0+/README
x ncbi-blast-2.9.0+/doc/
x ncbi-blast-2.9.0+/doc/execute it
x ncbi-blast-2.9.0+/doc/README.txt
x ncbi-blast-2.9.0+/LICENSE
x ncbi-blast-2.9.0+/bin/
x ncbi-blast-2.9.0+/bin/tblastx
x ncbi-blast-2.9.0+/bin/makemindex
x ncbi-blast-2.9.0+/bin/blastp
x ncbi-blast-2.9.0+/bin/makeblastdb
x ncbi-blast-2.9.0+/bin/windowmasker
x ncbi-blast-2.9.0+/bin/blastdbcheck
x ncbi-blast-2.9.0+/bin/legacy_blast.pl
x ncbi-blast-2.9.0+/bin/psiblast
x ncbi-blast-2.9.0+/bin/dustmasker
x ncbi-blast-2.9.0+/bin/convert2blastmask
x ncbi-blast-2.9.0+/bin/update_blastdb.pl
x ncbi-blast-2.9.0+/bin/blastdb_aliastool
x ncbi-blast-2.9.0+/bin/blastdbcmd
x ncbi-blast-2.9.0+/bin/blastn
x ncbi-blast-2.9.0+/bin/deltablast
x ncbi-blast-2.9.0+/bin/rpsblast
x ncbi-blast-2.9.0+/bin/makeprofiledb
x ncbi-blast-2.9.0+/bin/segmasker
x ncbi-blast-2.9.0+/bin/tblastn
x ncbi-blast-2.9.0+/bin/blastx
x ncbi-blast-2.9.0+/bin/get_species_taxids.sh
x ncbi-blast-2.9.0+/bin/rpstblastn
x ncbi-blast-2.9.0+/bin/blast_formatter
x ncbi-blast-2.9.0+/ChangeLog
→ BLAST

```

²<https://www.computerhope.com/unix/utar.htm>

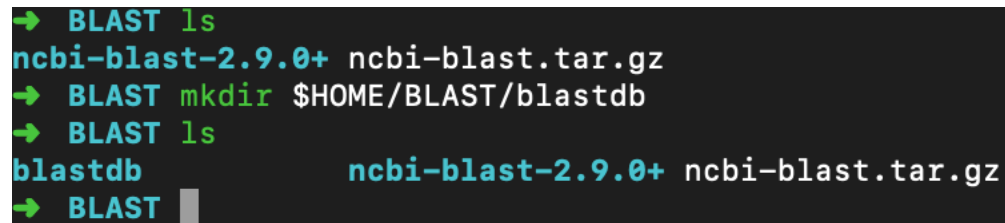
By default, the terminal does not know where BLAST has been installed so you need to set the path to your BLAST bin folder using the command:

```
export PATH=$PATH:$HOME/BLAST/ncbi-blast-2.9.0+/bin
```

Now you will need to create a location for BLAST to look for a database.

```
mkdir $HOME/BLAST/blastdb
```

\$HOME is the same as /Users/username

A terminal window with a dark background and light green text. The commands and output are as follows:
→ BLAST ls
ncbi-blast-2.9.0+ ncbi-blast.tar.gz
→ BLAST mkdir \$HOME/BLAST/blastdb
→ BLAST ls
blastdb ncbi-blast-2.9.0+ ncbi-blast.tar.gz
→ BLAST █

4.2 UNIX

Installing BLAST on Ubuntu is simple using the apt repository. Simply open a terminal with root privilege and enter the following command:

```
sudo apt-get install ncbi-blast+
```

Now you will need to create a location for BLAST to look for a database.

