BLAST Basic Local Alignment Search Tool

Blast Program Selection Guide

Table of Content

1. Introduction

NCBI has provided BLAST sequence analysis services for over a decade. For many users, the first question they often face is *"Which BLAST program should I use?"* In order to help users arrive at an answer to this question, we created this "BLAST Program Selection Guide."

This document first introduces the BLAST databases available from NCBI (in Section 2). The actual guide (Section 3) divides BLAST searches into several categories according to the *nature* and *size* of the input query and the primary goal of the search. Starting from the query sequence column on the left and cross-referencing to the right, a user will arrive at the specific BLAST program(s) best suited for that search.

This document is also available in PDF (163,516 bytes).

2. BLAST Database Content

A BLAST search has four components: query, database, program, and search purpose/goal. To discuss effective BLAST program selection, we first need to know what databases are available and what sequences these databases contain. In this section, we will first 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 contents for the three important model organisms, i.e., human, mouse, and rat, are described in Table 2.3. For other organisms, the content of their genome blast pages will be listed when these special BLAST pages are discussed.

Table 2.1 Content of Protein Sequence Databases

Database [1]	Content Description
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
refseq	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	Non-redundant CDS translations from env_nt entries.
Smart v4.0 [2]	663 PSSMs from Smart, no longer actively maintained.
Pfam v11.0 [2]	7255 PSSMs from Pfam, not the latest.

COG v1.00 [2]	4873 PSSMs from NCBI COG set.
KOG v1.00 [2]	4825 PSSMs from NCBI KOG set (eukaryotic COG equivalent).
CDD v2.05 [2]	11399 PSSMs from NCBI curated cd set.
NOTE: [1] default database is in bold. [2] These databases are searchable only from rpsblast page, actual version may vary.	

[Back to top]

Table 2.2 Nucleotide Databases for BLAST

Database	Content Description
nr [1]	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	mRNA sequences from NCBI Reference Sequence Project.
refseq_genomic	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 coresponding 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 **W**hole **G**enome **S**hotgun sequences.
env_nt	Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarsso Sea project. This does NOT overlap with nucleotide nr.
NOTE: [1] default database is in bold.	

[Back to top]

Table 2.3 Genome BLAST Databases and Contents [1]

Database [2]	Description
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 *ab initio* 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 *ab initio* 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 assemblid 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 assemblid 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, even though 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 and its availability. 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 ...

NOTE:
[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 ...

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

NOTE:
[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 uncheck 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	BLAST Page to Use	Explanation
Nucleotide: 11 bp or	- [2]	Compare two sequences directly	Align two sequences	Learn more ...

above 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	Find matches to sequences with MicroArray information	GEO BLAST	Learn more ...
	SNP	Find matches to human reference SNPs	SNP BLAST	Learn more ...
-	- [3]	To retrieve results for a search with its RID	Retrieve result for an RID	Learn more ...

Note:
[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.
[3] Requires valid RIDs that are assigned within the past 24 hours.

NOTE:
GenBank® and **BLAST®** are registered trademarks granted to NLM 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, discontiguous-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 discontiguous 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 Discontiguous 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.

A more sensitive search can be achieved by using the newly introduced discontiguous 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, discontiguous 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 discontiguous 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 discontiguous are available at:

www.ncbi.nlm.nih.gov/blast/discontiguous.html
www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html

Parameters unique for discontiguous megablast are:

- word size: retricted 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 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. NCBI provides 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 primers. 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 the reverse primer 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 sequence information, especially for organisms lacking a significant amount of regular mRNA or assembled genomic sequences. The sequence data 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 discontiguous 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 "low complexity filter". Unlike nucleotide BLAST, there is no comparable MEGABLAST for protein searches, so batch search via the web is not supported. To do batch protein BLAST, you can take a look at netblast (blastcl3). Document describing this tool is netblast.html.

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 or new members of a protein family. 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 a multiple 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 matches to new database sequences are found 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.html 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.

