

Basic Local Alignment Search Tool (BLAST) <http://www.ncbi.nlm.nih.gov/BLAST/>

Getting started:

The NCBI web site offers integrated access to databases including MEDLINE, GenBank, OMIM, and other molecular biology resources. The Entrez interface to GenBank is the same as the PubMed interface to MEDLINE, except for slight differences relating to the type of information stored in each database. While MEDLINE contains bibliographic citations, GenBank contains nucleic acid (DNA & RNA) sequences. GenBank is also linked to a protein sequence database (GenPept), a molecular structures database (MMDB) and gene locus information.

However, Entrez only offers keyword and field searching which is not an appropriate method for determining sequence similarity. Such searching requires a more complex software program (algorithm). The most utilized sequence similarity program is called BLAST®. BLAST(Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships.

Using BLAST to Search GenBank: Click on 'BLAST' in the NCBI menu:

The screenshot shows the NCBI website interface. At the top, the NCBI logo and the text "National Center for Biotechnology Information" are visible. Below this, a navigation bar contains links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The BLAST link is circled in red. Below the navigation bar, there is a search box with "GenBank" selected in a dropdown menu and a "Go" button. To the left, a "SITE MAP" section includes links for "About NCBI" and "What does NCBI do?". A red arrow points from the "What does NCBI do?" link to a detailed menu for BLAST. This menu lists various BLAST programs under "Nucleotide" and "Translated" categories. The "Nucleotide" section includes Discontiguous megablast, Megablast, Nucleotide-nucleotide BLAST (blastn), and Search for short, nearly exact matches. The "Translated" section includes Translated query vs. protein database (blastx), Protein query vs. translated database (tblastn), and Translated query vs. translated database (tblastx). The "Protein" section includes Protein-protein BLAST (blastp), PHI- and PSI-BLAST, Search for short, nearly exact matches, Search the conserved domain database (rpsblast), and Search by domain architecture (cdart). The "Genomes" section includes Human, mouse, rat; Fugu rubripes, zebrafish; Insects, nematodes, plants, yeasts, malaria; and Microbial genomes, other eukaryotic genomes.

There are a variety of BLAST programs to choose from depending on various factors such as the type of sequence and database (nucleic acid, protein) you wish to search, the size of your sequence, and whether you require a protein translation of a nucleotide sequence or database.

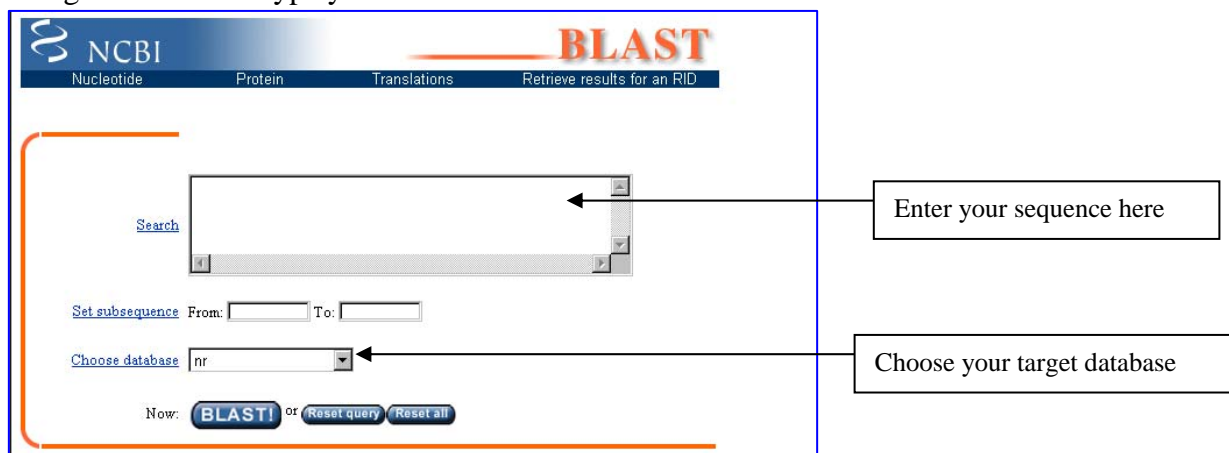
BLAST Programs

- Nucleotide: Standard, MEGABLAST, short matches
- Protein: Standard, PSI/PHI-BLAST, short matches
- Translated BLAST searches: blastx, tblastn, tblastx

