

Bioinformatics: Life Sciences on Your Computer

PSI-BLAST, Delta-BLAST

This is supplemental reading to the video on PSI-BLAST.

Other database searches beyond BLAST

Ensembl BLAST, WU-Blast: BLAST variations

PSI-BLAST: Based on PSSMs built from scratch

Delta-BLAST: Based on preexisting PSSMs

PHI-BLAST: Based on patterns

HMMER: Based on HMMs

BlastZ, MegaBLAST, Discontinuous MegaBLAST: For long genomic DNA

BLAT: alignment tool

EBI

The European Bioinformatics Institute ([EBI](http://www.ebi.ac.uk)) has multiple BLAST varieties available, as well as FASTA, and its own [ENA](http://www.ebi.ac.uk/Tools/sss/) program. Some slow exhaustive programs are also available. The extensive menu (below) can be found at <http://www.ebi.ac.uk/Tools/sss/>.

The screenshot shows the EBI Sequence Similarity Searching interface. At the top, there's a breadcrumb trail: EBI > Tools > Sequence Similarity Searching. Below this is a header 'Sequence Similarity Searching'. The main content is organized into sections: BLAST, FASTA, SSEARCH, PSI-Search, GGSEARCH, GLSEARCH, and FASTM. Each section includes a brief description of the tool and a list of available databases with corresponding search buttons. For example, under BLAST, there are buttons for Protein, Nucleotide, and Vectors. Under FASTA, there are buttons for Protein, Nucleotide, Proteomes, Genomes, Whole Genome Shotgun, ASD Protein, ASD Nucleotide, LGIC Protein, and LGIC Nucleotide. At the bottom, there's a section for ENA SEQUENCE SEARCH, which mentions that EBI has a new search tool faster than BLAST and provides a link to the ENA Sequence Search page.

EBI > Tools > Sequence Similarity Searching

Sequence Similarity Searching

BLAST

NCBI BLAST ⓘ NCBI BLAST Sequence Similarity Search using the NCBI BLAST (blastall) program. This tool is available for the following databases:

[Protein](#) [Nucleotide](#) [Vectors](#)

WU-BLAST ⓘ Sequence Similarity Search using the Washington University (WU) BLAST2 program (BLAST 2.0 with gaps). This tool is available for the following databases:

[Protein](#) [Nucleotide](#) [Parasites](#)

PSI-BLAST ⓘ Position Specific Iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile is automatically constructed from the first set of BLAST alignments.

[Launch PSI-BLAST](#)

FASTA

FASTA ⓘ Sequence Similarity Search using the FASTA program. This tool is available for the following databases:

[Protein](#) [Nucleotide](#) [Proteomes](#) [Genomes](#) [Whole Genome Shotgun](#)

[ASD Protein](#) [ASD Nucleotide](#) [LGIC Protein](#) [LGIC Nucleotide](#)

SSEARCH ⓘ Sequence Similarity Search using the SSEARCH program. This tool is available for the following databases:

[Protein](#) [Nucleotide](#) [Proteomes](#) [Genomes](#) [Whole Genome Shotgun](#)

[ASD Protein](#) [ASD Nucleotide](#) [LGIC Protein](#) [LGIC Nucleotide](#)

PSI-Search ⓘ PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST (blastpgp) iterative profile construction strategy to find distantly related protein sequences.

[Launch PSI-Search](#)

GGSEARCH ⓘ GGSEARCH performs a sequence search using alignments that are global in the query and global in the database (Needleman-Wunsch).

[Protein](#) [Nucleotide](#)

GLSEARCH ⓘ GLSEARCH performs a sequence search using alignments that are global in the query and local in the database.

[Protein](#) [Nucleotide](#)

FASTM ⓘ Peptide similarity searching using the FASTM/FASTS/FASTF programs. This tool is available for the following databases:

[Protein](#) [Nucleotide](#) [Proteomes](#) [ASD Protein](#) [LGIC Protein](#)

ENA SEQUENCE SEARCH

The EBI has a new search tool which is far faster than BLAST, with only a marginal loss in search sensitivity. Try it out at [ENA Sequence Search](#).

Alternate query types

Some alternate query types can be used with BLAST and other programs and are listed below. The last two are the basis for PSI-BLAST and PHI-BLAST.

Partial sequence: a subset or fragment of larger sequence, possibly a domain

Gapped sequence: from aligning original protein with others in family, gaps might help find other homologs

Consensus sequence: representative sequence from an alignment of homologs

Matrix: mathematical representation of family of sequences (PSI-BLAST)

Regular expression: qualitative description of a sequence pattern (PHI-BLAST)

Position-Specific Scoring Matrix (PSSM)

A PSSM is NOT a sequence. It is a matrix that represents frequency that various amino acids are found in certain positions. It is rarely used for nucleotides. The PSSM is derived from an ungapped block or conserved region that is common to two or more sequences.

PSI-BLAST

Position-Specific Iterated BLAST (PSI-BLAST) is a user-friendly version of a profile-based search.

***BLAST does not always find all homologs because no search program is 100% sensitive. To find putative homologs that BLAST does not find, search with mathematical representation of the protein family. It is important to note that PSI-BLAST finds putative homologs. It is best to confirm the homology in the wet lab or in the dry lab.*

PSI-BLAST (how to)

Iteration 1: BLAST search

The best hits from the BLAST search are used to build the PSSM.

Iteration 2: PSSM replaces query

Then continue iterations with new PSSMs until no more sequences are found (convergence). If the query is too broad or a non-homolog works its way into the PSSM, then unrelated sequences cause new sequences to be added every time (divergence).

The first iteration is a standard blastp search of a desired database. Different databases can be searched, and the output can be limited to one organism or group of organisms. The word size, matrix and gap costs can be changed.

For the second iteration, hit Run Two. The default settings can be changed, particularly the threshold. Usually, new sequences are added which help build the third iteration PSSM. You can choose to add or subtract sequences for the next iteration by checking or unchecking the check boxes.

Discussion topic: After watching the video PSI-BLAST lecture or this [YouTube video](#), any confusion about how to do a PSI-BLAST search?

PSI-BLAST summary

Select a query and search a protein database. PSI-BLAST does a multiple alignment, using homologs from iteration 1 and the query. It generates a PSSM which is used to query database for each successive iteration. PSI-BLAST estimates statistical significance, however the E value generated compares each sequence match to the PSSM (NOT to the query). Then repeat steps iteratively. For each search, a new

profile used as query.

It is VERY important to note that a strong E value in PSI-BLAST does NOT confirm homology. The E value is to the PSSM, and if there is any weakness in the PSSM, then unusual results can occur. PSI-BLAST finds **putative** homologs. Weakness in a PSSM can occur if a non-homolog accidentally gets into the sequences used to build the PSSM, or if errors in the multiple sequence alignment.

Discussion topic: How would you assess homology in putative homologs found by PSI-BLAST?

Delta-BLAST

The newest BLAST variation debuted in 2012 and is called DELTA-BLAST. The DELTA stands for "Domain enhanced lookup time accelerated BLAST." Yeah, they tried really hard to give it a Greek letter acronym. The idea is pretty simple. PSI-BLAST builds PSSMs after each iteration. Delta-BLAST searches a pre-existing PSSM library. The preexisting PSSMs come from the NCBI conserved domain database. In other words, these are PSSMs that are built from proteins that share a particular protein domain.

Discussion topic: In your opinion, what are the pluses and minuses for Delta-BLAST compared to PSI-BLAST?