

BLAST Program Selection Guide

By blast-help group, NCBI User Service
NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894

Table of Content

1. [Introduction](#)
2. [BLAST Database Content](#)
3. [Program Selection Table](#)
4. [Explanation for the program choices given in Tables 3.1 and 3.2](#)
5. [Explanation for the program choices given in Tables 3.3](#)
6. [Explanation on Special Purpose Pages](#)
7. [Appendices](#)

1. Introduction

NCBI has provided BLAST sequence analysis services for over a decade. For many users, the first question they face is *"Which BLAST program should I use?"*

In order to help users arrive at an answer to this question, we have constructed this table called the "BLAST Program Selection Guide." It is divided into several categories according to the *nature* and *size* of the query and the primary goal of the search. Starting from the query sequence on the left and cross-referencing to the right, an user will arrive the specific BLAST program best suited for that search.

This document is also available in [PDF](#) (1056656 bytes).

2. BLAST Database Content

To discuss BLAST program selection, we first need to know what databases are available and what sequences they contain. Here we will take a look at the common BLAST databases. According to their content, they are grouped into nucleotide and protein databases. These databases and their detailed compositions are listed in the two tables below.

NCBI also provides specialized BLAST databases such as the vector screening database, variety of genome databases for different organisms, and trace databases. The database names are being standardized. Their contents are also described here. This is mostly for the three model organisms, i.e. human, mouse, and rat. For other organisms, the content of their genome blast pages will be listed when the their special BLAST pages are discussed.

Table 2.1 Content of Protein Sequence Databases

Database	Content Description
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
refseq NEW	Protein sequences from NCBI reference sequence project.
swissprot	Last major release of the SWISS-PROT protein sequence database (no incremental updates).
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank.
env_nr NEW	Non-redundant CDS translations from env_nt entries.

[\[Back to top\]](#)

Table 2.2 Nucleotide Databases for BLAST

Database	Content Description
nr	All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" due to computational cost.
refseq_mrna NEW	mRNA sequences from NCBI Reference Sequence Project.
refseq_genomic NEW	Genomic sequences from NCBI Reference Sequence Project.
est	Database of GenBank + EMBL + DDBJ sequences from EST division.
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, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
pat	Nucleotides from the Patent division of GenBank.

pdb	Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are NOT the coding sequences for the corresponding proteins found in the same PDB record.
month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
alu_repeats	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
chromosome	Complete genomes and complete chromosomes from the NCBI Reference Sequence project. It overlaps with refseq_genomic.
wgs	Assemblies of Whole Genome Shotgun sequences.
env_nt NEW	Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarso Sea project. This does overlap with nucleotide nr.

[\[Back to top\]](#)

Table 2.3 Genome BLAST Databases and Contents

Database	Description
curated NT contigs	Assemblies of Finished Mouse BAC clones. These are annotated with SNPs and STSs and available in GenBank
genome (all assemblies)*	This database represents the current public build of the genome. The sequences in this database will have RefSeq accession numbers or type NT_?????? or NW_?????? and these represent either contigs (from a clone based assembly) or supercontigs (from a whole genome shotgun or composite assembly). The contigs in this database are from both the reference assembly and any alternate assemblies available for the genome. This database is generated at the time of a genome release.
genome (reference only)	This database represents the current public build of the genome. The sequences in this database will have RefSeq accession numbers or type NT_?????? or NW_?????? and these represent either contigs (from a clone based assembly) or supercontigs (from a whole genome shotgun or composite assembly). The contigs in this database are from only the reference assembly. This database is generated at the time of a genome release.

HTGS	This databases is a collection of all sequences in GenBank that have an HTG keyword. This allows users to search htgs_phase3 sequences (normally found in NR) and htgs_phase0, 1 and 2 sequences (normally found in HTGS) at the same time
RefSeq RNA	Collection of reference mRNAs generated by the NCBI RefSeq project . This database is generated daily.
RefSeq protein	Collection of reference proteins generated by the NCBI RefSeq project .This database is generated daily
Build RNA	Collection of reference mRNAs generated by NCBI as part of the genome annotation pipeline. This database is generated at the time of a genome release.
Build protein	Collection of reference proteins generated by NCBI as part of the genome annotation pipeline. This database is generated at the time of a genome release.
Ab Initio RNA	Collection of <i>ab initio</i> RNA predictions generated by NCBI as part of the genome annotation pipeline. This database is generated at the time of a genome release.
Ab Initio protein	Collection of <i>ab initio</i> protein predictions generated by NCBI as part of the genome annotation pipeline. This database is generated at the time of a genome release.
ESTs	Single pass sequence reads from cDNA libraries. This database is updated daily.
BAC ends	The end sequences of BAC clones. This database is generated daily
Traces-WGS	All of the raw organism WGS traces. This database is updated as needed.
Traces-ESTs	All of the raw organism EST traces. This database is updated as needed.
Traces-other	All of the raw organism non-WGS and non-EST traces. This database is updated as needed.
WGS contigs	If an organism was assembled using a whole genome shotgun (WGS) strategy, this database is available (if the WGS assembly is in GenBank). This database is updated as needed.
Gene Trap Clones (Mouse Only)	A collection of sequences generated by performing Gene Trap insertions. This database is updated weekly.
Reference Dog Assembly (boxer)	The supercontigs from the Whole Genome Shotgun (WGS) assembly from a 7.6X coverage whole genome library. This assembly was performed at the Broad Institute using the Arachne assembler

