A Quick Guide to the NCBI Blast

http://www.ncbi.nlm.nih.gov/blast

Blast (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is a sequence comparison algorithm optimized for speed and used to search sequence databases for optimal local alignments to a query. The NCBI implementation was established by the National Center for Biotechnology Information. The program can be used through the NCBI site or can be installed locally (stand alone blast).

This guide doesn't replace the entire documentation for Blast but can be used as a reference.

Where to start?

For beginners we suggest to first read the documentation of the Blast related to similarity searching (see link below). Other useful pages are available by following the links at the top of this page.
E.g., the glossary and the tutorials:
`http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/similarity.html`

Program selection – web interface options

BLASTN – used to search nucleotide databases with a nucleotide query sequence.

MEGABLAST – a version of BLAST specially designed to efficiently find very similar sequences in a database.

Discontiguous MEGABLAST – a version of MEGABLAST used to identify similar but not identical nucleotide sequences.

Search for short nearly exact matches – used to search for primer or short nucleotide motifs in nucleotide sequences or short peptides in protein sequences.

BLASTP – used to search protein databases with a protein query sequence.

PSI-BLAST (Position-Specific Iterated BLAST) – used to search protein databases with increased sensitivity potentially locating distant homologies. A position-specific scoring matrix is created after each iteration using the selected results from the previous search.

PHI-BLAST (Pattern-Hit Iterated BLAST) – a version similar to PSI-BLAST, but including a user-defined pattern limiting the output to sequences matching the pattern. The patterns must follow the pattern syntax conventions from PROSITE.

BLASTX – makes a six-frame nucleotide query search against a protein database, finding proteins similar to those encoded by the query. Useful when the reading frame of the query is unknown or when it contains errors that may lead to frame shifts.

TBLASTN – makes a protein query search against a dynamically translated nucleotide database. Useful when searching for a specific protein against an unannotated nucleotide database, like HTGs or ESTs databases.

TBLASTX – searches all six-frame query translations against all six-frame database translations. Effectively performs a more sensitive blastp search without doing manual translations.

CDD-Search (Conserved Domain Database Search) – used to identify conserved protein domains.

CDART (Conserved Domain Architecture Retrieval Tool) – explores the domain architectures of proteins.

Blast 2 sequences – direct comparison of two sequences.

VecScreen – screens DNA sequence queries for vector contamination using a database of known vectors.

Main databases (available at NCBI)

Protein *nr* (non-redundant + PDB + SwissProt + PIR + PRF); *swissprot* (latest major release of the SWISS-PROT); *pat* (proteins from patent division of GenBank); *month* (new data released in the last 30 days); *pdb* (3-dimensional structure records from Protein Data Bank).

Nucleotide *nr* (GenBank + EMBL + DDBJ + some PDB); *est* (GenBank + EMBL + DDBJ from EST division); *pat* (nucleotides from patent division); *pdb* (3-dimensional structure records); *month* (new data released in the last 30 days); *chromosome* (complete genomes and chromosomes); *est human* (human subset of EST); *est mouse* (mouse subset of EST); *est others* (subset of EST other than human or mouse); *gss* (Genome Survey Sequence); *htgs* (Unfinished High Throughput Genomic Sequences); *alu_repeats* (select Alu repeats from REPBASE); *dbsts* (STS division + EMBL + DDBJ); *wgs* (assemblies of whole genome shotgun sequences).

LOCAL BLAST INSTRUCTIONS

Format source databases

formatdb formats protein or nucleotide source databases before they can be searched by blastall, blastpgp or megablast. The source database may be in either FASTA or ANS.1 format.

Selected formatdb arguments:

-t [string]	title for database (opt).
-i [file in]	input file for formatting.
-l [file out]	logfile name (opt; def = formatdb.log).
-p [T/F]	type of file (opt; T = protein (def); F = nucleotide).
-o [T/F]	parse options (opt; T = parse SeqID and create indexes; F = no parse, no indexes (def)). Obs.: the first word on the fasta definition line should be a unique identifier (SeqID).
-v [integer]	size of the volume in millions of letters (opt; def = 0). Obs.: This option breaks up large FASTA files into 'volumes' (each with a maximum size of 2 billion characters). I.e.: -v 2000.
-n [string]	base name for BLAST files (opt).

Fasta from databases

fastacmd retrieves FASTA formatted sequences from a BLAST database, if it was formatted using the '-o' option.

Selected fastacmd arguments:

-d [string]	database (def = nr).
-s [string]	search string.
-i [string]	input file with GIs/accessions/locuses for batch retrieval (opt).
-l [integer]	line length for sequence (def = 80, opt).

Stand-alone blast

blastall performs all five flavors of blast comparison.

Selected blastall arguments:

-p [string]	program name (input should be one of "blastp", "blastn", "blastx", tblastn" or "tblastx").
-d [string]	database (def = nr). Obs.: Multiple database names will be accepted if quoted. E.g., -d "nr est".
-i [file in]	query file (def = stdin). Obs.: Query should be in FASTA format. If multiple FASTA entries are in the input file, all queries will be searched.
-e [real]	expectation value threshold (def = 10.0).
-o [file out]	BLAST report output file (opt; def = stdout).
-F [string]	filter query sequence (def = T). Obs.: T = DUST for blastn or SEG for others, and F = no filtering. To change SEG options, use: `-F "S 10 1.0 1.5"`, where 10 = window value, 1.0 = low cut and 1.5 = high cut. For coiled-coil filter: `-F "C 28 40.0 32"`, where 28 = window, 40.0 = cut off and 32 = linker. To use both SEG and coiled-coil: `-F "C;S"`.
-b [integer]	number of alignments (def = 250).
-v [integer]	number of one-line description (def = 500).
-Q [integer]	query genetic code (def = 1).

