BLAST

1 Resources

We will learn how to locally run BLAST from command line - Blast+ user manual: https://www.ncbi.nlm.nih.gov/books/NBK1762

While learning basic BLAST commands, we will also encounter these new programs/commands/concepts:

- ftp
- wget
- md5sum
- tar
- echo
- $PATH
- pipes
- sort
- uniq

2 Installation

In case that a file cannot be downloaded from the provided URL, all files can be downloaded, using wget, from: http://bioinformatics.med.uoc.gr/bioinfo_grad/

2.1 Download

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

1. **Connect** using FTP (BLAST can also be downloaded using wget but for the purpose of the lesson we will be using ftp)

```
ftp ftp.ncbi.nlm.nih.gov (user:anonymous   pswd:your email)
```

2. **List** contents of remote server and go to

```
ls
```

3. Go to **BLAST directory**

```
cd blast/executables/LATEST/
```

4. List contents in order to find **latest release**

```
ls
```

5. **get** the latest appropriate release and the hash file

```
get ncbi-blast-2.7.1+-x64-linux.tar.gz
get ncbi-blast-2.7.1+-x64-linux.tar.gz.md5
```

6. **quit**

```
quit
```

Alternatively, all the above can be completed using wget:

```
wget ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.7.1+-
x64-linux.tar.gz

wget ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.7.1+-
x64-linux.tar.gz.md5
```

7. **Check integrity**:

```
md5sum -c ncbi-blast-2.7.1+-x64-linux.tar.gz.md5
```

md5sum calculates 128-bit MD5 hashes, which are used as compact digital identifiers for files. It SHOULD NOT be used for security related tasks, but it is commonly used for verifying integrity of file transfers.

2.2 Install

```
tar zxvpf ncbi-blast-2.7.1+-x64-linux.tar.gz (the z unzips)
```

-z --- put the file though gzip or use gunzip on the file first.
-x --- extract the files from the tarball.
-v --- verbose, give more output, show what files are being worked with (extracted or added).
-p --- preserves dates, permissions of the original files.
-f --- file (create or extract from file) - should always be the last option otherwise the command will not work.

All this arguments are difficult to remember. Use an alias:

```
alias tarf="tar zxvpf"
```

2.3 Configure

We need to make it so we can run blast commands from any directory. Currently, we have to run them using their full path e.g.:

```
~/BLAST/ncbi-blast-2.7.1+/bin/blastp
```

Commands that can be ran anywhere without invoking the absolute path, are in directories specified by the PATH variable ($PATH).

```
echo $PATH
```

We have to either store our command in one of these directories or, in our case, add the directory to $PATH:

```
export PATH=$PATH:~/BLAST/ncbi-blast-2.6.0+/bin/
Additionally, create a directory for blast databases e.g. blastdb
export BLASTDB=~/blastdb
```

3 BLAST Databases

3.1 Download a DB from NCBI

```
ftp://ftp.ncbi.nlm.nih.gov/blast/db/
```

in our case 16SMicrobial.tar.gz (ftp://ftp.ncbi.nlm.nih.gov/blast/db/16SMicrobial.tar.gz)

and

swissprot.tar.gz.md5 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz)

Download both files into the blastdb directory and extract them

3.2 Check downloaded database:

- General info

```
blastdbcmd -info -db 16SMicrobial
blastdbcmd -info -db swissprot
```

this works because we set the BLASTDB variable

- Print all records (only titles, gi, taxonomy id) (16SMicrobial)

```
blastdbcmd -db 16SMicrobial -entry all -outfmt "%t %g %T"
```

- Sort and display number of sequences for each species (16SMicrobial)

```
blastdbcmd -db 16SMicrobial -entry all -outfmt "%T"|sort|uniq -c
blastdbcmd -db 16SMicrobial -entry all -outfmt "%T"|sort|uniq -c|sort -
n
```

- Print only specific sequences

```
blastdbcmd -db 16SMicrobial -entry 645321669 -outfmt "%t %g %T"
blastdbcmd -db swissprot -entry 112690,112672
blastdbcmd -db swissprot -entry 112690,112672 > query.fa
```

4 Run BLAST

- Basic BLAST run

```
blastp -query query.fa -db swissprot
```

- BLAST run with alignment view options

```
blastp -query query.fa -db swissprot -outfmt "7 qseqid sseqid pident
ppos evalue bitscore"
```

5 Make your own database

- Download a FASTA file

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/HUMAN_MICROBIOM/Bacteria/Escherichia
_coli_SE11_uid18057/NC_011415.ffn
```

- create BLAST db named ecoli

```
makeblastdb -in NC_011415.ffn -out blastdb/ecoli -dbtype nucl
```