Celera Dog Assembly (Poodle)	This database is a collection of the Whole Genome Shotgun (WGS) contigs assembled from a 1.5X coverage whole genome library. A description of this assembly can be found in Kirkness et al (2003) .
Celera Dog Extra (Poodle)	This database a collection of Whole Genome Shotgun (WGS) reads that were not assembled into contigs (the Celera Dog Assembly). A description of the assembly can be found in Kirkness et al (2003) .
Ref Chimp Assembly	This database a collection of Whole Genome Shotgun (WGS) contig assembled using the program Arachne. These contigs were assembled from a 4.5X coverage set of WGS reads. A publication describing this data should occur in early 2004
Alt Chimp Assembly	This database a collection of Whole Genome Shotgun (WGS) contig assembled using the program PCAP. These contigs were assembled from a 4.5X coverage set of WGS reads. A publication describing this data should occur in early 2004
Celera CSA	Celera January 2001 compartmental shotgun assembly (CSA) of the human genome. It was generated from the 27 million reads of Celera's 5.3X whole genome shotgun data and 16 million 'reads' of shredded GenBank data from other human genome projects (Nature 2001. 409:860-921). It was generated by the Celera Assembler applied to 3800 separate compartments of Celera and GenBank data associated by inferred sequence overlaps and Celera read pairs. It relied on Celera's paired reads and the BAC end reads for long range order and orientation. See Istrail et al (2004) .
Celera cWGA	This is the November 2000 combined whole genome shotgun assembly (WGA) of the human genome. It was generated by the Celera Assembler applied to the 27 million reads of Celera's 5.3X whole genome shotgun data and 16 million 'reads' of shredded GenBank data from other human genome projects (Nature 2001. 409:860-921). It relied on Celera's paired reads and BAC end reads from GenBank for long range order and orientation. See Istrail et al (2004) .
Celera WGA	This is the December 2001 whole genome shotgun assembly (WGSA) of the human genome. It was generated by the Celera Assembler applied to shotgun data only: the 27 million reads of Celera's 5.3X whole genome shotgun data and 104,000 BAC end sequence pairs from GenBank from other human genome projects (Nature 1996. 381:364-366 ; Genomics 2000. 63:321-332). It relied on Celera's paired reads and the BAC end reads for long range order and orientation. See Istrail et al (2004) .

hsc_tcag	The Hospital for Sick Children Center for Applied Genomics assembly of Human Chromosome 7. This is a combination of WGS sequence data generated at Celera and HTGS sequence generated by the Human Genome Sequencing Consortium. An analysis of this assembly was published by Scherer et al (2003) .
----------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Note

¹ Table content is derived from <http://www.ncbi.nlm.nih.gov/genome/seq/Database.html>.

² Database nomenclature is being standardised for genome blast pages.

³ This table is most relevant to human, mouse, and rat genome BLAST pages.

[\[Back to top\]](#)

3. Program Selection Tables

The appropriate selection of a BLAST program for a given search is influenced by the following three factors 1) the nature of the query, 2) the purpose of the search, and 3) the database intended as the target of the search. The following tables provide recommendations on how to make this selection.

Table 3.1 Program Selection for Nucleotide Queries

Length ¹	Database	Purpose	Program	Explanation
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast , megablast , or blastn	Learn more ...
		Find sequences similar to query sequence	discontiguous megablast or blastn	Learn more ...
		Find similar sequence from the Trace archive	Trace megablast , or Trace discontiguous megablast	Learn more ...
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)	Learn more ...
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)	Learn more ...

7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches	Learn more ...
-----------	----------------------------	----------------------------------------------------------	--------------------------------------------------------	--------------------------------

¹ The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST.

[\[Back to top\]](#)

Table 3.2 Program Selection for Protein Queries

Length ¹	Database	Purpose	Program	Explanation
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)	Learn more ...
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST	Learn more ...
		Find proteins similar to the query around a given pattern	PHI-BLAST	Learn more ...
		Find conserved domains in the query	CD-search (RPS-BLAST)	Learn more ...
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)	Learn more ...
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)	Learn more ...
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches	Learn more ...

¹ The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

[\[Back to top\]](#)

As genomic and other specialized sequence information is made available to the public, NCBI creates specialized BLAST pages for those sequences. The table below provides a general guide on how to select and use those special BLAST databases.

Table 3.3 Search against Organism Specific or Genome Databases ¹

Query ²	Database	Purpose	Program Pages to Use ³	Explanation	
Nucleotide: 20 or 28 bp and above	Human Genome	Map the query sequence	Human	Learn more ...	
	Mouse Genome		Mouse	Learn more ...	
	Rat Genome		Rat	Learn more ...	
	Chicken, Cow, or Dog Genome		Chicken , Cow , or Dog	Learn more ...	
	Cat, Sheep, or Pig Genome		Cat , Sheep , or Pig	Learn more ...	
	Fugu (Pufferfish)		Fugu rubripes	Learn more ...	
	Protein: 15 residues and above	Zebrafish	Determine the genomic structure	Zebrafish	Learn more ...
		Insects (flies and honeybees)	Identify novel genes	Insects	Learn more ...
		Nematodes (worms)		Nematodes	Learn more ...
		Plants	Find homologs	Plants	Learn more ...
		Fungi Genomes (including yeasts)	Other data mining	Fungi	Learn more ...
		Malaria		Malaria	Learn more ...
Environmental Samples	Environmental Samples	Learn more ...			
Other Lower Eukaryotic Genomes	Other eukaryotes genomes	Learn more ...			
	Microbial Genomes		Microbial genomes	Learn more ...	

¹ Those pages access the genome database consisting of contig assemblies and other sequences specific to the organisms. Not all organisms listed here have genome assemblies available.

² Sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable searches with a short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, **-F F -e 20000 -W 2** should be used. For nucleotide comparison, use **-F F -e 1000 -W 7**. This also requires the unchecking of the megablast checkbox.

³ Available databases and their contents are described in Section 5.

[\[Back to top\]](#)

BLAST pages for special purposes are listed under Special and Meta sections. Their functions are described in Table 3.4 below.

Table 3.4 Function of Special BLAST Pages under Special/Meta Sections

Query ¹	Database	Purpose	Program Pages to Use	Explanation
Nucleotide: 11 bp or above	- ²	Compare two sequences directly	Align two sequences	Learn more ...
Protein: 15 or above	Immunoglobulin sequences	Find matches to curated immunoglobulin sequences	igBLAST	Learn more ...
Nucleotide: 20 or 28 bp and above	UniVec	Screen for vector contamination	VecScreen	Learn more ...
	GEO NEW	Find matches to sequences with MicroArray information	GEO BLAST	Learn more ...
	human SNP	Find matches to human reference SNPs	SNP BLAST	Learn more ...
-	Retrieve Results	To retrieve results for a search with its RID	Retrieve result for an RID	Learn more ...

¹ The query sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable better handling of short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, **-F F -e 20000 -W 2** should be used. For nucleotide comparison, use **-F F -e 2000 -W 7**.

² "Align two sequences" treats the second sequence as the database.

NOTE:

GenBank® and **BLAST®** are registered trademarks granted to NCBI by USPTO.

For questions and suggestions about BLAST, please write to: blast-help@ncbi.nlm.nih.gov

For general questions about NCBI resources, please write to: info@ncbi.nlm.nih.gov

NCBI User Services can also be reached by phone at:(301)496-2475.