```
mkdir $HOME/BLAST/blastdb
```

\$HOME is the same as /Users/username

5. THE SETS

To perform the BLAST search, you will need two sequences; the subject set and the query set. The subject sequence is the database that the query sequence is compared to. Subject sets will eventually create a personal database or library of subject sequences that are unique to the local BLAST environment. For the purpose of this manual the subject set will be that of the stickleback genome which was used in earlier assignments. The query will be that of MSX1 from the salamander *Ambystoma mexicanum*. The question that will be asked is simply will this homeobox gene sequence from a salamander be found in the genome of the fish stickleback? If it is, how similar is it?

5.1 WHERE TO FIND A SUBJECT SET

Many sequences can be located on Ensembl³. There are plenty of other sources for subject sets as well. Any sequence or number of sequences can be used as a subject set, even but generally the user would want them to be long. The subject set for stickleback can be downloaded directly to the database folder using the command:

```
curl ftp://ftp.ensembl.org/pub/release-98/fasta/gasterosteus_aculeatus/dna/Gasterosteus_aculeatus.BROADS1.dna.toplevel.fa.gz --output $HOME/BLAST/blastdb/broadS1.fa.gz
```

Note the name has been saved as broadS1.fa.gz. Unzip the database with

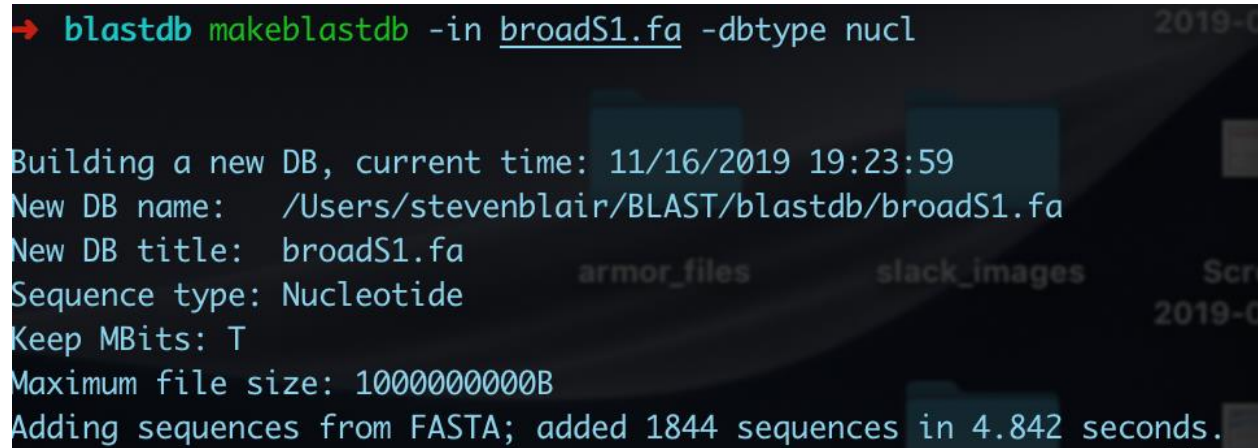
```
gzip -d broadS1.fa.gz
```

³<https://useast.ensembl.org/>

Now you will need to tell BLAST+ that this is a subject set. You do this by using the `makeblastdb` command like this.

```
makeblastdb -in broadS1.fa -dbtype nucl
```

This is the most basic method in creating a database. Many arguments can be used that are explained further in the BLAST+ manual⁴.

A terminal window with a dark background. The command `blastdb makeblastdb -in broadS1.fa -dbtype nucl` is entered at the prompt. The output shows the database being built, including the current time (11/16/2019 19:23:59), the new DB name and title, the sequence type (Nucleotide), and the maximum file size. It concludes by stating that 1844 sequences were added from the FASTA file in 4.842 seconds.