-D [integer] DB genetic code (def = 1).
-M [string] matrix (def = BLOSUM62).
-T [T/F] produces HTML output (def = F).
-U [T/F] uses lower case filtering (def = T) Obs.: T = any lower-case letter in input FASTA file should be masked.

Position Specific Iterated BLAST

PSI-BLAST is a variant of blast that searches a query against a database using a position-specific scoring matrix created by PSI-BLAST. First run **blastpgp** to create and save a position-specific scoring matrix, then run **blastpgp** again to search iteratively with the previously saved matrix. e.g.,

```
blastpgp –i ff.chd –d yeast –C ff.chd.ckp
blastpgp –i ff.chd –d nr –j 3 –R ff.chd.ckp
```

Selected blastpgp arguments for PSI-BLAST:

-j [integer] maximum number of iterations (def = 1).
-h [number] E-value threshold for including sequences in the score matrix model (def = 0.001).
-C [file out] stores the query and frequency count ratio matrix in a file (opt).
-Q [file out] output file for PSI-BLAST matrix in ASCII (opt).
-R [file in] restarts from a file stored previously with -C.
-B [file in] input alignment for restart.

Pattern-Hit Initiated BLAST

PHI-BLAST is a search program that combines the matching of regular expressions with local alignments surrounding the match. E.g.,:

```
blastpgp –i query.file –k pattern.file –p patseedp
```

Select blastpgp arguments for PHI-BLAST:

-i [file in] input sequence file in FASTA format.
-k [file in] pattern (syntax follows the PROSITE conventions).
-p [string] usage mode (def = blastpgp). Obs: use 'patseedp', if pattern occurs only once, and 'seedp', if it occurs more than once per protein.

Obs.: You can integrate a PSI-BLAST search after the PHI-BLAST search, using the argument "-j". E.g.,

```
blastpgp –i query –k pattern –p patseedp –j 2
```

Mega BLAST

Mega BLAST uses a greedy algorithm optimized for aligning sequences that differ slightly as a result of sequencing or other similar «errors». When a larger word size is used, it is up to 10 times faster than more common sequence similarity programs. It is also able to efficiently handle much longer DNA sequences than the blastn program.

Selected megablast arguments:

-D [integer] type of megablast output (def = 0 = alignment endpoints and score; 1 = all ungapped segments endpoints; 2 = traditional BLAST output; 3 = tab-delimited one line format).
-M [integer] maximal total length of queries for a single search (def = 20000000).
-f [T/F] shows full Ids in the output (def = F, only GIs or accessions).
-p [real] identity percentage cut off (def = 0).
-s [integer] minimal hit score to report (def = 0).

To compare two sequences

bl2seq performs a pairwise comparison between two sequences.

Selected bl2seq arguments:

-i [file in] first sequence.
-j [file in] second sequence.
-p [string] program name (as in blastall; def = blastp).
-o [T/F] alignment output (def = stdout).
-G [integer] cost to open a gap (def = 0; zero invokes default behavior).
-E [integer] cost to extend a gap (def = 0; zero invokes default behavior).
-W [integer] wordsize (def = 0; zero invokes default behavior).
-M [string] matrix (def = BLOSUM62).
-F [string] filters query sequence (def = T).
-e [real] expectation value E (def = 10.0).
-T [T/F] produces HTML (def = F).

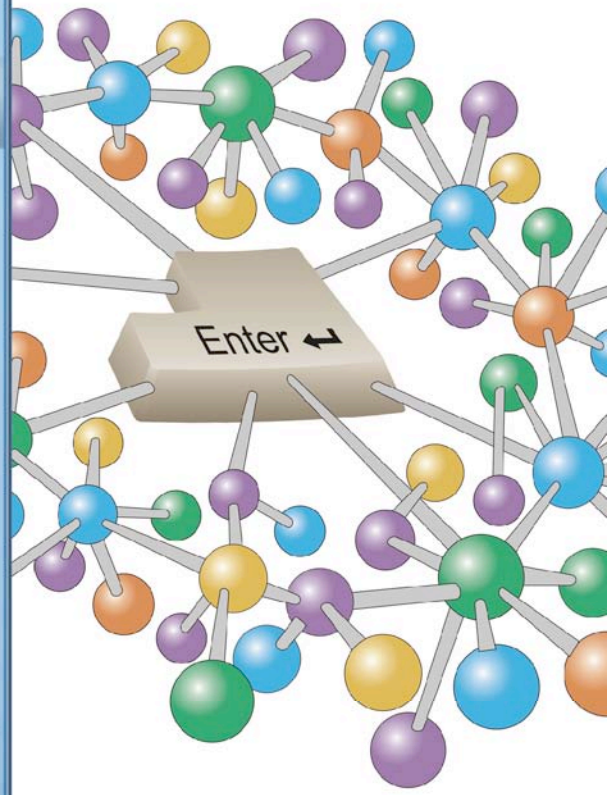
This document was written and designed by Eduardo Fernandes Formighieri with the help of Marcos Renato R. Araújo, Marcelo Falsarella Carazzolle and Gonçalo A. Guimarães Pereira from the Brazilian EMBnet node and distributed by the P&PR Publications Committee of EMBnet.

EMBnet – European Molecular Biology network – is a network of bioinformatics support centers situated primarily in Europe. Most countries have a national node, which can provide training courses and other forms of help for users of bioinformatics software.

http://www.embnet.org/

A Quick Guide to NCBI Blast
First edition © 2004

EMBnet



A Quick Guide
BLAST