[\[Back to top\]](#)

4. Explanation for the program choices given in Tables 3.1 and 3.2

4.1 MEGABLAST is the tool of choice to identify a nucleotide sequence.

The best way to identify an unknown sequence is to see if that sequence already exists in a public database. If the database sequence is a well-characterized sequence, then one will have access to a wealth of biological information. MEGABLAST, discontinuous-megablast, and blastn all can be used to accomplish this goal. However, MEGABLAST is specifically designed to efficiently find long alignments between very similar sequences and thus is the best tool to use to find the identical match to your query sequence. In addition to the expect value significance cut-off, MEGABLAST also provides an adjustable percent identity cut-off for the alignment, which provides cut-off in addition to the significance cut-off threshold set by Expect value.

Web MEGABLAST and discontinuous megablast pages can also accept batch queries, the only web BLAST pages with this capability. Please refer to the "[Batch Search](#)" section for details.

[\[Back to top\]](#)

4.2 Discontinuous MEGABLAST is better at finding nucleotide sequences similar, but not identical, to your nucleotide query.

The BLAST nucleotide algorithm finds similar sequences by breaking the query into short subsequences called words. The program identifies the exact matches to the query words first (word hits). BLAST program then extends these word hits in multiple steps to generate the final gapped alignments.

One of the important parameters governing the sensitivity of BLAST searches is the length of the initial words, or word size as it is called. The most important reason that blastn is more sensitive than MEGABLAST is that it uses a shorter default word size (11). Because of this, blastn is better than MEGABLAST at finding alignments to related nucleotide sequences from other organisms. The word size is adjustable in blastn and can be reduced from the default value to a minimum of 7 to increase search sensitivity.

The search sensitivity can further improved by using the newly introduced discontinuous megablast page. This page uses an algorithm with the same name, which is similar to that reported by [Ma et.al](#). Rather than requiring exact word matches as seeds for alignment extension, discontinuous megablast uses non-contiguous word within a longer window of template. In coding mode, the third base wobbling is taken into consideration by focusing on finding matches at the first and second codon positions while ignoring the mismatches in the third position. Searching in discontinuous MEGABLAST using the same word size is more sensitive and efficient than standard blastn using the same word size. For this reason, it is now the recommended tool for this type of search. Alternative non-coding patterns can also be specified if desired. Additional details on discontinuous are available at:

<http://www.ncbi.nlm.nih.gov/blast/discontinuous.html>

<http://www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html>

Parameters unique for discontinuous megablast are:

- word size: restricted to two options, i.e., 11 or 12
- template: only three options are available, 16, 18, or 21
- template type: coding (0), non-coding (1), or both (2)

It is important to point out that nucleotide-nucleotide searches are not the best method for finding homologous protein coding regions in other organisms. That task is better accomplished by performing searches at the protein level, by direct protein-protein BLAST searches or by translated BLAST searches. This is because of the codon degeneracy, the greater information available in amino acid sequence, and the more sophisticated algorithm and scoring matrix used in protein-protein BLAST.

[\[Back to top\]](#)

4.3 "Search for short nearly exact matches" is useful for primer or short nucleotide motif searches.

Short sequences (less than 20 bases) will often not find any significant matches to the database entries under the standard nucleotide-nucleotide BLAST settings. The usual reasons for this are that the significance threshold governed by the Expect value parameter is set too stringently and the default word size parameter is set too high.

You can adjust both the word size and the expect value on the standard BLAST pages to work with short sequences. However, we do provide a BLAST page with these values preset to give optimal results with short sequences. This page ("Search for short nearly exact matches") is linked under the nucleotide BLAST section of the main BLAST page.

Table 4.3.1 Parameter settings for standard blastn and "Search for short and nearly exact matches"

Program	Word Size	DUST Filter Setting	Expect Value
Standard blastn	11	On	10
Search for short nearly exact matches	7	Off	1000

A common use of this page is to check the specificity of PCR or hybridization. A useful way to check a pair of PCR primers is to first concatenate them by inserting string of 20 or more N's in between the two primers, and then search the concatenated pair as one sequence. Since BLAST looks for local alignments and automatically searches both strands, there is no need to reverse complement one of the primers before doing the concatenation or the search.

The query sequence should contain no ambiguous bases. Consensus motifs with [degenerate bases](#), such as AACNNNNNNRTAYG (StySQI recognition site) or TGGNNNNNNGCCAA (NF-1 binding site) will not work for this type of search.

[\[Back to top\]](#)

4.4 Use the Trace Archive BLAST page to search raw primary sequence trace files.

Trace data files are not official entries of the GenBank database and have no associated feature annotations. Despite this limitation, they are still a rich source of information, especially for organisms lacking a significant amount of regular mRNA or assembled genomic sequences. The sequences come from a variety of projects and sequencing strategies, including Whole Genome Shotgun (WGS), BAC end sequencing, and EST sequencing. The trace data are single pass sequencing reads not trimmed for quality or vector contamination. Their average lengths are between 500 to 700 bp.

A search against the Trace Archive can use MEGABLAST or discontinuous MEGABLAST. The former is better for identifying exact matches in intra-species searches, such as looking for extra mRNA sequences or the genomic counterparts for a given gene, while the latter is better for identifying similar coding sequences from different species. Information on the Trace Archive is available from the [Trace documentation page](#).

[\[Back top\]](#)

4.5 Standard protein BLAST is designed for protein searches.

Standard protein-protein BLAST (blastp) is used for both identifying a query amino acid sequence and for finding similar sequences in protein databases. Like other BLAST programs, blastp is designed to find local regions of similarity. When sequence similarity spans the whole sequence, blastp will also report a global alignment, which is the preferred result for protein identification purposes.

For clear result in identification search, try taking off both "low complexity filter". Unlike nucleotide BLAST, there is no comparable MEGABLAST for protein searches, so batch search via the web is not possible. To do batch protein BLAST, you can take a look at [netblast](#) (blastcl3). Document describing this tool is [netblast.txt](#).

4.6 PSI-BLAST is designed for more sensitive protein-protein similarity searches.

Position-Specific Iterated (PSI)-BLAST is the most sensitive BLAST program, making it useful for finding very distantly related proteins. Use PSI-BLAST when your standard protein-protein BLAST search either failed to find significant hits, or returned hits with descriptions such as "hypothetical protein" or "similar to...".