- Conserved Domain Searches
- BLAST2Sequences
- Genomic BLAST pages
- Specialized BLAST pages: VecScreen, IgBLAST, Trace BLAST

For example, choose “nucleotide BLAST” if you wish to compare a nucleotide sequence to a nucleotide database. Within the nucleotide BLAST choices, you can use the standard search, the MEGABLAST search, or a short sequence search. Other specialized BLAST searches include translational searches, a conserved protein domain search, and searches against specific genomes or types of sequences. Use the "?" links to learn more about the different search options.

Once you have chosen a BLAST search, you will see a screen that allows you to input your target sequence and to select certain search parameters. Many of the parameters will be pre-set for you, depending on the search type you chose.



First enter your target sequence or an accession number or gi identifier. The sequence can be entered either as plain sequence or in a FASTA format. The FASTA format begins with a comment line (starting with the ">" character) that you can utilize to identify your sequence.

Next, it is necessary to choose the database to be searched with your query sequence. Click on the hypertext link “*Choose Database*” to see a list of available databases. These include both protein and nucleic acid databases of various types. Here are some examples:

<i>Database</i>	<i>Contains</i>	<i>Type</i>
nr - protein	non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF (Note: no longer truly non-redundant, but name of db not changed)	aa
nr -nucleotide	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). Also no longer "non-redundant".	nt
swissprot	SWISS-PROT	aa
Drosophila genome	Drosophila genome: protein OR DNA sequences	Both
Kabat	Kabat's database of immunological sequences	Both
dbest	GenBank+EMBL+DDBJ sequences from EST Division	nt
Vector	Vector subset of GenBank	nt
Mito	mitochondrial sequences	nt

Select the desired database using the drop-down menu. Note that the program and the database selected must correspond. A protein search must be conducted on a protein database; likewise, a nucleic acid search must be conducted on a nucleic acid database.

Scroll down the page to see additional options for refining your search. The default options will vary depending on your initial choice of type of BLAST search. It is usually best to rely upon the defaults for a first try and then modify them if you are not satisfied with your results. All of the

Options for advanced blasting

[Limit by entrez query](#) or select from: (none)

[Choose filter](#) Low complexity Human repeats Mask for lookup table only Mask lower case

[Expect](#)

[Other advanced](#)

blue underlined links can be used to learn more about the available options. Note that you have a dropdown menu that can be used to specify organism and that you can limit the BLAST search by any Entrez query term and/or field code combination.

Further down this page, you will find options for adjusting the display of the BLAST search results.

Format

Show [Graphical Overview](#) [NCBI-gi Alignment](#) in [format](#)

Number of: [Descriptions](#) [Alignments](#)

[Alignment view](#)

[Autoformat](#)

Alignment view options can be seen in a dropdown menu and include pairwise, query-anchored, and flat query-anchored. The default pairwise alignment shows you an alignment for each individual match to your sequence. The others show a grouped alignment. Remember that **all** BLAST alignments are pairwise, so even the grouped formats are actually showing a series of pairwise alignments and not a true multiple sequence alignments. Other programs must be used to create multiple sequence alignments.

Once you have set all of the options, click on the 'BLAST' button to conduct the search. You will then see a page generated which displays your description line and gives you a search request ID #. You may want to note that number down as it can be used to retrieve your results later, if necessary.

NCBI **Advanced BLAST** BLAST Entrez ?

Your request has been successfully submitted and put into the Blast Queue.

Query= gi|532319|pir|TVFV2E|TVFV2E envelope protein
(243 letters)

You are using a new system that allows users to retrieve results at their convenience and format their results multiple times with different formatting options. This system also allows the NCBI to more efficiently use computational resources, better serving the community. A [description](#) of the queuing system is available.

