



Need help?

www.ebi.ac.uk/Tools/sss/

Support: www.ebi.ac.uk/support/

EMBL-EBI

Wellcome Trust Genome Campus
Cambridge
CB10 1SD, UK

Phone: +44 (0) 1223 494444

Twitter: @emblembl

Further reading

Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acid Research* 38: W695-9 (2010).

Support

European Commission under FELICS; core funding - European Molecular Biology Laboratory; European Patent Office.

Sequence Searching at EBI

www.ebi.ac.uk/Tools/sss/

With sequence databases increasing in size, it is often difficult to find the sequences you need. Different sequence-based search tools have been developed over the years, each with its own advantages and disadvantages. No one tool can do everything, therefore selecting the right tool for the job is important. The European Bioinformatics Institute provides a broad range of search tools you can use to search a comprehensive collection of sequence databases.

Which search tool do I use?

Choosing the tool that best meets your search needs means your results will be more relevant. The choice of search algorithm depends on the size of the database you wish to search, the length of your query sequence, and what type of information you wish to find. Below is a table of the different search tools available at EBI and how they work, so you can better choose the right tool for the job you want done.

| Search tool | Best use | Alignment | Cautions |
|----------------------------------------|--------------------------------------------------------------------------------------------|-----------|-----------------------------------------------------------------------------------------------------|
| NCBI BLAST (*suite of tools) | Searching protein and nucleotide databases. | Query | Better for protein searches than nucleotide searches. Not as good for very short query sequences. |
| WU-BLAST (*suite of tools) | Searching protein and nucleotide databases. | Query | Better for protein searches than nucleotide searches. Not as good for very short query sequences. |
| PSI-BLAST | Finding remote homologues . | Query | Always check sequences to be included in alignments before starting the next round of searching. |
| FASTA (*suite of tools) | Searching protein and nucleotide databases. | Query | Not as good for very short query sequences. |
| SSEARCH | Rigorous search for smaller databases. Good for short sequences . | Query | Slow when querying large databases. |
| PSI-SEARCH | Finding remote homologues . Uses rigorous SSEARCH. | Query | Always check sequences to be included in alignments before starting the next round of searching. |
| GGSEARCH | Finding sequences similar in length to the query sequence (global-global matches). | Query | Limits its search to matches that are between 80-125% of the length of the query sequence. |
| GLSEARCH | Matching the full-length query sequence to longer sequences (global-local matches). | Query | Limits its search to target sequences that are longer than 80% of the length of the query sequence. |
| FASTM | Matching a set of short peptide or nucleotide sequences to longer sequences. | Query | Sequences can match query peptides in any order. |
| FASTF | Specialised FASTM for ordered peptides (Edman sequencing). | Query | Limits search to sequences matching the order of the query peptides. Specialised statistics. |
| FASTS | Specialised FASTM for unordered peptides (mass spectrometry sequencing). | Query | Sequences can match query peptides in any order. Specialised statistics. |

*The BLAST and FASTA programs have different search tools to choose from:

| NCBI BLAST | WU-BLAST equivalent | FASTA equivalent | Best use |
|------------|---------------------|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| BLASTP | BLASTP | FASTA | Searches a protein database with a protein query sequence. |
| BLASTN | BLASTN | FASTA | Searches a nucleotide database with a nucleotide query sequence. |
| BLASTX | BLASTX | FASTX | Searches a protein database with a nucleotide query sequence (6-frame translations). |
| - | - | FASTY | Searches a protein database with a nucleotide query sequence (6-frame translations; allows frameshifts within codons). Slower than FASTX. |
| TBLASTN | TBLASTN | TFASTX | Searches a nucleotide database (6-frame translations) with a protein sequence. Very slow. |
| - | - | TFASTY | Searches a nucleotide database (6-frame translations; allows frameshifts within codons) with a protein sequence. Slower than TFASTX. |
| TBLASTX | TBLASTX | - | Searches a nucleotide database (6-frame translations) with a nucleotide sequence (6-frame translations). Extremely slow and cpu-intensive. |

TOP TIP: BLAST, FASTA and SSEARCH look for regions of **local** alignment between the query and target sequences; therefore **always check alignments** to see where the two sequences align.

What databases are available to search?

Using the EBI website, the above tools can be used to search a number of different databases, including the **European Nucleotide Archive**, **UniProt**, as well as a variety of specialty databases. There are two points to consider when choosing a database to search: whether to search a protein or a nucleotide database, and whether to search the entire database or just part of it.

TOP TIP: When possible, search a protein sequence database over a nucleotide sequence one.

Protein searches are more sensitive than nucleotide ones, because they take account of both codon degeneracy and the similarity between the substituted amino acids. By contrast, nucleotide searches only score match/mismatch. As a result, protein sequence searches can find homologues 5-10 times further back in evolution than nucleotide sequence searches.

TOP TIP: Search the smallest database that you think will have the sequences you want.

Size matters. The larger the database, the less significant the matches. This is often a problem when querying a large database with very short sequences: you may fail to get the matches you expect simply because they are above threshold. If you can, try searching only part of the database; for example, if you are interested in sequences of a specific taxonomy, restrict the search to that taxonomic division. With both ENA and UniProt, there is the option of restricting the search to subsections of these databases.

Do I need to change parameters?

The parameters are set for searching an average length protein or coding sequence. If you wish to search using a short query sequence, then you will need to change the parameters you use, especially the matrix (or match/mismatch for nucleotide queries) and the gap penalties. The table on the right gives the best parameters to use based on the length of your query sequence:

| Query sequence length | Matrix | Gap open | Gap extend |
|-----------------------|----------|----------|------------|
| >300 | BLOSUM50 | -10 | -2 |
| 85-300 | BLOSUM62 | -7 | -1 |
| 50-85 | BLOSUM80 | -16 | -4 |
| >300 | PAM250 | -10 | -2 |
| 85-300 | PAM120 | -16 | -4 |
| 35-85 | MDM40 | -12 | -2 |
| < = 35 | MDM20 | -22 | -4 |
| < = 10 | MDM10 | -23 | -4 |

Enhancing your results

Sequence search results using EBI tools include links to multiple resources, which provide extra biological information about the sequences. Database links include sequence databases, genomic information, gene expression, ontologies, molecular interactions, reactions & pathways, enzymatic information, protein classification structural databases and literature.

In addition, functional predictions are provided for protein sequence matches. Using the 'Functional Predictions' tab, you are able to view the **InterPro** protein signature matches for each UniProt protein sequence in your search results. These results are identical to those obtained using **InterProScan** software. Protein signatures predict family classification and domain annotation. Signatures are also useful for identifying partial matches and potential mismatches, as well as splice variants.