The first round of PSI-BLAST is a standard protein-protein BLAST search. The program builds a position-specific scoring matrix (PSSM or profile) from an alignment of the sequences returned with Expect values better (lower) than the inclusion threshold (default=0.005). The PSSM will be used to evaluate the alignment in the next iteration of search. Any new database hits below the inclusion threshold are included in the construction of the new PSSM. A PSI-BLAST search is said to have converged when no more new database sequences are added in subsequent iterations. You can add database hits that fall outside the inclusion threshold to your PSSM for the next round by checking the box next to the hit. Already selected hits can also be removed from the selection by uncheck the checkbox.

PSSM is query specific. You can save a PSSM created during a PSI-BLAST search of one database and use it to search a different database with the same query. To do this, change "Alignment" to "PSSM" in a pull-down menu in the Format section of a "Formatting BLAST" page (at any iteration after the first). Then format the search, copy the resulting ascii encoded PSSM and paste it into the PSSM window of a new PSI-BLAST search page.

Web PSI-BLAST cannot generate the PSSM in human readable form. You can use the -Q file option in standalone version (blastpgp) for this purpose. See [blast.txt](#) for more information.

[\[Back to top\]](#)

4.7 PHI-BLAST can do a restricted protein pattern search.

Pattern-Hit Initiated (PHI)-BLAST is designed to search for proteins that contain a pattern specified by the user AND are similar to the query sequence in the vicinity of the pattern. This dual requirement is intended to reduce the number of database hits that contain the pattern, but are likely to have no true homology to the query.

To run PHI-BLAST, enter your query (which contains one or more instances of the pattern) into the "Search" box, and enter your pattern into the "PHI pattern" box in the "Options" section of the page. Patterns must follow the syntax conventions of PROSITE. Only one pattern can be used in a given search. Pattern syntax is described [here](#).

The sample query sequence and a sample pattern in ProSite format are given below for test run with PHI-BLAST.

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGNFYVIERGTYDILVTKDNQTRSVGQYDNRGS
FGELALMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQVEEIARCHKGQYFGEALALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLQ
```