An example query sequence and a sample pattern in ProSite format are given below for test run with PHI-BLAST. Pattern occurrence in the query is underlined.

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGQ
```

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

You can click this example search link to get to a PHI-BLAST page with the above query and pattern preloaded to see how they are entered to the PHI-BLAST page.

[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. Very similar to that for short primer searches, 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

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 (as well as nucleotide) pattern search, "seedtop" from NCBI's standalone BLAST package is a much better choice. This tool is described in seedtop.html. The standalone BLAST packages, as the blast initialed archives for different platforms, are under ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/.

[Back to top]

4.9 "Translated query vs protein database (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 "Protein query vs translated database (tblastn)" 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 "Translated query vs translated database (tblastx)" is useful for identifying novel genes in error prone nucleotide 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 should be used only as last resort. Searching with large genomic queries is **NOT** recommended. For users with regular or batch need for this time of searches, the best way 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.ncbi.nlm.nih.gov/blast/documents/formatdb.html
ftp.ncbi.nlm.nih.gov/blast/documents/netblast.html

[Back to top]

4.12 "Search the Conserved Domain Database" uses RPS-BLAST to identify 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 and additional information are available from:

ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/
www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml

[Back to top]

4.13 "Protein homology by domain architecture (cdart)" explores the domain architectures of proteins.

CDART allows you to examine the domain structure of all proteins in theprotein nr database. The CDART tool first searches a query sequence for the presence of conserved domains using RPS-BLAST. It then retrieves proteins that share one or more protein domains in common with your query. The result is sorted according to taxonomy classification, and can 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.

[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.

This page centralizes the access to human specific databases. The default databases are the current NCBI human genome build and alternative assemblies from Celera (2001) and HSC in Toronto, Canada.

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 hits can be visualized and analyzed in a genomic context to see their relationship to other map elements such as Transcript, SNPs, and Gene. Database nomenclature (for higher organims) is 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 "all assemblies" is analogous to that in the human page. Hits are linked to corresponding MapViewer for visualization.

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 Chimp, 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 Chimp, Chicken, Cow, and Dog

Organism	Databases currently available
Chimp	genome (all assemblies), genome (reference only), HTGS, RefSeq RNA, RefSeq protein, Non-RefSeq RNA, Non-RefSeq protein, Build RNA, Build protein, Ab initio RNA, Ab initio protein, ESTs, Clone end sequences, Traces-WGS, Traces- other, Arachne Chimp WGS Contigs, PCAP Chimp WGS Contigs, SNPs
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	genome, 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 are 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 (WGS) 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 WGS 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 rapidly and this page is frequently updated to reflect the changes.

