NCBI Program Selection Guide BLAST

BLAST Program Selection Guide

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

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 table is also provided in PDF format, one can click the icon to download it.

2. Common BLAST databases

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 content of those databases will be listed when the relevant special BLAST pages are discussed.

Table 2.1 Content of Protein Sequence Databases

Database Name	**Content Description**
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF.
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.

[Back to top]

Table 2.2 Nucleotide Databases for BLAST

Database Name	**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.
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.
wgs	Assemblies of **w**hole **g**enome **s**hotgun sequences.

[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	discontiguos megablast or blastn	Learn more ...
		Find similar sequence from the Trace arcchive	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 Genome or Special Databases

Query [1]	**Database**	**Purpose**	**Program Pages to Use** [3]	**Explanation**
	- [2]	Compare two sequences directly	Align two sequences	Learn more ...
	Human Genome		Human	Learn more ...
	Mouse Genome		Mouse	Learn more ...
	Rat Genome		Rat	Learn more ...

Nucleotide: **20 or 28 bp and above** **Protein:** **15 residues and above**	Fugu (Pufferfish)	Map the query sequence Determine the genomic structure Identify novel genes Find homologs Other data mining	Fugu rubripes	Learn more ...
	Zebrafish		Zebrafish	Learn more ...
	Insects (fruit fly, mosquito, and honeybees)		Insects	Learn more ...
	Nematodes (worms)		Nematodes	Learn more ...
	Plants		Plants	Learn more ...
	Fungi Genomes (including yeasts)		Fungi	Learn more ...
	Malaria		Malaria	Learn more ...
	Other Lower Eukaryotic Genomes		Other eukaryotes genomes	Learn more ...
	Microbial Genomes		Microbial genomes	Learn more ...
	Immunoglobulin sequences	Find matches to curated immunoglobulin sequences	Trace MEGABLAST	Learn more ...
Nucleotide: **20 or 28 bp and above**	UniVec	Screen for vector contamination	VecScreen	Learn more ...

[1] This is similar to what is in Table 3.1 and Table 3.2. 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 2000 -W 7. This also requires the unchecking of the megablast checkbox.

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

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

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 2.1 to 2.2

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 overrides the significance threshold set by Expect value parameter.

Web MEGABLAST can also accept batch queries, the only web BLAST tool with this capability. Please refer to the "Batch Search" section for details.

[Back to top]

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

The BLAST nucleotide algorithm finds similar sequences by breakin 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. 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 of 11 to a minimum of 7 to increase search sensitivity.

The search sensitivity can further improved 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 work 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:

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

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 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 optimum 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 KMKGSMGYYGSNNNNNNGCTYRGCWCSYTC or CNNGAANNTCCNNG 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 discontiguous MEGABLAST. The former is better for indentifying exact matches in intra-species searches, such as looking for extra mRNA sequences or the genomic counterparts for a given gene, while the later is better for indentifying similar coding sequences from different species. Information on the Trace Archive is available from the Trace documentation page.

[Back to 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" and "Composition based statistics" function. Unlike nucleotide BLAST, there is no comparable MEGABLAST for protein searches, so batch search via the web is not possible.

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). In the second iteration the PSSM becomes the query in the 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.

You can also save a PSSM created during a PSI-BLAST search of one database and use it to search a different database. 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 PSSM and paste it into the PSSM window of a new PSI-BLAST search page.

[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. The documentation on pattern syntax is at:

http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html

Sample query sequence, with modified defline and highlighted pattern occurance, and a sample pattern in ProSite format are given below:

```
>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].
```

[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 less) 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	Composition based Statistics	Score Matrix
Standard Protein Blast	3	On	10	On	BLOSUM62
Search for short and nearly exact matches	2	Off	20000	Off	PAM30

[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 the translation 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.

4.10 The "Protein query - Translated db [tblastn]" search is useful for finding protein homologs in unnannotated 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, KOG, and LOAD. 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. 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 "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. 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 5.1.1 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

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.

