



Searching NCBI Databases with BLAST

BLAST, the Basic Local Alignment Search Tool, is a set of algorithms that find regions of similarity between biological sequences. BLAST can be used to search sequence databases such as NCBI's Nucleotide and Protein databases and calculate the statistical similarity of matches. Access BLAST at <http://blast.ncbi.nlm.nih.gov>.

- BLAST can be used to:
- Identify a sequence
 - Find related sequences:
 - to infer function
 - to infer species relatedness
 - to perform phylogenetic analysis

How BLAST works

BLAST works by breaking a query sequence into a series of "words" of a set number of letters. These words are compared to words from sequences in the database. Once a match is found, the sequences are aligned and scored.

The number of letters per word can be changed by using the "Algorithm parameters" link on the BLAST screen.

Insertions and deletions in sequences result in gaps when sequences are aligned. These gaps are assigned a certain score penalty for their existence and for their extension. These scores can be changed from their defaults by using the "Algorithm parameters" link on the BLAST screen.

Limit by Entrez Query

The Entrez Query box on the BLAST page allows you to limit your results using Entrez terms, e.g. "last 30 days"[MDAT] for records created or modified in the last 30 days.

- Note: if using the operator NOT without another search term, use the "all" filter first:
e.g. **all[filter] NOT mitochondrion[filter]** to exclude sequences known to be mitochondrial.

Filters and Masking

BLAST can filter out low complexity regions that may yield spurious matches with unrelated sequences. It can also filter nucleotide repeat regions for specific species and mask lower-case letters (e.g. if the sequence input is in FASTA format). Filtering and Masking can be turned on and off by using the "Algorithm parameters" link on the BLAST screen.

Nucleotide Sequence Queries (DNA or RNA)

BLAST Algorithms

megablast is the default. It is designed to find nearly identical sequences and is ideal to identify an unknown sequence.

discontiguous megablast is designed to find more distantly related sequences (e.g., in other organisms). It takes into account third base wobbling, and allows words to be non-contiguous.

blastn allows a very short word size to find distantly related sequences.

blastx translates a nucleotide query in all 6 reading frames and uses it to search the selected protein database.

tblastx translates a nucleotide query to protein sequence in all 6 reading frames and searches against a nucleotide database which has been translated to protein sequences in all 6 reading frames.

Nucleotide Databases (partial list)

Note: no single nucleotide database contains all NCBI nucleotide sequences.

Human Genomic plus transcript (Human G+T)

- This is the default database. It contains human DNA and RNA sequences.

Nucleotide collection (nr/nt)

- Contains all GenBank and PDB sequences except Expressed Sequence Tags, Sequence Tagged Sites, Genomic Survey Sequences and unfinished High Throughput Genomic Sequences (all of which can be searched separately) Note: this database is NOT non-redundant.

Reference mRNA sequences (refseq_rna) and Reference genomic sequences (refseq_genomic)

- Standardized, curated sequences from the NCBI's Reference Sequence project.

Protein Data Bank (pdb)

- Sequences derived from the 3-dimensional structure records from Protein Data Bank.

High throughput genomic sequences (htgs)

- Contains unfinished High Throughput Genomic Sequences; finished HTG sequences are in nr/nt

Environmental samples (env_nt)

- Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples, e.g. the Sargasso Sea project. These sequences are NOT in nr/nt.

Protein Sequence Queries

BLAST Algorithms

blastp is the default. It is a standard BLAST protein sequence search.

PSI-BLAST (*Position Specific Iterative BLAST*) is run in multiple iterations. The first iteration uses the results of a *blastp* search to create a position specific score matrix (PSSM). This matrix is used instead of the standard matrices in subsequent iterations to generate more specific results.

PHI-BLAST (*Pattern Hit Initiated BLAST*) searches the database for matches to your query that contain a specific pattern from the query that you specify.

tblastn uses a protein query to search against a nucleotide database which has been translated in all 6 reading frames.

Matrices for Protein BLAST

BLAST uses a simple match/mismatch scoring system to compare nucleotide sequences. Scoring protein sequences is more complex. Scoring matrices can be changed by using the "Algorithm parameters" link on the BLAST screen.

BLOSUM62 is the default matrix, determined to be efficient for detecting most weak protein similarities. *BLOSUM80* is often more useful to detect very similar sequences, and *BLOSUM45* may be better for more divergent sequences.

PAM is the older family of matrices, which may be more useful than BLAST for very short queries. *PAM30* is useful for very similar short sequences, while *PAM70* is better for more divergent short sequences.

Protein Databases

Non-redundant protein sequences (nr)

- Non-redundant database of GenBank coding sequence translations and Protein Data Bank, SwissProt, Protein Information Resource and Protein Research Foundation sequences, excluding those in *env_nr*

Reference proteins (refseq_protein)

- Standardized, curated protein sequences from the NCBI Reference Sequence project

Swiss-prot protein sequences (swissprot)

- Last major release of the SwissProt protein sequence database

Patented protein sequences (pat)

- Proteins from the Patent division of GenBank

Protein Data Bank proteins (pdb)

- Sequences derived from 3-dimensional structure records in the Protein Data Bank

Environmental samples (env_nr)

- Non-redundant coding sequence translations of nucleotide sequences from the environmental samples database

Results Note: the BLAST results display may be changing soon

Conserved Domains

With protein searches, BLAST automatically searches for conserved domains that resemble known functional or structural regions and displays them graphically while your search continues. You can also see the conserved domains on the results page by clicking "Show Conserved Domains".

Results Display

Graphic Summary

The BLAST results screen presents a graphical overview of the matches found, showing bars aligned with the matching region and color-coded according to the alignment score. Mouse-over a bar to see the accession number and name of the hit, and click the bar to see an alignment with your query sequence.

Descriptions

This brief results list shows accession numbers, names, scores and e-values of BLAST hits. It also provides links to databases such as Gene, UniGene and Structure when the hit has records in those databases.

Alignments

An extended list of results shows pairwise alignments for the most significant hits.

Other Reports

Links at the top of the screen to *Other reports* include color-coded distance trees and lists of results organized taxonomically.

Manipulating your Results

The *Formatting options* link allows you to filter your results and change the display formatting. The *Edit and Resubmit* link takes you back to the BLAST search page, where you can change parameters and rerun the search.

E-value

The e-value is the best determinant of the value of your match. It is essentially a false-positive rate: it indicates the number of results you would expect to get by chance with a score as good or better. The score is useful in calculating the e-value but is not the best measure of sequence matching – it depends heavily on sequence length and algorithm parameters, and shouldn't be used to compare a match to the results of other BLAST searches.