Unfinished genomes not submitted to NCBI as wgs entries are no longer supported.

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 this page overlap with those found in Protozoa, Fungi, Insects, and Nematodes BLAST pages. For better visualization of BLAST hits in Map Viewer (if available), access the BLAST pages through the Map Viewer home page.

Tips for Microbial and Other Eukaryotes BLAST Page:

- "green" represents unfinished genomes with wgs contigs
- "yellow" represents finished genomes
- red "P" presents availablity of protein database
- "Adv. BLAST" button allows better control over the search

[Back to top]

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

This page provides access to environmental sequences from two specific projects: Sagarsso 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 Zebrafish or Fugu genome BLAST page to search against the fish genome.

The zebrafish genome blast page provides access to the recently released zebrafish geneome assembly built on WGS contigs from Sanger. It also provides access to the refseq mRNA and proteins databases plus other sequence databases specific for this organism. Detailed list of available databases is in the table below.

Table 5.8.1 Available Zebrafish Specific Databases

Database	Content
genome (reference only)	RefSeq genomic contigs annotated by NCBI
HTGS	HTG BAC clone sequences, Phase 0-3
RefSeq RNA	RefSeq mRNAs ¹
RefSeq Protein	RefSeq proteins ¹
Non-RefSeq RNA	GenBank mRNAs ¹

Non-RefSeq Protein	GenPept proteins [1]
Build RNA	RefSeq mRNAs from the annotation
Build Protein	RefSeq protein from the annotation
Ab initio RNA	Ab initio predicted RNAs
Ab initio Protein	Ab initio predicted proteins
ESTs	Zebrafish ESTs
Clone End Sequences	Zebrafish GSS entries
Traces- WGS	Zebrafish WGS trace reads
Traces- ESTs	Zebrafish EST trace reads
Traces- Others	Other Zebrafish trace reads
Sanger WGS contigs	WGS contigs from Sanger
SNPs	Zebrafish SNPs from dbSNP

NOTE:
[1] Updated daily.

The Fugu genome blast page provide access to the draft genome (dated 2002) 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 the latest genome assembly can also be done from their Fugu BLAST page.

[Back to top]

5.9 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, mRNAs, 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 [1]	Content
Arabidopsis thaliana (mustard)	Genome assembly, mRNAs, and Proteins
Avena sativa (Oat)	Currently mapped nucleotide entries
Glycine max (soy bean)	Currently mapped nucleotide entries
Hordeum vulgare (Barley)	Currently mapped nucleotide entries
Oryza sativa	Currently genomic assemblies, mRNAs, and proteins
Oryza sativa indica (Indian Rice)	WGS contig assemblies, mRNAs, and Proteins
Oryza sativa ssp. indica WGS contigs (not mapped)	WGS contigs, not yet mapped
Tricicum aestivum (Wheat)	Currently mapped nucleotide entries, plus available ESTs
Zea mays (Corn)	Currently mapped nucleotide entries
Lycopersicon esculentum (Tomato)	Currently mapped nucleotide entries
Mapped sequences from all listed plants	All of the mapped DNA sequences from above.

NOTE:
[1] Organisms with complete genome and MapViewer linked BLAST results are in bold.

[Back to top]

5.10 The Nematode BLAST page.

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

5.11 The Fungi Genome BLAST page provides access to multiple fungal genomes.

This page provides access to complete genomes for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* as well as genomes for other fungi in various finished stage. Protein sequences from the genome annotation are also provided when available. One can search them individually or in combination. Hits are not 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
Ajellomyces capsulatus NAm1	W
Aspergillus clavatus NRRL 1	W