[Back to top]

5.2 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. 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. Names for the databases are being standardized, refer to Table 5.3.1 for details on database content.

[Back to top]

5.3 Use the Mouse Genome BLAST page to search current assemblies and other mouse sequences.

The organization of this page 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 "genome" is analogous to the human genome database. Hits are linked to Mouse Genome MapViewer for visulization. The content of databases available for searching is given below. [Content derived from http://www.ncbi.nlm.nih.gov/genome/seq/Database.html]

Table 5.3.1 Mouse Genome BLAST Databases and Contents

curated NT contigs	Assemblies of Finished Mouse BAC clones. These are annotated with SNPs and STSs and available in GenBank
HTGS	Mouse phase 0, phase 1, phase 2 or phase 3 sequence. These are the original BAC sequences as submitted by the sequencing centers.
genome (default)	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 they 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.
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
Ref RNA	Collection of reference mRNAs generated by the NCBI RefSeq project. This database is generated daily.
Ref 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. Not available for human.
Gene Trap	

Clones (Mouse Only)	A collection of sequences generated by performing Gene Trap insertions. This database is updated weekly.

[Back to top]

5.4 Use the Rat BLAST page to search rat genome assembly and other rat related sequence databases.

This page provides access to BLAST databases specific for rat. Databases available are similar to those for mouse with the exception that the contig accessions are NW_ initialed and there is no "Gene Trap Clone" database. Hits are linked to visual display in Rat Genome Mapview. For database information, refer to Table 5.3.1.

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. Due to the lack of annotation, the protein dataset may not be available for unfinished genomes. 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 flaking 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.

[Back to top]

5.7 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 DOE Joint Genome Institute (http://www.jgi.doe.gov/). For details on the databases and its release policy, please go to http://genome.jgi-psf.org/fugu6/fugu6.home. Similar BLAST searches against this genome assembly can also be done there (http://aluminum.jgi-psf.org/prod/bin/runBlast.pl?db=fugu6).

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

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

Table 5.8.1 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, phase 1, phase 2 or phase 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]

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

Currently, only nucleotide sequences are available from this page for a limited number of green plants. For this reason, only blastn and tblastn searches are available.

Arabidopsis thaliana is the only exception, for which protein data is available and hits identified are linked to Mapviewer.

Table 5.9.1 Plants Genome BLAST Database Content

Database Name	Content
Arabidopsis thaliana (mustard)	Genome assembly, mRNAs, and Proteins
Avena sativa (Oat)	Currently available genomic clones, GSS, and STS entries
Glycine max (soy bean)	Currently available genomic clone, GSS, EST, and STS entries
Hordeum vulgare (Barley)	Currently available genomic clones, ESTs, GSSs, and STSs
Oryza sativa (Rice)	Currently available genomic clone, EST, GSS, htg, and STS entries
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 available genomic clone, EST, GSS, and STS entries
Zea mays (Corn)	Currently available genomic clone and EST entries
Lycopersicon esculentum (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]

5.10 The Nematode BLAST page.

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

5.11 Yeasts 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 5.11.1 Database list for Yeasts page

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

[Back to top]

5.12 Use the Flies 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) and Drosophila melanogaster (fruitfly) 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.

[Back to top]

5.13 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: http://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. Appendix