The request ID is

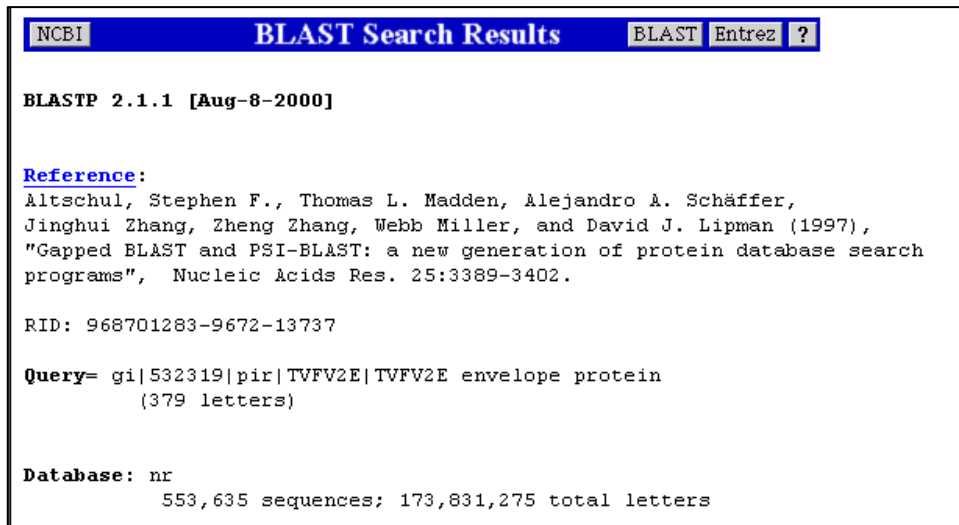
The results are estimated to be ready in **0 min 43 sec** but may be done sooner.
Please press "Format results" when you wish to check your results.

Also, this page will give you an estimated finish time for your search. The time required to complete a search depends on several parameters including the length of your query sequence, the database and program you have chosen, and the number of searches being conducted simultaneously.

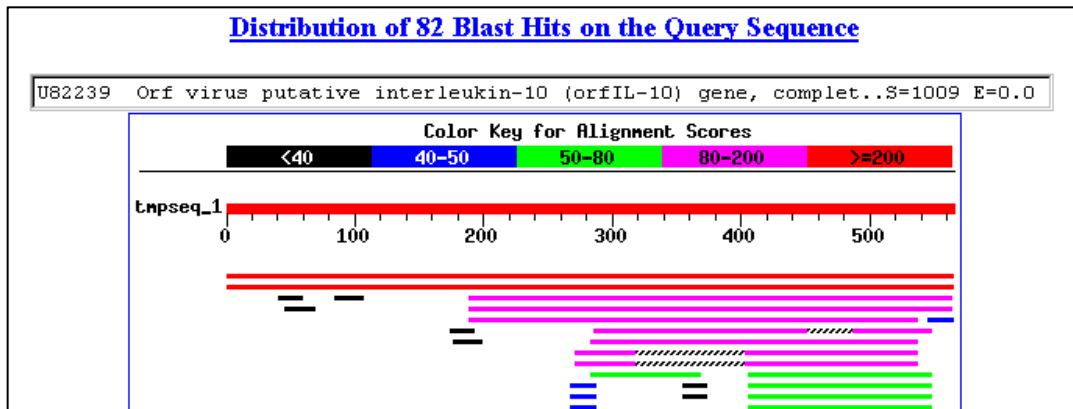
At any time, you may click on the 'format results' button to see if your results are ready. If they are not, you will see a new browser window that says:



Repeat this until your results are displayed in a new browser window. The top of that page will look like this:



It will list for you the request ID number and your query description line. Also, it will display the name and content information for the database you searched. Scrolling further down the page will show you the search matches in graphical format, in a list ordered by match score.



The graphical display shows the matches to your query sequence using colored bars. The longer the bar, the more complete the match. The color is a representation of the match score. Holding the mouse cursor over one of the bars will display the definition line for that sequence in the text box above the graph.