Aspergillus flavus NRRL3357	W
Aspergillus fumigatus Af293	F+P
Aspergillus nidulans FGSC A4	W+P
Aspergillus terreus ATCC 20542	W
Aspergillus terreus NIH2624	W
Botryotinia fuckeliana B05.10	W
Chaetomium globosum CBS 148.51	W
Coccidioides immitis RS	W
Gibberella moniliformis 7600	W
Gibberella zeae PH-1	W+P
Magnaporthe grisea 70-15	W
Neosartorya fischeri NRRL 181	W
Neurospora crassa	W+P
Phaeosphaeria nodorum SN15	W
Sclerotinia sclerotiorum 1980	W
Trichoderma reesei QM9414	W
Uncinocarpus reesii 1704	W
Candida albicans	F+P (Chromosome 7 only)
Candida albicans SC5314	W+P
Candida glabrata CBS138	F+P
Candida tropicalis MYA-3404	W
Clavispora lusitaniae ATCC 42720	W
Debaryomyces hansenii CBS767	F+P
Eremothecium gossypii	F+P
Kluyveromyces lactis NRRL Y-1140	F+P
Kluyveromyces waltii NCYC 2644	W
Pichia guilliermondii ATCC 6260	W
Saccharomyces bayanus 623-6C	W
Saccharomyces bayanus MCYC 623	W
Saccharomyces castellii NRRL Y-12630	W
Saccharomyces cerevisiae	F+P
Saccharomyces cerevisiae RM11-1a	W
Saccharomyces cerevisiae YJM789	W
Saccharomyces kluyveri NRRL Y-12651	W
Saccharomyces kudriavzevii IFO 1802	W
Saccharomyces mikatae IFO 1815	W
Saccharomyces paradoxus NRRL Y-17217	W
Yarrowia lipolytica CLIB122	F+P
Schizosaccharomyces pombe 972h-	F+P
Coprinopsis cinerea okayama7#130	W

Cryptococcus neoformans R265	W
Cryptococcus neoformans var. grubii H99	W
Cryptococcus neoformans var. neoformans B-3501A	W+P
Cryptococcus neoformans var. neoformans JEC21	F+P
Phanerochaete chrysosporium RP-78	W
Ustilago maydis 521	W+P
Encephalitozoon cuniculi	F+P
Encephalitozoon cuniculi GB-M1	F+P
Rhizopus oryzae RA 99-880	W
NOTE: W=wgs; P=protein; F=complete	

[Back to top]

5.12 Use the Protozoa BLAST page to search the protozoa genomes.

This page provides access to the finished and wgs assembly of several medically important protozoan genomes. Available databases are list below in Table 5.12.1. There is no direct link from hits to MapViewer.

Table 5.12.1 Protozoan Genome Databases and Type of Searches Available		
Organism	Databases	BLAST searches
Cryptosporidium hominis	W+P	blastn, blastp, blastx, tblastn
Cryptosporidium parvum	F+P	blastn, blastp, blastx, tblastn
Plasmodium berghei	W+P	blastn, blastp, blastx, tblastn
Plasmodium chabaudi	W+P	blastn, tblastn
Plasmodium falciparum 3D7	F+P	blastn, tblastn
Plasmodium vivax	W	blastn, tblastn
Plasmodium yoelii yoelii	W+P	blastn, tblastn
Theileria parva	F+P	blastn, tblastn
Leishmania major strain Friedlin	F+P	blastn, blastp, blastx, tblastn
Trypanosoma brucei	W	blastn, tblastn
Trypanosoma brucei TREU927	F+P	blastn, blastp, blastx, tblastn
Trypanosoma cruzi	W+P	blastn, blastp, blastx, tblastn
Cyanophora paradoxa	F+P, Mitochondrion only	blastn, blastp, blastx, tblastn
Dictyostelium discoideum	F+P	blastn, blastp, blastx, tblastn
Entamoeba histolytica HM-1:IMSS	W+P	blastn, blastp, blastx, tblastn
Giardia lamblia ATCC 50803	W+P	blastn, tblastn
Tetrahymena thermophila SB210	W	blastn, tblastn
Thalassiosira pseudonana CCMP1335	W	blastn, tblastn
Trichomonas vaginalis	W	blastn, tblastn
NOTE: W=wgs; P=protein; F=complete		

5.13 Use the Insects BLAST page to search the available genomes for various insects.

This page provides access to the genomes for several insects. For a subset of the genomes, protein sequences translated from the genome annotation are also available. *Anopheles gambiae*, *Drosophila melanogaster*, and *Apis melliferacome* do have Map Viewer display available. However, BLAST searches performed through this page will not have direct link to MapViewer display.

Table 5.13.1 Service Availability From Insect Blast Page			
Organism	Databases	BLAST searches	MapViewer Link [1]

Aedes aegypti	W	blastn, tblastn	-
Anopheles gambiae str. PEST	F+P	blastn, blastp, blastx, tblastn	Yes
Apis mellifera	W+P	blastn, blastp, blastx, tblastn	Yes
Bombyx mori	W	blastn, tblastn	-
Drosophila melanogaster	F+P	blastn, blastp, blastx, tblastn	Yes