```
[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
```

[\[Back to top\]](#)

4.8 The protein "Search for short nearly exact matches" is optimized to find matches to a short peptide.

A short peptide (10-15mer or shorter) often will not find any significant matches to the database under the standard protein-protein BLAST settings. The usual reasons for this are that the significance threshold governed by the expect value parameter is set too stringently and the default word size parameter is set too high.

You could adjust both the word size and the expect value on the standard BLAST pages to make it work with short query sequences. NCBI provides a separate BLAST page with these values preset to optimize blastp searches with short query sequences. This page, "Search for short nearly exact matches", is available via a link under the Protein BLAST section of the BLAST home page. In addition, the more stringent PAM30 is used in lieu of BLOSUM62, and the composition-based statistics which takes the amino acid composition of the query sequence into account when calculating the score and significance of the alignments.

Composition based statistics takes the amino acid composition of the query and subject sequence into account when calculating the score and significance of the alignments. It can have a large effect on searches using queries with a biased amino acid composition. By definition, short peptides will have a biased compositions and should not be used with composition based statistics.

Due to the requirement that the query needs to be at least twice the word size, a query shorter than 5 residues is not recommended even though it can be as short as 4 residues when the word size is set to 2. In addition, since ambiguous residues break the query sequence, there should be no ambiguities in the query to ensure that the entire sequence can be used as seeds for the initial search.

Table 4.8.1 Parameter settings for standard blastp and "Search for short and nearly exact matches"				
Program	Word Size	SEG Filter	Expect Value	Score Matrix
Standard Protein Blast	3	On	10	BLOSUM62
Search for short and nearly exact matches	2	Off	20000	PAM30

For protein pattern search, one can take a look at the standalone tool "**seedtop**". This is described in [blast.txt](#). The program is part of the standalone blast package, or the blast initialed archive under <ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>.

[\[Back to top\]](#)

4.9 The "Nucleotide query - Protein db [blastx]" is useful for finding similar proteins to those encoded by a nucleotide query.

Translated BLAST services are useful when trying to find homologous proteins to a nucleotide coding region. Blastx compares translational products of the nucleotide query sequence to a protein database. Because blastx translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames, it is particularly useful when the reading frame of the query sequence is unknown or it contains errors that may lead to frame shifts or other coding errors. Thus blastx is often the first analysis performed with a newly determined nucleotide sequence and is used extensively in analyzing EST sequences. This search is more sensitive than nucleotide blast since the comparison is performed at the protein level.

4.10 The "Protein query - Translated db [tblastn]" search is useful for finding protein homologs in unannotated nucleotide data.

A tblastn search allows you to compare a protein sequence to the six-frame translations of a nucleotide database. It can be a very productive way of finding homologous protein coding regions in unannotated nucleotide sequences such as expressed sequence tags (ESTs) and draft genome records (HTG), located in the BLAST databases est and htgs, respectively.

ESTs are short, single-read cDNA sequences. They comprise the largest pool of sequence data for many organisms and contain portions of transcripts from many uncharacterized genes. Since ESTs have no annotated coding sequences, there are no corresponding protein translations in the BLAST protein databases. Hence a tblastn search is the only way to search for these potential coding regions at the protein level. The HTG sequences, draft sequences from various genome projects or large genomic clones, are another large source of unannotated coding regions.

Like all translating searches, the tblastn search is especially suited to working with error prone data like ESTs and draft genomic sequences from HTG because it combines BLAST statistics for hits to multiple reading frames and thus is robust to frame shifts introduced by sequencing error.

[\[Back to top\]](#)

4.11 The "Nucleotide query - Translated db [tblastx]" is useful for identifying novel genes in error prone query sequences.

tblastx takes a nucleotide query sequence, translates it in all six frames, and compares those translations to the database sequences dynamically translated in all six frames. This effectively performs a more sensitive blastp search without doing the manual translation.

tblastx gets around the potential frame-shift and ambiguities that may prevent certain open reading frames from being detected. This is very useful in identifying potential proteins encoded by single pass read ESTs. In addition, it can be a good tool for identifying novel genes.

This type of search is computationally intensive and searches with large genomic queries are not recommended. The best way to do this is to install standalone blast and perform the search locally. For more information on standalone blast, please read the documents for formatdb and standalone BLAST at:

<ftp://ftp.ncbi.nih.gov/blast/documents/formatdb.txt>

<ftp://ftp.ncbi.nih.gov/blast/documents/netblast.txt>

[\[Back to top\]](#)

4.12 The Conserved Domain Database (CDD) search service uses RPS-BLAST to identify conserved protein domains.

Reverse Position Specific BLAST (RPS-BLAST) is a more sensitive way of identifying conserved domains in proteins than standard BLAST searching. It compares a protein sequence against a database of position specific scoring matrices (PSSMs). The PSSMs used in CDD search capture the substitution frequencies at each position in the multiple sequence alignments of recognized conserved domains. The conserved domain alignments are from the NCBI's CDD, which contains alignments from protein domain databases: Smart, Pfam, COG, and cd. There is no batch search function available for RPS-BLAST page. For that, you can use the rpsblast program from the standalone blast package. The preformatted database for this program is downloadable from:

<ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/>

For additional information, refer to CDD help document at:

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml

[\[Back to top\]](#)

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

CDART allows you to examine the domain structure of all proteins in the default BLAST protein database. The CDART tool first searches a query sequence for the presence of conserved domains using RPS-BLAST. It then allows you to retrieve proteins that share one or more protein domains in common with your query. The result is sorted according to taxonomy classification, and be further manipulated to display only subsets. Because CDART relies on RPS-BLAST, these searches are more sensitive than ordinary BLAST searches.

If the query does not contain any conserved domains, CDART will not report any result. This search is also linked from web CDD result page.

[\[Back to top\]](#)

5. Explanation for Program Choices Given in Table 3.3

5.1 The Human Genome BLAST page is for comparing a query against the NCBI's assembly of human genome, plus its derivative and related databases.

Like other BLAST search pages in this Genomes section, this page provides a centralized page to access specialized databases. In this case, the databases are the current NCBI human genome build and those derived from or related to it. In addition, alternative assemblies, from Celera (2001) and HSC, are also available. All flavors of BLAST, except tblastx, are available with MEGABLAST set as default. Default filters are DUST and human repeats. The BLAST output links directly to the Human Genome MapViewer, where database hits can be visualized/analyzed in a genomic context, such as their relationship to other map elements like Transcript, SNPs, and Gene_seq. Database nomenclature is being standardized, and their contents are described in Table 2.3.

To download the sequences and human genome mapview data, please visit ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/

[\[Back to top\]](#)

5.2 Use the Mouse and Rat Genome BLAST page to search current assemblies and other sequences specific to those two organisms, respectively.

The organization of these two two pages is similar to that of Human Genome BLAST page. Due to the same concern, tblastx is not provided. MEGABLAST is the default algorithm and "DUST" plus "rodent repeats" are default filters. The default database "all_contigs" is analogous to the human "all_contigs" database. Hits are linked to corresponding MapViewer for visulization.

For Rat, the contigs are NW_ initialed and there is no "Gene Trap Clone" database. The coverage for Rat may be less comprehensive than that for mouse and human. For database information on databases, refer to Table 2.3 above.

To download the sequences and mouse and human genome mapview data, please visit ftp.ncbi.nlm.nih.gov/genomes/M_musculus/
ftp.ncbi.nlm.nih.gov/genomes/R_norvegicus/

5.3 Use the Chicken, Cow, or Dog Genome BLAST pages to search specific sequences from these organisms.

These pages provide access to BLAST databases specific to the organisms listed. The sequence databases include the whole genome shotgun assemblies (wgs) as well as ESTs, HTGs, and Traces databases. The details are listed below. Links to MapView from BLAST result is limited to chicken at this time. For dog, two assemblies are available - boxer breed from Broad Institute and poodle from TIGR.

Table 5.3.1 Databases available for Pig, Sheep, and Cat

Organism	Databases currently available
Chicken	genome, HTGS, Ref RNA, Ref Protein, Build RNA, Build protein, Ab initio RNA, Ab initio protein, ESTs, BAC end sequences, Traces-WGS, Traces-Other, WGS contigs
Cow	WGS contigs, HTGS, RefSeq RNA, RefSeq Proteins, ESTs, BAC end sequences, Traces- WGS, Traces- ESTs, Traces- other
Dog	Reference Dog Assembly (boxer), HTGS, Traces- WGS, Traces- other, TIGR Dog Assembly (poodle), TIGR Dog Extra (poodle)

5.4 Use the Pig, Sheep, Cat Genome BLAST pages to search specific sequences from these organisms.

These pages provide access to BLAST databases specific to the organisms listed. There is no genomic assemblies due to the lack of publicly available genomic sequences. There is also no link to MapViewer, where only maps for physical markers are available. The sequence databases are limited to EST, HTG, and Traces as listed below.

Table 5.4.1 Databases available for Pig, Sheep, and Cat

Organism	Databases currently available
Pig	HTGS, ESTs, BAC end sequences, Traces-other
Sheep	HTGS, ESTs, Traces-other
Cat	HTGS, ESTs, BAC end sequences, Traces-other

[\[Back to top\]](#)

5.5 The Microbial page provides centralized access to complete and unfinished bacterial/archaeal genomes.

This page provides access to many complete and some unfinished bacterial/archeal genomes. The available genomes are listed in the page. The primary dataset is the genome(s), with protein as the derivative dataset. The availability of protein database is marked by red "P" in front of the genome name. Due to the lack of annotation, the protein dataset may not be available for unfinished genomes, which are marked by "green" background. One can choose to search against all the genomes or a selected subset of them, and all flavors of BLAST programs are available.

This is a very dynamic page since the number of available genomes is increasing steadily, and the page is frequently updated to reflect the changes. For BLAST hits to an unfinished genome, only matched region and its immediate flanking sequences are available for downloading.

5.6 The Other eukaryotes BLAST page provides access to genomic sequences of other eukaryotic organisms.

Genomic sequences for many other lower eukaryotes are available from this page. The exact sequences available for BLAST search vary depending on the stage of the sequencing projects. The databases in the page overlaps with Malaria, Fungi, Insects, and Nematodes BLAST pages. The difference is that results from those organism specific pages are hot linked to their corresponding MapViewer for better visualization, which provides more information on genomic context.