6.1 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
CTTTTCCCCCCAGGTTGAATTGACCAAAGCAATGGTGATGGAGAAGCCTAGTCCCCTGCTGGTCGGGCGG
GAATTTGTGAGACAGTATTACACACTGCTGAACCAGGCCCCAGACATGCTGCATAGATTTTATGGAAAGA
ACTCTTCTTATGTCCATGGGGGATTGGATTCAAATGGAAAGCCAGCAGATGCAGTCTACGGACAGAAAGA
AATCCACAGGAAAGTGATGTCACAAAACTTCACCAACTGCCACACCAAGATTCGCCATGTTGATGCTCAT
GCCACGCTAAATGATGGTGTGGTAGTCCAGGTGATGGGGCTTCTCTCTAACAACAACCAGGCTTTGAGGA
GATTCATGCAAACGTTTGTCCTTGCTCCTGAGGGGTCTGTTGCAAATAAATTCTATGTTCACAATGATAT
CTTCAGATACCAAGATGAGGTCTTTGGTGGGTTTGTCACTGAGCCTCAGGAGGAGTCTGAAGAAGAAGTA
GAGGAACCTGAAGAAAGACAGCAAACACCTGAGGTGGTACCTGATGATTCTGGAACTTTCTATGATCAGG
CAGTTGTCAGTAATGACATGGAAGAACATTTAGAGGAGCCTGTTGCTGAACCAGAGCCTGATCCTGAACC
AGAACCAGAACAAGAACCTGTATCTGAAATCCAAGAGGAAAAGCCTGAGCCAGTATTAGAAGAAACTGCC
CCTGAGGATGCTCAGAAGAGTTCTTCTCCAGCACCTGCAGACATAGCTCAGACAGTACAGGAAGACTTGA
GGACATTTTCTTGGGCATCTGTGACCAGTAAGAATCTTCCACCCAGTGGAGCTGTTCCAGTTACTGGGAT
ACCACCTCATGTTGTTAAAGTACCAGCTTCACAGCCCCGTCCAGAGTCTAAGCCTGAATCTCAGATTCCA
CCACAAAGACCTCAGCGGGATCAAAGAGTGCGAGAACAACGAATAAATATTCCTCCCCAAAGGGGACCCA
>Sequence-2
TCTGCACTGGAGAGTCTGGAGCTGGGAAGACGGAAAACACCAAGAAGGTCATCCAGTACCTCGCCCACGT
GGCATCGTCTCCAAAGGGCAGGAAGGAGCCGGGTGTCCCCGGTGAGCTGGAGCGGCAGCTGCTTCAGGCC
AACCCCATCCTAGAGGCCTTTGGCAATGCCAAGACAGTGAAGAATGACAACTCCTCCCGATTCGGCAAAT
TCATCCGCATCAACTTTGATGTTGCCGGGTACATCGTGGGCGCCAACATTGAGACCTACCTGCTGGAGAA
GTCGCGGGCCATCCGCCAGGCCAAGGACGAGTGCAGCTTCCACATCTTCTACCAGCTGCTGGGGGGCGCT
GGAGAGCAGCTCAAAGCCGACCTCCTCCTCGAGCCCTGCTCCCACTACCGGTTCCTGACCAACGGGCCGT
CATCCTCTCCCGGCCAGGAGCGGGAACTCTTCCAGGAGACGCTGGAGTCGCTGCGGGTCCTGGGATTCAG
CCACGAGGAAATCATCTCCATGCTGCGGATGGTCTCAGCAGTTCTCCAGTTTGGCAACATTGCCTTGAAG
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 equivilant. Similar to using sequences, a text file containing those ID numbers can be uploaded through the "Browse" button.

Accessions	NCBI GIs
U12345 F12564 BH023812	540023 708563 14647366

For other alternative means of batch BLAST search, refer to " Other Alternative means for Batch BLAST" (Section 6.3) for more details.

[Back to top]

6.2 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.

BLAST treats the ambiguities in an accepted nucleotide query as mismatches in alignments. 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 6.2.1 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 6.2.2 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 anbiguous codes as mismatches in the alignment.

[Back to top]

6.3 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 6.3 Alternatives Means for Batch BLAST Searches

Alternatives	Pros	Cons	Links
blastcl3	• No database maintainance • Simple to set up	• server/network fluctuation • Relative low throughput • No graphic output	document program
URL-API	• Versatility • No database maintainance	• Custom scripts needed • Load restrictions	document
Standalone BLAST	• No server fluctuation • Custom databases • High throughput	• Needs database update • No graphic	document program

[Back to top]

Updated on Sat Feb 14 14:12:00 2004