Drosophila persimilis	W	blastn, tblastn	-
Drosophila pseudoobscura	W+P	blastn, blastp, blastx, tblastn	-
Drosophila sechellia	W	blastn, tblastn	-
Drosophila simulans	W	blastn, tblastn	-
Drosophila yakuba	W	blastn, tblastn	-
Tribolium castaneum	W	blastn, tblastn	-
NOTE: [1] Graphic visualization of BLAST hits on the genome through MapViewer available. Accessing BLAST through Map Viewer Home Page is recommended.			

[Back to top]

6. Explanation on Special Purpose Pages

6.1 "Align Two Sequences" page 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 should go there. Unlike the other BLAST programs, there is no need to format the database sequence in any special way. Recent changes removed the need of separate input box for GI or Accession. GI and Accession should be pasted in the same text window as the FASTA sequences.
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 [1]	Program to Use
Nucleotide	Nucleotide	blastn, megablast, or tblastx
Nucleotide	Protein	blastx
Protein	Nucleotide	tblastn
Protein	Protein	blastp
NOTE: [1] Larger sequence should be used as second query. Use GI or Accession and subsequence coordinates whenever possible for better feature display.		

Tips:

- If the database sequence or second query is present in an NCBI database, using the GI or 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.
- If the GI or accession represents a large sequence, one should use the subsequence feature (marked by "from" and "to" fields) to help reduce the time needed for sequence retrieval and actual search.
- 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 or splign.

[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 UniVec, which contains a non-redundant set of unique vector sequence segment from a large number of known 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 "E" gif icons links 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. The actual sequence alignment becomes secondary in this case.

6.4 igblast is for identify matches to curated human and mouse immunoglobin sequences.

This page accesses the curated human and mouse immunoglobin gemline 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 reference SNP entries from various organisms and identifies potential matches to known SNPs.

This page accesses the curated SNPs from NCBI's dbSNP database. The access has expanded to cover several organisms in addition to human. Default is nucleotide search using megablast with DUST and human repeat filter. Translated search using tblastn is also supported, which requires an input protein query. SNP Blast result is displayed in "Pairwise with identity" format, which highlights the mismatches in red. In certain cases, change the display to "Query anchored with identity" format may be more informative.

[Back to top]

6.6 "Retrieve result for an RID" provides multiple accesses to the same result in various formats.

For each successfully submitted BLAST search request, a unique request ID (RID) is issued. This RID 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 formats to emphasize different aspect of the result and bring out the features, such as identity or variation across the matches, that otherwise would not stand out. Those representative display formats are described below.

Table 6.6.1 Selected display formats and their description

Format [1]	Description
Pairwise	Query aligned to individual matched sequence, one pair at a time
Pairwise with identity	Similar to "Pairwise", but identities are 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 tab delimitted format

XML	Result in XML format ideal for parsing (controled by "Format" field)
NOTE: ¹ CDD and CDART searches do not have RIDs or alternative display format.	

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 (defline) 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 two sequences is given below.