Tips for Microbial and Other Eukaryotes BLAST Page:

Information available is indicated by color of background and letter prefix:

- "green" - unfinished
- "blue" - finished
- red "P" - protein database available

To gain more control of the search

- click "Adv. BLAST" button after database selection

[\[Back to top\]](#)

5.7 Environmental Samples page is for finding matches in Sagarso Sea and Mine Drainage Samples.

This page provides access to environmental sequences from two specific projects: Sagarso Sea and Mine Drainage. The dataset overlaps in part with the [env_nt](#) and [env_nr](#) databases accessible through main nucleotide and protein blast page, respectively.

5.8 Use the Fugu genome BLAST page to search against the draft Fugu rubripes (Puffer fish) genome.

This page provides access to the draft genome and the protein translation of *Fugu rubripes* (Japanese Puffer fish), an assembly provided by [Joint Genome Institute](#). For details on the databases and its release policy, please go to [Fugu home page](#). Similar BLAST searches against this genome assembly can also be done [there](#). The current assembly was built in 2002.

5.9 Use the Zebrafish Genome BLAST page to search against Zebrafish specific sequences.

Currently, there are no finished genomic contigs are available from NCBI for Zebrafish. The databases provided in this page are a subset of sequences from this organism present in different NCBI sequence databases.

Table 5.9.1 Content of Zebrafish Genome Blast Databases

Database Name	Content Description
mRNAs	All Zebrafish mRNAs in GenBank.
ESTs	Single pass sequence reads from numerous Zebrafish cDNA libraries.
HTGS	Zebrafish phase 0, 1, 2 or 3 sequences. These are the original BAC sequences as submitted by the sequencing centers.
WGS Traces	All of the raw Zebrafish WGS and BAC Traces from Trace archive.
EST Traces	All of the raw Zebrafish EST Traces from Trace archive.
Reference mRNAs	Zebrafish reference mRNAs generated by the NCBI RefSeq project.

Reference Proteins	Zebrafish reference proteins generated by the NCBI RefSeq project.
--------------------	--------------------------------------------------------------------

[\[Back to top\]](#)

5.10 Use the Plants genome BLAST pages to search against green plant genomes.

This page accesses sequences from a limited number of green plants. For most of the organisms listed, only nucleotide sequences and blastn and tblastn searches are available. Genomic contigs as well as protein sequences are available for *Arabidopsis thaliana* and *Oryza sativa*, and matches are linked to MapView.

Table 5.10.1 Plants Genome BLAST Database Content

Database Name	Content
<i>Arabidopsis thaliana</i> (mustard)	Genome assembly, mRNAs, and Proteins
<i>Avena sativa</i> (Oat)	Currently available genomic clones, GSS, and STS entries
<i>Glycine max</i> (soy bean)	Currently available genomic clone, GSS, EST, and STS entries
<i>Hordeum vulgare</i> (Barley)	Currently available genomic clones, ESTs, GSSs, and STSs
<i>Oryza sativa</i> (Rice)	Currently available genomic clone, EST, GSS, htg, and STS entries
<i>Oryza sativa indica</i> (Indian Rice)	WGS contig assemblies, mRNAs, and Proteins
<i>Oryza sativa ssp. indica</i> WGS contigs (not mapped)	WGS contigs, not yet mapped
<i>Tricum aestivum</i> (Wheat)	Currently available genomic clone, EST, GSS, and STS entries
<i>Zea mays</i> (Corn)	Currently available genomic clone and EST entries
<i>Lycopersicon esculentum</i> (Tomato)	Currently available genomic clone, GSS, EST, and STS entries
Mapped sequences from all listed plants	All of the DNA sequences above.

[\[Back to top\]](#)

5.11 The Nematode BLAST page.

In this page, one can access the *Caenorhabditis* genome and the derivative databases. Matches are linked to MapView. In addition, sequence database for *Caenorhabditis briggsae* is also available (genomic sequences only).

5.12 The Yeast Genome BLAST page provides access to multiple yeast genomes.

This page provides access to different yeast genomes and their protein translations. Sequences of other yeast strains are also available in addition to that for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. The databases for the two well known strains can be searched individually or together. Hits are linked to MapView. All flavors of BLAST, with the exception of tblastx, are available.

Table 5.12.1 Database list for Yeasts page

Organism	Sequence Databases Available
<i>Schizosaccharomyces pombe</i>	Genome, mRNAs, and Proteins
<i>Saccharomyces cerevisiae</i>	Genome, mRNAs, and Proteins
<i>Saccharomyces paradoxus</i> NRRL Y-17217	Nucleotide only
<i>Saccharomyces mikatae</i> IFO1815(MIT)	Nucleotide only
<i>Saccharomyces mikatae</i> IFO1815(WashU)	Nucleotide only
<i>Saccharomyces bayanus</i> MCYC623(MIT)	Nucleotide only
<i>Saccharomyces bayanus</i> MCYC623(WashU)	Nucleotide only
<i>Saccharomyces castellii</i> NRRL Y-12630	Nucleotide only
<i>Saccharomyces kluyveri</i> NRRL Y-12651	Nucleotide only
<i>Saccharomyces kudriavzevii</i> IFO 1802	Nucleotide only
All YEAST Genomes	Genomes, mRNAs, and Proteins
<i>Neurospora crassa</i>	Genome, mRNAs, and proteins
<i>Magnaporthe grisea</i>	Genome, mRNAs, and proteins
<i>Aspergillus nidulans</i>	Genome, mRNAs, and proteins
All Species	Genomes, mRNAs, and Proteins

[\[Back to top\]](#)

5.13 Use the Insects BLAST page to search the *Anopheles gambiae*, *Drosophila melanogaster*, and *Apis mellifera* genomes.

This page provides access to the genome scaffold of *Anopheles gambiae* (mosquito), *Drosophila melanogaster* (fruitfly), and *Bombyx mori* (silkworm/moth) chromosomes. The proteins translated from the genome annotation are also available. The data available for *Anopheles gambiae* is from a NIAID publicly funded project with the sequencing and assembly performed by Celera Corporation. The data for *Drosophila melanogaster* come from FlyBase. Hits are linked to corresponding MapViewer pages, providing additional information. *Apis mellifera* sequences were added recently, which has no MapViewer link yet.

Organism	Service Availability		
	Databases	BLAST searches	MapViewer Link
<i>Drosophila melanogaster</i>	genomic, mRNA, Protein	blastn, blastp, blastx, tblastn	yes
<i>Anopheles gambiae</i>	genomic, mRNA, Protein	blastn, blastp, blastx, tblastn	yes
<i>Apis mellifera</i> (honeybees)	genomic, mRNA, Protein	blastn, blastp, blastx, tblastn	??
<i>Drosophila pseudoobscura</i>	genomic only	blastn, tblastn	no
<i>Bombyx mori</i> (silkworm/moth)	genomic only	blastn, tblastn	no

[\[Back to top\]](#)

6. Explanation on Special Purpose Pages

6.1 "BLAST 2 Sequences" is designed for direct comparison of two sequences.

This program takes two input sequences and compares them directly. "Aligning Two Sequences" regards the second sequence as the database, longer sequence of the generally goes there. Unlike the other BLAST programs, there is no need to format the database sequence in any special way. Since translated BLAST programs are incorporated in this program, the second sequence can be of different type so long as an appropriate BLAST program is selected. Appropriate query/program combination is listed in the table below.

Table 6.1.1 Appropriate Query/Program Combinations for "BLAST 2 Sequences"

First Query	Second Query	Program to Use
Nucleotide	Nucleotide	blastn, megablast, or tblastx
Nucleotide	Protein	blastx
Protein	Nucleotide	tblastn
Protein	Protein	blastp

Tips:

- If the database sequence or second query is present in an NCBI database, using the GI/Accession instead of the FASTA sequence allows the program to incorporate the translation and other sequence features, found in that record, into the final alignment making it more informative.
- To matching primer to a given sequence, you need to increase the Expect to 1000 or higher since it is calculated on the actual size of nr. Unchecking filter checkbox and decrease word size may also be needed.
- For mapping mRNA on to its genomic counterpart, "Align two sequences" page is not the ideal tool since it is not aware of the consensus splicing sites. Instead you should take a look at [Spidey](#).


[\[Back to top\]](#)

6.2 The VecScreen page is for identifying vector sequence contamination in a query sequence.

VecScreen, under special section, is a rapid screening tool that checks the query sequence against a non-redundant vector database, UniVec, which contains one copy of every unique sequence segment from a large number of cloning vectors. In addition, UniVec contains sequences for adapters, linkers, stuffers, and primers that are commonly used in the cloning and manipulation of cDNA or genomic DNA. Detailed information on UniVec is at: www.ncbi.nlm.nih.gov/VecScreen/UniVec.html.

This page is generally used to screen for vector contamination in sequences before their submission to GenBank. The color-coded graphics in the result page makes the result easy to understand.

6.3 The GEO Blast page can be used to search for expression information in the GEO database.

This blast page allows you to blast a given set of sequences to find matches to those sequences/genes represented by entries in the GEO database. Matching hits will have  links linking to corresponding entries in GEO. Different from text query on Entrez/Geo database, this page provides a way to search and retrieval of expression data through sequence similarity search by way of BLAST.

6.4 igblast is for identify matches to curated Immunoglobulin sequences.

This page accesses the curated human and mouse immunoglobulin germline sequences. The databases are updated regularly. Both protein and nucleotide sequences are available for blastn and blastp searches. In part, it functions as a replacement of the now defunct Kabat database, with extra sequence similarity search capability. Help document is linked off this page:

www.ncbi.nlm.nih.gov/igblast/.

6.5 SNP Blast page searches the human dbSNP entries and identifies potential matches to known SNPs.

This page accesses the curated human SNPs available from NCBI's dbSNP database. The search can be against the complete set or limited to those on one or more of the selected chromosomes. For SNP Blast result, displaying the result in "Query anchored with identity" format is likely to be more informative in identify the matching SNPs. Currently, this is not listed in the blast home page yet.

6.6 Retrieve result for an RID, can be used to format a BLAST result in different format.

For each successfully submitted BLAST search request, a unique RID (request ID) is issued. This ID will be valid for 24 hours. Within this period of time, you can use the RID to retrieve the result multiple times. More importantly, the RID can be used to retrieve and display the result in different display formats to emphasize different aspect of the result and bring out the features such as identity or variation across the matches, that would otherwise not stand out. Those representative display formats are described below.

Table 6.6.1 Selected display formats and their description

Format	Description
Pairwise	Query aligned to individual matched sequence, one pair at a time
Pairwise with identity	Similar to "Pairwise", with identity replaced by "."
Query anchored with identity	Pseudo multiple sequence alignment format. Identity in matched subject is replaced by "."
Flat query anchored with identity	Similar to above with whole alignment kept flat by inserting "-" into query if needed.
Hit table	Statistics for each HSP summarized in table format
XML	Result in XML format for parsing (controled by "Format" field)

Footnote: CDD and CDART searches do not have RID issued.

7. Appendices

7.1 Web MEGABLAST can accept batch queries.

MEGABLAST is the only BLAST web service that can accept multiple queries. There are two ways to enter batch queries in MEGABLAST. If the query sequences are not present in the NCBI Entrez system, those sequences need to be provided in FASTA format, one after another with no blank lines in between sequences. The FASTA definition line (or title) of each sequence should be on a single line all by itself. If those sequences are already saved as a text file in proper format, the file can be uploaded using the "Browse" button. An example query file with multiple sequences is given below.

>Sequence-1

