

**Constructing blast database**

For list of all program options:

```
>formatdb --help
```

Frequently used options:

```
-p    T or F (T= protein, F=nucleotide)
-i    input file
-t    title of database e.g. "Entire Y. pestis KIM genome"
-o    T or F (T=parse names or do not=F)
-n    name for database files
```

Example:

Make database out of NC\_004088.faa multisequence fasta file:

```
>formatdb -p T -i NC_004088.faa -t "All Y. pestis KIM
proteins" -o T -n YP_AA
```

**Simple homology searches**

For list of all program options:

```
>blastall --help
```

Frequently used options:

```
-p    blast algorithm to use blastp, blastn, blastx, tblastx
      tblastn
-d    blast database to use (your own or general use)
-i    input file (e.g. gene_p.faa)
-o    output file (e.g. gene_p.br)
-e    E value threshold (e.g. 0.00001 or 10E-5)
-F    T or F Filtering of low complexity regions (e.g.
      AAAAAAAAA)
-b    number of alignments to show (default 250)
-v    number of one line descriptions to show (default 500)
-m    output style (0-9) (default 0)
```

Example:

```
>blastall -p blastp -d YP_AA -i gene.faa
```

prints results to screen

```
>blastall -p blastp -d YP_AA -i gene.faa -o gene.br -e
10e-10 -m 9 -b 20 -v 20
```

prints results to tab separated file and limited to top 20 hits with E-values less than  $10^{-10}$ .

**Retrieval of sequences from blast database**

For list of all program options:

```
>fastacmd --help
```

Frequently used options:

```
-d  database to query
-p  T or F (T= protein, F=nucleotide)
-s  comma separated string of database entries, entered on
    the command line
-i  input file containing list of database entries
-D  Dump entire contents of database
-L  start,stop coordinates - Return part of an entire
    database entry
-o  output file
-S  1=forward strand, 2=reverse strand
```

Details of database:

```
>fastacmd -d YP_AA -I
```

Retrieve two entries:

```
>fastacmd -d YP_AA -p T -s 22123923,22123924 -o 2genes.faa
```

Retrieve part of one entry:

```
>fastacmd -d YP_AA -p T -s 22123922 -L 1,1000 -o
part_seq.ffn
```

**Running alignments on the command line**

Follow command line prompts:

```
>clustalw
```

or

```
>mafft
```

Automatically start run:

```
>clustalw -infile=input.faa
```

other options:

```
-OUTPUT=NEXUS GCG,FASTA, PHYLIP    output file format
-OUTFILE=output                    rename output file as you like
```

Mafft command line options:

```
>mafft --help
```

Automatically start run:

```
>mafft input.faa > output.aln
```

other options:

```
--clustalout  output as clustal instead of fasta
--reorder     reorder sequences based on similarity
```

**Running a homology search between two specific sequences**

For list of all program options:

```
>bl2seq --help
```

Frequently used options:

```
-i  Query sequence
-j  Subject sequence
-e  E value threshold (e.g. 0.00001 or 10E-5)
-p  blast algorithm to use blastp, blastn, blastx, tblastx
    tblastn
-o  output file
-D  Output type 0=standard, 1=tab delimited (like -m 9)
```

**Sequence file manipulation and sequence information:**

EMBOSS software tools loaded on amadeus.biosci.arizona.edu

List all programs:

```
>wosname
```

Finds programs by keywords in their one-line documentation

Text to search for, or blank to list all programs:

List details of a program(e.g. man page of each program)

```
>tfm <program name>
```

```
>tfm infoseq
```

Get tab delimited file describing multisequence fasta file (NT or AA)

```
>infoseq genes.faa
```

Get %G+C for a NT multisequence fasta file

```
>geecee genes.ffn -outfile genes.gc
```

Find a NT pattern in a multisequence fasta file

```
>fuzznuc -c -outfile output -pattern GGCCGACATGT -pmismatch
```

```
0 NC_009793.fna
```

```
-c          check reverse complement
```

```
-pattern    pattern to search for
```

```
-pmismatch  number of mismatches allowed in pattern
```

Convert NT multisequence fasta file to AA multisequence fasta file

```
>transeq genes.ffn genes.faa
```