Below this graphical display, all of the matches will be listed in order from highest match score to lowest score.

Sequences producing significant alignments:		Score (bits)	E Value
gb U60552.1 OVU60552	Orf virus putative interleukin-10 (orf...	1096	0.0
gb U82239.1 OVU82239	Orf virus putative interleukin-10 (orf...	1009	0.0
emb Z29362.1 OAINLE10	O.aries mRNA for interleukin 10	141	4e-31
gb U11421.1 OAU11421	Ovis aries interleukin 10 (IL-10) mRNA...	133	9e-29
gb U11767.1 CEU11767	Cervus elaphus interleukin-10 (IL-10),...	98	5e-18
gb U00799.1 U00799	Bos taurus charolais interleukin-10 mRNA...	98	5e-18
gb U38200.1 ECU38200	Equus caballus interleukin-10 (IL-10) ...	86	2e-14
ref NC_001650.1 	Equine herpesvirus 2, complete genome	84	7e-14
gb U20824.1 EHVU20824	Equine herpesvirus 2, complete genome	84	7e-14
ref NC_001345.1 	Human herpesvirus 4, complete genome	78	5e-12
emb V01555.1 EBV	Epstein-Barr virus (EBV) genome, strain B95-8	78	5e-12

There are four pieces of information given in the ordered list:

- **database accession number**
- **description of sequence**
- **match score**
- **E value**

The **database accession number** is provided as a hypertext link that will bring up the GenBank record for that sequence. The **description** is simply the definition line from the GenBank record. The **match score** represents a mathematical summation of base pair matches, mismatches and gap penalties, based on the BLAST algorithms and matrices. The higher the score, the better the match. The **E value** is the 'Expect Value', ie the probability that this match occurred by random chance and is, thus, not significant. The lower the E value, therefore, the better the match. The best matches will show an E value of 0 or very close to 0. The E values are displayed in base ten exponents. Read the E value of "5e-18" as 5×10^{-18} .

Below the ordered list, the sequence alignment will be displayed. If you chose the default pairwise alignment, each individual match will be given in complete sequence detail. Again, the best matches are displayed first.

```
>gb|U82239.1|OVU82239 Orf virus putative interleukin-10 (orfIL-10) gene, complete cds
Length = 778

Score = 1009 bits (509), Expect = 0.0
Identities = 552/565 (97%), Gaps = 1/565 (0%)
Strand = Plus / Plus

Query: 1 atgtcgaagaacaaaattctggtgtgtgttggaattattcttacttatacattatacaca 60
      |||
Sbjct: 108 atgtcgaagaacaaaattctggtgtgtgttgcgattattcttacttatacattatacaca 167

Query: 61 gatgcgctattgtgttgagtatgaggaaagtgaggaagataaacaacagtcggtagtagt 120
      |||
Sbjct: 168 gatgcgctattgtgttgagtattagaaagtggggaagatgaacaacagtcggtagtagt 227

Query: 121 agtaatttctctgcgagtttaccgcacatgcttagagaactcagggcagcgttcggaaag 180
      |||
Sbjct: 228 agtaatttctctgcgagtttaccgcacatgctcagagaactcagggcagcgttcggaaag 287

Query: 181 gtaaaaacttttctccagatgaaagaccaactgaacagtagctactcaccacagtcgct 240
      |||
Sbjct: 288 gtaaaaactt-tcttccagatgaaagaccaactgaacagtagctactcaccacagtcgct 346
```

The results can be printed or selected and copy/pasted into another program such as a word processor. The links to the database records for each sequence match allow further examination

1.5-5X that of transversion mutations (ie, purine to pyrimidine). (See Wakely, *Mol Biol Evol* 11(3):436-42, 1994)

Protein BLAST

When comparing amino acid sequences, it is vital to consider the similarities and differences between any two amino acids, the level of evolutionary conservation, and the frequency that a particular substitution occurs within or between species. Scientists have constructed several different scoring matrices for amino acid sequence comparisons based on these factors. The most common two types of scoring matrices are PAM and BLOSUM.