```
>EST_Clone_DW1
ATAGTACAAACTTAGGGCTCTTTATTCAGGCAGTAAAGTAAGGAACAGCAAAGTGGGAGGGCTACACCAT
CACCATGGCAACAGAAAGCCTCAAAAACATAAAGTCCCTCGACTTATGTCGGGTAGACTCTTCCTAGCTC
AGGAGAAACACATTTTAACTGGCTGAGGACAAGGCCAGGCAGCCTGGCCACACTGCGGAAGGGCAGNTGG
ACGCGCGGCCTCTGGTCAGTCCTGGAAGTGCTTGGTGAGGGCTTCCAGCAGCTCCTGCTTCTTCAGACCA
CTCTTCAGCCCGTAAGCCCGGCAGGCCTCTTTCAGCATGGGCACAAGTGAACTTGCCCAGCGTACCCTTG
CTGATGTGGGTCTTCAGCTCCTCTTCTGAATACTCCACCTTGGGCCTTTTGCTTCCAGAACCTTCATTAT
CGNGTTTCTCTTGGTAACTTTCCCTTCAGGAATGTAATCTGGTGGGTAAACAAGCTCCTTAAACTATTCA
CCAAGGANCCANGTTTTTATCAATTGNTTCAACCTTGGCAATGTAAGGNCACTGGTTGGTNCCGGTTCAT
AAAATCCAAGGCCAAGGCCTTCCAGTT
>EST_Clone_AI2
GCACGAGGGTCATTTCCTTTCTTCATGTACCAGATGCTGAAATACTATGAGATAAAGATTTTAGGTTTCA
ATTGTAAAGAGAGAGAAGTGGATAAATCAGTGCTGCTTTCTTTAGGACGAAAGAAGTATGGAGCAGTGGG
ATCACTTTCACAATCAACAGGAGGACACTGATAGCTGCTCCGAATCTGTGAAATTTGATGCTCGCTCAAT
GACAGCTTTGCTTCCTCCGAATCCTAAAAACAGCCCTTCCCTTCAAGAGAAACTGAAGTCCTTCAAAGCT
GCACTGATTGCCCTTTACCTCCTCGTGTTTGCAGTTCTCATCCCTCTCATTGGAATAGTGGCAGCTCAAC
TCCTGAAGTGGGAAACGAAGAATTGCTCAGTTAGTTCAACTAATGCAAATGATATAACTCAAAGTCTCAC
GGGAAAAGGAAATGACAGCGAAGAGGAAATGAGATTTCAAGAAGTCTTTATGGAACACATGAGCAACATG
GAGAAGAGAATCCAGCATATTTTAGACATGGAAGCCAACCTCATGGACACAGAGCATTTCCAAAATTTCA
GCATGACAACTGATCAAAGATTTAATGACATTCTTCTGCAGCTAAGTACCT
```

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
BC006850	13905125
NM_000586	28178860
AY414236	39770198

Click the following three links to see Megablast pages preloaded with multiple queries in various format:

Two FASTA sequences
Three Accessons
Three GIs

For other 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 from IUPAC. These codes are often used to represent degenerate bases in the third position of codons, in degenerate oligo-nucleotide primers, or sequence motifs. However, BLAST treats them all the same - as ambiguous mismatch like N.

Even though BLAST can take nucleotide queries with ambiguities, BLAST web pages have a built-in functionality that screens query sequences and reject those with too many ambiguities. In alignments, BLAST will treat 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. Any attempt to identify consensus patterns using BLAST will likely fail for this reason.

Table 7.2.1 Single Letter Nucleotide Code

Code[1]	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)
-[2]	gap(s)		

[1] Fully accepted nucleotide code are in bold.
[2] Dash(s) in the query will not be accepted. They will be removed before the search is submitted. To better represent a gap, 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 [2]	any residue
M	methionine	*	translation stop
N	asparagine	- [3]	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] Blastp treats the ambiguous codes as mismatches in the alignment.
[3] Dash(s) in the query will be removed before the search is submitted. To better represent a gap, use a string of

X's.

Protein queries, consisting of mostly ACGTN, may be rejected by BLAST for their similarity to nucleotide query. For example, the peptide below will be rejected due to its extreme G-biased composition:

>
>gi|295808:58-99 glycine-rich protein [Hordeum vulgare subsp. vulgare]
> GGGGYGGGGGYGGGGGGYPGGGGGYGGGGGGYPGHGGEGGGG

To make it acceptable by protein BLAST pages, we can append a string of X's to it

>
>gi|295808:58-99 glycine-rich protein [Hordeum vulgare subsp. vulgare]
> GGGGYGGGGGYGGGGGGYPGGGGGYGGGGGGYPGHGGEGGGGXXXXXXXXXX

Again to search for patterns or motifs, seedtop in standalone is a much better tool. See primer search section for more information.

[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.

Table 7.3 Alternatives Means for Batch BLAST Searches

Alternatives	Pros	Cons	Links
blastcl3	• No db maintenance • Simple to set up • Batch capable	• server/network fluctuation • Relative low throughput • No graphic interface or output	netblast.html program
URL-API	• Versatility • No database maintenance	• Custom scripts needed • Load restrictions • Server fluctuation	urlapi.html Sample script
wwwblast	• GUI/graphical output • Easy remote access • Custom databases	• Web server installation • Database Maintenance	wwwblast.html program
Standalone BLAST	• No server fluctuation • Custom databases • High throughput	• Needs database update • No graphic	blast.html program

[Back to top]

Updated on 01/07/2009 17:05:13