```
→ blastdb makeblastdb -in broadS1.fa -dbtype nucl

Building a new DB, current time: 11/16/2019 19:23:59
New DB name:   /Users/stevenblair/BLAST/blastdb/broadS1.fa
New DB title:  broadS1.fa
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 1844 sequences in 4.842 seconds.
```

5.2 WHERE TO FIND A QUERY SET

Query sets can be found in many places. For the purpose of this manual the query will be MSX-1, also known as msh homeobox 1. The nucleotide sequence is provided at the end of this manual and was found in the publication *The Journal of Experimental Zoology* (Koshiba, 1998).

6. SIMPLE BLAST OF DOWNLOADED EXAMPLES

Once a database has been created and a query is available it is time to BLAST the sets to look for alignments. The simplest method to do this is to use the following command.

- `blastn -db broadS1.fa -query msx1 -out results.txt`
- `blastn` performs a BLAST comparing nucleotides.
- `-db` points to the database that is named “broadS1.fa”
- `-query` points to the file that has the sequence in it “msx1”.
- `-out` is the directory and filename that the results will be written in.

⁴<https://www.ncbi.nlm.nih.gov/books/NBK279680/>

```

Query= MSX-1
Length=1619

Sequences producing significant alignments:

groupVII dna:group group:BR0ADS1:groupVII:1:27937443:1 REF          259          5e-67
groupVI  dna:group group:BR0ADS1:groupVI:1:17083675:1 REF          228          1e-57

>groupVII dna:group group:BR0ADS1:groupVII:1:27937443:1 REF
Length=27937443

Score = 259 bits (140), Expect = 5e-67
Identities = 198/227 (87%), Gaps = 0/227 (0%)
Strand=Plus/Plus

Query  428          GGAAGCACAAAGACCAACCGGAAGCCGCGGACGCCGTTCACCACGTCGCAAGCTGCTGGCCC 487
Sbjct  22854142      GGAACACAAAGACCAACAGGAAGCCCGCACTCCCTTCACCACGTCCTAGCTCTCGGGCC 22854201

Query  488          TGGAGCGGAAGTTCGGCAGAAGCAGTACCTGTCCATCGCCGAGCGCGCGGAGTTCTCGG 547
Sbjct  22854202      TGGAGCGGAAGTTCGGCAGAAGCAGTACCTGTCCATCGCCGAGCGGGCCGAGTTCTCCT 22854261

Query  548          GCTCCCTCAGCCTGACCGAGACGCAAGGTCAAGATCTGGTTCCAGAACC CGCCGCGCAAGG 607
Sbjct  22854262      CGTCCTTGACCTTGACAGAGACTCAAGTGAAGATCTGGTTCCAGAATCGCCGGGCGAAAG 22854321

Query  608          CCAAGCGGCTCGAGGAGGCCGAGCTGGAGAAGCTCAAGATGGCCGCC 654
Sbjct  22854322      CCAAGAGGCTCCAGGAGGCCGAGCTGGAGAACTCAAGATGGCCGCC 22854368

```

7. FURTHER READING

BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Installation. Available from:

<https://www.ncbi.nlm.nih.gov/books/NBK279671/>

BLAST Frequently asked questions:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ

8. CITATIONS AND REFERENCES

Altschul, S. "Basic Local Alignment Search Tool." *Journal of Molecular Biology*, vol. 215, no. 3, 1990, pp. 403–410., doi:10.1006/jmbi.1990.9999

Marchler-Bauer, A., et al. "CDD: a Conserved Domain Database for the Functional Annotation of Proteins." *Nucleic Acids Research*, vol. 39, no. Database, 2010, doi:10.1093/nar/gkq1189

Jones, Tim M. "Conversing through the Internet with CURL and Libcurl." IBM, 8 Sept. 2009, www.ibm.com/developerworks/library/os-curl/index.html

Koshiba, Kazuko, et al. "Expression Of Msx Genes in Regenerating and Developing Limbs of Axolotl." *The Journal of Experimental Zoology*, vol. 282, no. 6, 1998, pp. 703–714., [https://doi.org/10.1002/\(SICI\)1097-010X\(19981215\)282:6<703::AID-JEZ6>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1097-010X(19981215)282:6<703::AID-JEZ6>3.0.CO;2-P)

Ye, J., et al. "BLAST: Improvements for Better Sequence Analysis." *Nucleic Acids Research*, vol. 34, no. Web Server, 2006, doi:10.1093/nar/gkl164

Zhang, Zheng, et al. "A Greedy Algorithm for Aligning DNA Sequences." *Journal of Computational Biology*, vol. 7, no. 1-2, 2000, pp. 203–214., doi:10.1089/10665270050081478

9. MSX1 CONTENTS

Copy and paste the following gene sequence into a file called msx1

>MSX-1

