Five ways to run BLAST

- locally from the shell command line;
- locally from a Python script or interactive Python session;
- locally using Biopython;
- through the NCBI web server using Biopython;
- using your browser and the BLAST web page.

What are the advantages of running BLAST locally?

- you can search a query sequence in a customised database, e.g. in a newly sequenced genome you are studying, or a set of protein sequences of your interest (e.g. only protein kinases).
- you may want to insert the program in a pipeline
- only by running BLAST locally you have full control over the sequence database and by that, reproducibility of your search

Running Blast locally

- Download and install the BLAST+ package (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download).
- http://www.ncbi.nlm.nih.gov/books/NBK1762/.

- The downloaded files are unpacked into a BLAST directory, and you have to add the path of this directory to the PATH environment variable of your computer (i.e. tell your system where to look for the installed BLAST programs)
- Otherwise, you have to change to the BLAST directory on the shell and run BLAST from there.
- Inform the BLAST programs which directory to search for the databases.
- In other words you have to modify two environment variables: PATH and BLASTDB.

Modify the PATH environment variable

- When you install the program from source, you will have to place the downloaded package under a desired directory, e.g. `/home/john`
- When you unpack the package, a BLAST directory will appear in `/home/john` (e.g. `/home/john/ncbi-blast-2.2.23+`).
- You have to add to the PATH environment variable the bin directory under this BLAST directory

Under the bash shell (`.bash_profile`)

```
PATH=$PATH:/home/john/blast-2.2.23+/bin
export PATH
```

Under the tcsh shell (`.cshrc`):

```
setenv PATH ${PATH}:/home/john/ncbi-blast-2.2.23+/bin
```

Notice that when you use the dmg disk to install BLAST+ on Mac OS X (10.4 or higher), all BLAST+ programs will be installed under `/usr/local/ncbi/blast/bin`.

Modify the BLASTDB environment variable

- Create the BLAST database directory `/blast/db` in your home directory

```
mkdir /home/john/blast/db
```

- This is the directory where you will put all the databases (either downloaded from the BLAST website or your custom ones) that you will use with BLAST.
- Save at least a database in `/home/john/blast/db`.
- If you want to download a database from NCBI, go to ftp://ftp.ncbi.nlm.nih.gov/blast/db.

Create a `.ncbirc` text file in your home directory having the following path specification

```
; Start the section for BLAST configuration
[BLAST]
; Specifies the path where BLAST databases are installed
BLASTDB=/home/john/blast/db
```

The semicolon at the beginning of the first and third lines indicates a comment.

- Unless you use a pre-formatted database downloaded from the NCBI ftp site, you will need to format your custom sequence file.
- `makeblastdb` produces BLAST databases from FASTA files:

`makeblastdb –in genome.fasta -parse_seqids -dbtype prot`

- `-in` is the option for the input file,
- `-parse_seqids` enables parsing of sequence ids
- `-dbtype` type of input molecules (nucl or prot).

The query sequence can be in FASTA format and this is the structure of the command line:

`blastProgram -query InSeq.fasta -db Database -out OutFile`

For example,

`blastp -query P05480.fasta -out blast_output -db nr.00`

- `blastp` aligns protein sequences, nr.00 is the name of the BLAST-formatted database,
- `blast_output` is the name you have chosen for the BLAST output
- `P05480.fasta` is a file that contains your query sequence in FASTA format.

Program	Task Name	Description
blastp	blastp	Traditional BLASTP to compare a protein query to a protein database
	blastp-short	BLASTP optimized for queries shorter than 30 residues
blastn	blastn	Traditional BLASTN requiring an exact match of 11
	blastn-short	BLASTN program optimized for sequences shorter than 50 bases
	megablast	Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
	dc-megablast	Discontiguous megablast used to find more distant (e.g., interspecies) sequences

Exit Code	Meaning
0	Success
1	Error in query sequence(s) or BLAST options
2	Error in BLAST database
3	Error in BLAST engine
4	Out of memory
5	Network error connecting to NCBI to fetch sequence data
6	Error creating output files
255	Unknown error

option	**type**	**default value**	**description and notes**
db	string	none	BLAST database name.
query	string	stdin	Query file name.
query_loc	string	none	Location on the query sequence (Format: start-stop)
out	string	stdout	Output file name
evalue	real	10.0	Expect value (E) for saving hits
subject	string	none	File with subject sequence(s) to search.
subject_loc	string	none	Location on the subject sequence (Format: start-stop).
show_gis	flag	N/A	Show NCBI GIs in report.

The option for the output format is **`-outfmt`**

0 = pairwise,
1 = query-anchored showing identities,
2 = query-anchored no identities,
3 = flat query-anchored, show identities,
4 = flat query-anchored, no identities,
5 = XML Blast output,
6 = tabular,
7 = tabular with comment lines,
8 = Text ASN.1,
9 = Binary ASN.1
10 = Comma-separated values