Percent Accepted Mutation

PAM is a unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. A 1.0 PAM unit is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins [Info from the NCBI Glossary – PAM definition].

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	5																								
A	1	2																							
V	-1	0	4																						
L	-4	-2	2	6																					
I	-3	-1	4	2	5																				
P	0	1	-1	-3	-2	6																			
S	1	1	-1	-3	-1	1	2																		
T	0	1	0	-2	0	0	1	3																	
D	1	0	-2	-4	-2	-1	0	0	4																
E	0	0	-2	-3	-2	-1	0	0	3	4															
N	0	0	-2	-3	-2	0	1	0	2	1	2														
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4													
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5												
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6											
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6										
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9									
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10								
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17							
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6						
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12					
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3			
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1		
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1	*

PAM 250

BLOSUM - Blocks Substitution Matrix.

A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. [Info from the NCBI Glossary – BLOSUM definition].

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	6																								
A	0	4																							
V	-3	0	4																						
L	-4	-1	1	4																					
I	-4	-1	3	2	4																				
P	-2	-1	-2	-3	-3	7																			
S	0	1	-2	-2	-2	-1	4																		
T	-2	0	0	-1	-1	-1	1	5																	
D	-1	-2	-3	-4	-3	-1	0	-1	6																
E	-2	-1	-2	-3	-3	-1	0	-1	2	5															
N	0	-2	-3	-3	-3	-2	1	0	1	0	6														
Q	-2	-1	-2	-2	-3	-1	0	-1	0	2	0	5													
K	-2	-1	-2	-2	-3	-1	0	-1	-1	1	0	1	5												
R	-2	-1	-3	-2	-3	-2	-1	-1	-2	0	0	1	2	3											
H	-2	-2	-3	-3	-3	-2	-1	-2	-1	0	1	0	-1	0	8										
F	-3	-2	-1	0	0	-4	-2	-2	-3	-3	-3	-3	-3	-1	6										
Y	-3	-2	-1	-1	-1	-3	-2	-2	-3	-2	-2	-1	-2	-2	2	3	7								
W	-2	-3	-3	-2	-3	-4	-3	-2	-4	-3	-4	-2	-3	-3	-2	1	2	11							
M	-3	-1	1	2	1	-2	-1	-1	-3	-2	-2	0	-1	-1	-2	0	-1	-1	5						
C	-3	0	-1	-1	-1	-3	-1	-1	-3	-4	-3	-3	-3	-3	-2	-2	-1	9							
B	-1	-2	-3	-4	-3	-2	0	-1	4	1	3	0	0	-1	0	-3	-3	-4	-3	-3	4				
Z	-2	-1	-2	-3	-3	-1	0	-1	1	4	0	3	1	0	0	-3	-2	-3	-1	-3	1	4			
X	-1	0	-1	-1	-1	-2	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-1	-1		
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

BLOSUM 62

About Choosing a Matrix

- BLOSUM matrices tend to be more sensitive to distant relationships than PAM.
- BLOSUM tends to give
 - higher scores to hydrophilic amino acids substitutions
 - lower scores to hydrophobic amino acids substitutions
- Substitutions of rare amino acids are more tolerated by BLOSUM.
- BLOSUM is generally preferred BUT it is generally better to perform multiple searches with various matrix choices at first

General Rules for Choosing a Matrix

- Use higher PAM or lower BLOSUM matrices for more divergent sequences.
- Use lower PAM or higher BLOSUM matrices for more closely related sequences.
- BLOSUM62 is usual default choice

Accounting for Gaps

- Alignment will be more accurate
- Alignments will reflect biological relationships more closely
- Distinct local regions of similarity will not be diluted or missed

```

GMILVKDSKTNQLVPEVLEYNV
G++++ + + F+V+E+N
GLMIMPNGQ-----PKVIEFNC
  
```

The scoring for gaps in BLAST must reflect both that gaps can be frequent and may be long and that the introduction of gaps can introduce errors in alignments. Therefore, gap penalties include large existence (establishment) penalties and smaller extension penalties. These penalties can be adjusted by the user within limits.