```
GAGTGGCAGTTATATAGACCGGCGGCGGAGCACGCGTGTGTGCGGACGCAGTTGCGTGAGGGGTTTGTAC
TATCCTCGGTGCTGTGGTGCAGAGCTAGTTCCTCTCCAGCTCAGCCGCGTAGGTTTGGACATATTTGACT
CTTTTCCCCCAGGTTGAATTGACCAAAGCAATGGTGATGGAGAAGCCTAGTCCCCTGCTGGTCGGGCGG
GAATTTGTGAGACAGTATTACACACTGCTGAACCAGGCCCCAGACATGCTGCATAGATTTTATGGAAAGA
ACTCTTCTTATGTCCATGGGGGATTGGATTCAAATGGAAAGCCAGCAGATGCAGTCTACGGACAGAAAGA
AATCCACAGGAAAGTGATGTCACAAAACCTTACCAACTGCCACACCAAGATTCGCCATGTTGATGCTCAT
GCCACGCTAAATGATGGTGTGGTAGTCCAGGTGATGGGGCTTCTCTCTAACAACAACCAGGCTTTGAGGA
GATTCATGCAAACGTTTGTCTTGTCTCCTGAGGGTCTGTTGCAAATAAATTCATGTTTACAATGATAT
CTTCAGATAACCAAGATGAGGTCTTTGGTGGGTTTGTCACTGAGCCTCAGGAGGAGTCTGAAGAAGAAGTA
GAGGAACCTGAAGAAAGACAGCAAACACCTGAGGTGGTACCTGATGATTCTGGAACCTTCTATGATCAGG
CAGTTGTGAGTAATGACATGGAAGAACATTTAGAGGAGCCTGTTGCTGAACCAGAGCCTGATCCTGAACC
AGAACCAGAACAAGAACCTGTATCTGAAATCCAAGAGGAAAAGCCTGAGCCAGTATTAGAAGAACTGCC
CCTGAGGATGCTCAGAAGAGTCTTCTCCAGCACCTGCAGACATAGCTCAGACAGTACAGGAAGACTTGA
GGACATTTTCTTGGGCATCTGTGACCAGTAAGAATCTTCCACCCAGTGGAGCTGTTCCAGTTACTGGGAT
ACCAC'TCATGTTGTTAAAGTACCAGCTTACAGCCCCGTCCAGAGTCTAAGCCTGAATCTCAGATTCCA
CCACAAAGACCTCAGCGGGATCAAAGAGTGCGAGAACAACGAATAAATATTCCTCCCCAAAGGGGACCCA
```

>Sequence-2