```
TCGGAGTGAAGGCCGAGGAGTCGCCCCGTCTAAGCAAGCAGAGGATGCAGACCGGCCTGA
GCTCCGGGGCGGACCGAGGAGCCCCAGAAACCCAAGCTGCCGGCCATCCTGCCATTTAGC
GTGGAGGCCCTCATGGCTGACCGCAGGCCGACGGTCAGAGACCGTGAGCGGTGCAGCCCC
GCGGGGACCCAGCTGCCCCGGGCCCTCGCAAACCAGCCCCAGGCTAGGGGGGCACCTCTCA
GGACCGGAGTCCCCTGGATCCGCTCTCCATGAACAGACACTATTCCATGGGTGGCTTACT
GCACTTACCAGAAGAGGCTCTTGCGAAGCCGAGAGCCCGGACAGCCAGGAGAGGAACCCG
TGGATGCAGAGCCCCAAATTCTCCCCACCCTCAGCAAGGAGGCTGAGCCCACCGGCCTGC
ACTCTCCGGAAGCACAAAGACCAACCGGAAGCCGCGGACGCCGTTACCACGTCGCAGCTG
CTGGCCCTGGAGCGGAAGTTCCGGCAGAAGCAGTACCTGTCCATCGCGGAGCGCGCCGAG
TTCTCGGGCTCCCTCAGCCTGACCGAGACGCAGGTCAAGATCTGGTTCCAGAACCGCCGC
GCCAAGGCCAAGCGGCTGCAGGAGGCCGAGCTGGAGAAGCTCAAGATGGCCGCCAAGCCC
ATGATGCCGCCGGCCTTCGGCATCTCCTTCCCCCTCGGCTCTCCAGTGCACGCGGCCTCC
CTGTACGGGGCCCTCCGGCCCCCTCCACAGACCCAGCATGCCCATGTCGCCCATGGGACTG
TACGCCGCTCACATGGGCTACAGCATGTACCACCTGACATAAGGGCGCCGCAGACCCACC
ACAGACCATTTCATGCAGCACTTTTCTGATGTTGGGCCCTGCCACGTCTGCCATTGGTGG
CACTCAGGCATGCATGCCAACCACGTTGGAAAGAACCGAGAGCGTGATTCCGGTGGCAGGA
AGAGGGGGGTTGTGCATGCCATTGGCTCTCATCGCAATGAAGGAACGCTATGCCAGGCA
TTGCACACCTTTAACAAGTTGAACAAGGACAATGTTTTGTGTCGTGAAGGAGCGCCTCCC
ACTTCTGAATAATAGAGAGATGGCATGTGTGCACCAGCCTGAAATACGCCAGGCGTTTTG
GATTTTCACAGTGTGTTCAACACCTGTAGAGGGGAACTGAAACATATTTGTGAGAAGTTCA
CGTTTGGACATACAGTTCCTCACACGTGGTTTACAGAAAAGTCCAGCATTTTCAGCAGCTC
AACCTGGCTCAGCACCATTCAATACAGAAAGCCCGACATCTTGTTGTATGGCCGCATGAA
TTAGTTCACATCACCGGGAAATGTCATGAGTTCTAAGAAGATGACTTTTTATAAATAAAG
CGCTATCGAAAATGCTCCTCAAAAGTGCCACCAGACACACGTGGAAAGGCAACAGAACTT
GTCAACGAATCACTGTGCTTCACTGTTTCCCTTGCCTGTGGATGTTTCTACACTCGTCCC
TTGGGAGCAGGGGATCCGTACTATGTAATATACTGTATATTTGAAAAAATATTATCATT
TATATTATAGCTATATTTGTTAAATAAATTAATTTTAAGCTAAAAAAAAAAAAAAAAAAAA
```