```
TCTGCACTGGAGAGTCTGGAGCTGGGAAGACGGAAAACACCAAGAAGGTCATCCAGTACCTCGCCCACGT
GGCATCGTCTCCAAAGGGCAGGAAGGAGCCGGGTGTCCCCGGTGAGCTGGAGCGGCAGCTGCTTCAGGCC
AACCCCATCCTAGAGGCCTTTGGCAATGCCAAGACAGTGAAGAATGACAACCTCCTCCCGATTTCGGCAAT
TCATCCGCATCAACTTTGATGTTGCCGGGTACATCGTGGGCGCCAACATTGAGACCTACCTGCTGGAGAA
GTCGCGGGCCATCCGCCAGGCCAAGGACGAGTGCAGCTTCCACATCTTCTACCAGCTGCTGGGGGGCGCT
GGAGAGCAGCTCAAAGCCGACCTCCTCCTCGAGCCCTGCTCCACTACCGGTTCTGACCAACGGGCCGT
CATCCTCTCCCGGCCAGGAGCGGGAACCTTCCAGGAGACGCTGGAGTCGCTGCGGGTCTGGGATTTCAG
CCACGAGGAAATCATCTCCATGCTGCGGATGGTCTCAGCAGTTCCTCCAGTTTGGCAACATTGCCTTGAAG
AGAGAACGGAACACCGATCAAGCCACCATGCCTGACAACACAGCTGCACAGAAGCTCTGCCGCCTCTTGG
GACTGGGGGTGACGGATTTCTCCCGAGCCTTGCTCACCCCTCGCATCAAAGTTGGCCGAGACTATGTGCA
GAAAGCCCAGACTAAGGAACAGGCTGACTTCGCGCTGGAGGCCCTGGCCAAGGCCACCTACGAGCGCCTC
```

If the query sequences are already present in an Entrez Nucleotide database, their GI or Accession numbers can be pasted into the search box, one identifier per line.

For example, the two groups of identifiers given in the table are equivalent. Similar to using sequences, a text file containing those ID numbers can be uploaded through the "Browse" button.

Accessions	NCBI GIs
U12345	540023
F12564	708563
BH023812	14647366

For other alternative means of batch BLAST search, refer to "[Other Alternative means for Batch BLAST](#)" (Section 7.3) for more details.

[\[Back to top\]](#)

7.2 Degenerate bases and ambiguity codes are treated as mismatches by BLAST.

Uncertainties in a nucleotide sequence can be represented by a standard set of single-letter codes. These codes are often used to represent degenerate bases in the third position of codons, in degenerate oligo-nucleotide primers, or sequence motifs. Even though ambiguities in the query are accepted by BLAST, BLAST web pages have a built-in functionality that screens query sequences. Too many ambiguities in a nucleotide query could make the BLAST page mistake a nucleotide query as a protein. This will prevent the search from going through and result an error message.

In alignments, BLAST treats the ambiguities in an accepted nucleotide query as mismatches. In short queries, these ambiguous bases may break the query in such a way that no valid word is available for BLAST to index the query and identify initial word hits, thus preventing BLAST from finding any matches in the database.

Table 7.2.1 Single Letter Nucleotide Code

Code	Meaning (Base)	Code	Meaning (Base)
A	adenosine (A)	M	amino (A or C)
C	cytidine (C)	S	strong (G or C)
G	guanine (G)	W	weak (A or T)
T	thymidine (T)	B	not A (G or T or C)
U	uridine (U)	D	not C (G or A or T)
R	purine (G or A)	H	not G (A or C or T)
Y	pyrimidine (T or C)	V	not T (G or C or A)
K	keto (G or T)	N	any base (A or G or C or T)
- 1	gap(s)		

¹ Dash(s) in the query will not be accepted. They will be removed before the search is submitted. To represent gaps, use a string of N's.

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the IUPAC based amino acid codes are given in the table below.

Table 7.2.2 Single Letter Amino Acid Code

Code	Residue	Code	Residue
A	alanine	P	proline
B	aspartate or asparagine	Q	glutamine
C	cysteine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U ¹	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate or glutamine
L	leucine	X	any residue

M	methionine	*	translation stop
N	asparagine	-2	gap of indeterminate length

¹ BLAST cannot handle U properly in protein alignment since it was not specified in the scoring matrices used by blastp. To partially resolve this, U in the query is replaced by an X before the search is performed.

² Dash(s) in the query will be removed before the search is submitted. To represents a gap, a string of X's should be used instead.

³ Blastp treats the ambiguous codes as mismatches in the alignment.

[\[Back to top\]](#)

7.3 Other alternative means for batch BLAST searches.

Even though BLAST home page does not offer batch searches other than blasn via MEGABLAST, we do provide alternatives to users who would like to batch their blastp or other types of BLAST searches. The options and their pros and cons are summarized in the table below.

Alternatives	Pros	Cons	Links
blastcl3	<ul style="list-style-type: none"> - No database maintenance - Simple to set up 	<ul style="list-style-type: none"> - server/network fluctuation - Relative low throughput - No graphic output 	document program
URL-API	<ul style="list-style-type: none"> - Versatility - No database maintenance 	<ul style="list-style-type: none"> - Custom scripts needed - Load restrictions 	document
wwwblast	<ul style="list-style-type: none"> - GUI/graphical output - Easy remote access - Custom databases 	<ul style="list-style-type: none"> - Web server installation - Database Maintenance 	document program
Standalone BLAST	<ul style="list-style-type: none"> - No server fluctuation - Custom databases - High throughput 	<ul style="list-style-type: none"> - Needs database update - No graphic 	document program

[\[Back to top\]](#)

Updated on